Using Only Link-Local Addressing Inside an IPv6 Network
draft-behringer-lla-only-01

Abstract

This document proposes to use only IPv6 link-local addresses on infrastructure links between routers, wherever possible. It discusses the advantages and disadvantages of this approach to aide the decision process for a given network,

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on January 17, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.
Table of Contents

1. Introduction .................................................. 3
   1.1. Requirements Language .................................. 3
2. Using Link-Local Address on Infrastructure Links ............... 3
   2.1. The Suggested Approach ................................... 3
   2.2. Advantages .................................................. 4
   2.3. Caveats and Possible Workarounds ......................... 5
   2.4. Summary ..................................................... 6
3. Security Considerations ....................................... 6
4. IANA Considerations ............................................ 6
5. Acknowledgements .............................................. 6
6. References ..................................................... 7
   6.1. Normative References ...................................... 7
   6.2. Informative References .................................... 7
Authors’ Addresses ................................................. 7
1. Introduction

An infrastructure link between a set of routers typically does not require global or even unique local addressing [RFC4193]. Using link-local addressing on such links has a number of advantages, for example that routing tables do not need to carry link addressing, and can therefore be significantly smaller. This helps to decrease failover times in certain routing convergence events. An interface of a router is also not reachable beyond the link boundaries, therefore reducing the attack horizon.

We propose to configure neither globally routable IPv6 addresses nor unique local addresses on infrastructure links of routers, wherever possible. We recommend to use exclusively link-local addresses on such links.

This document discusses the advantages and caveats of this approach.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] when they appear in ALL CAPS. These words may also appear in this document in lower case as plain English words, absent their normative meanings.

2. Using Link-Local Address on Infrastructure Links

This document proposes to use only link-local addresses (LLA) on all router interfaces on infrastructure links. Routers typically do not need to be reached from users of the network, nor from outside the network. For a network operator, there may be reasons to send packets to an infrastructure link for certain monitoring tasks; we suggest that many of those tasks could also be handled differently, not requiring routable address space on infrastructure links.

2.1. The Suggested Approach

Neither global IPv6 addresses nor unique local addresses are configured on infrastructure links. In the absence of specific global or unique local address definitions, the default behavior of routers is to use link-local addresses. These link-local addresses MAY be hard-coded to prevent the change of EUI-64 addresses when changing of MAC address (such as after changing a network interface card).

ICMPv6 [RFC4443] error messages (packet-too-big...) are required for
routers, therefore a loopback interface MUST be configured with a
global scope IPv6 address. This global scope IPv6 address MUST be
used as the source IPv6 address for all generated ICMPv6 messages.

The effect on specific traffic types is as follows:

- Control plane protocols, such as BGP, ISIS, OSPFv3, RIPng, PIM
  work by default or can be configured to work with link-local
  addresses.

- Management plane traffic, such as SSH, Telnet, SNMP, ICMP echo
  request ... can be addressed to loopback addresses of routers with
  a global scope address. Router management can also be done over
  out-of-band channels.

- ICMP error message can also be sourced from the global scope
  loopback address.

- Data plane traffic is forwarded independently of the link address
  type.

- Neighbor discovery (neighbor solicitation and neighbor
  advertisement) is done by using link-local unicast and multicast
  addresses, therefore neighbor discovery is not affected.

We therefore conclude that it is possible to construct a working
network in this way.

2.2. Advantages

Smaller routing tables: Since the routing protocol only needs to
carry one loopback address per router, it is smaller than in the
traditional approach where every infrastructure link addresses are
 carried in the routing protocol. This reduces memory consumption,
and increases the convergence speed in some routing failover cases.
Note: smaller routing tables can also be achieved by putting
interfaces in passive mode for the IGP.

Reduced attack surface: Every globally routable address on a router
constitutes a potential attack point: a remote attacker can send
traffic to that address, for example a TCP SYN flood, or he can
intent SSH brute force password attacks. If a network only uses
loopback addresses for the routers, only those loopback addresses
need to be protected from outside the network. This significantly
eases protection measures, such as infrastructure access control
lists. See also [I-D.ietf-grow-private-ip-sp-cores] for further
discussion on this topic.
Lower configuration complexity: LLAs require no specific configuration, thereby lowering the complexity and size of router configurations. This also reduces the likelihood of configuration mistakes.

Simpler DNS: Less address space in use also means less DNS mappings to maintain.

2.3. Caveats and Possible Workarounds

Interface ping: If an interface doesn’t have a globally routable address, it can only be pinged from a node on the same link. Therefore it is not possible to ping a specific link interface remotely. A possible workaround is to ping the loopback address of a router instead. In most cases today it is not possible to see which link the packet was received on; however, RFC5837 [RFC5837] suggests to include the interface identifier of the interface a packet was received on in the ICMP response; it must be noted that there are little implemented of this extension. With this approach it would be possible to ping a router on the loopback address, yet see which interface the packet was received on. To check liveness of a specific interface it may be necessary to use other methods, for example to connect to the router via SSH and to check locally.

Traceroute: Similar to the ping case, a reply to a traceroute packet would come from a loopback address with a global address. Today this does not display the specific interface the packets came in on. Also here, RFC5837 [RFC5837] provides a solution.

Hardware dependency: LLAs are usually EUI-64 based, hence, they change when the MAC address is changed. This could pose problem in a case where the routing neighbor must be configured explicitly (e.g. BGP) and a line card needs to be physically replaced hence changing the EUI-64 LLA and breaking the routing neighborship. But, LLAs can be statically configured such as fe80::1 and fe80::2 which can be used to configure any required static routing neighborship.

NMS toolkits: If there is any NMS tool that makes use of interface IP address of a router to carry out any of NMS functions, then it would no longer work, if the interface is missing globally routable address. A possible workaround for such tools is to use the globally routable loopback address of the router instead.

MPLS and RSVP-TE [RFC3209] allows establishing MPLS LSP on a path that is explicitly identified by a strict sequence of IP prefixes or addresses (each pertaining to an interface or a router on the path). This is commonly used for FRR. However, if an interface uses only a link-local address, then such LSPs can not be established. A
possible workaround is to use loose sequence of IP prefixes or addresses (each pertaining to a router) to identify an explicit path along with shared-risk-link-group (to not use a set of common interfaces).

2.4. Summary

Using link-local addressing only on infrastructure links has a number of advantages, such as a smaller routing table size and a reduced attack surface. It also simplifies router configurations. However, the way certain network management tasks are carried out has to be adapted to provide the same level of detail, for example interface identifiers in traceroute.

3. Security Considerations

Using LLAs only on infrastructure links reduces the attack surface of a router: Loopback addresses with globally routed addresses are still reachable and must be secured, but infrastructure links can only be attacked from the local link. This simplifies security of control and management planes. The proposal does not impact the security of the data plane. This proposal does not address control plane [RFC6192] attacks generated by data plane packets (such as hop-limit expiration).

As in the traditional approach, also this approach relies on the assumption that all routers can be trusted due to physical and operational security.

4. IANA Considerations

There are no IANA considerations or implications that arise from this document.

5. Acknowledgements

The authors would like to thank Salman Asadullah, Janos Mohacsi and Wes George for their useful comments about this work.

6. References
6.1. Normative References


6.2. Informative References


Authors’ Addresses

Michael Behringer
Cisco
400 Avenue Roumanille, Bat 3
Biot, 06410
France

Email: mbehring@cisco.com
Eric Vyncke
Cisco
De Kleetlaan, 6A
Diegem, 1831
Belgium

Email: evyncke@cisco.com
Abstract

This document provides guidance and suggestions for Internet Content Providers and Application Service Providers who wish to offer their service to both IPv6 and IPv4 customers. Many of the points will also apply to any enterprise network preparing for IPv6 users.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on August 26, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
Table of Contents

1. Introduction ..................................................... 3
2. General Strategy .................................................. 3
3. Education and Skills .............................................. 5
4. Arranging IPv6 Connectivity ..................................... 6
5. IPv6 Infrastructure ............................................... 6
   5.1. Address and subnet assignment ............................. 6
   5.2. Routing ..................................................... 7
   5.3. DNS ......................................................... 8
6. Load Balancers ................................................... 8
7. Proxies ........................................................... 9
8. Servers .......................................................... 9
   8.1. Network Stack ................................................ 9
   8.2. Application Layer .......................................... 10
   8.3. Geolocation ................................................ 10
9. Coping with Transition Technologies ............................. 11
10. Content Delivery Networks ...................................... 12
11. Business Partners ............................................... 12
12. Operations and Management ...................................... 13
13. Security Considerations ........................................ 13
14. IANA Considerations ............................................. 15
15. Acknowledgements ............................................... 15
16. Change log [RFC Editor: Please remove] ....................... 15
17. References ..................................................... 15
   17.1. Normative References ..................................... 15
   17.2. Informative References ................................... 16
Authors’ Addresses .................................................. 18
1. Introduction

The deployment of IPv6 [RFC2460] is now in progress, and users with no IPv4 access are likely to appear in increasing numbers in the coming years. Any provider of content or application services over the Internet will need to arrange for IPv6 access or else risk losing large numbers of potential customers. The time for action is now, while the number of such customers is small, so that appropriate skills, software and equipment can be acquired in good time to scale up the IPv6 service as demand increases. An additional advantage of early support for IPv6 customers is that it will reduce the number of customers connecting later via IPv4 "extension" solutions such as double NAT, which will otherwise degrade the user experience.

Nevertheless, it is important that the introduction of IPv6 service should not make service for IPv4 customers worse. In some circumstances, technologies intended to assist in the transition from IPv4 to IPv6 are known to have negative effects on the user experience. A deployment strategy for IPv6 must avoid these effects as much as possible.

The purpose of this document is to provide guidance and suggestions for Internet Content Providers (ICPs) and Application Service Providers (ASPs) who wish to offer their services to both IPv6 and IPv4 customers. For simplicity, the term ICP is mainly used in the body of this document, but the guidance also applies to ASPs. Many of the points in this document will also apply to enterprise networks that do not classify themselves as ICPs. Any enterprise or department that runs at least one externally accessible server, such as an HTTP server, may also be concerned. Although specific managerial and technical approaches are described, this is not a rule book; each operator will need to make its own plan, tailored to its own services and customers.

2. General Strategy

The most important advice here is to actually have a general strategy. Adding support for a second network layer protocol is a new departure for most modern organisations, and it cannot be done casually on a day-by-day basis. Even if it is impossible to write a precisely dated plan, the intended steps in the process need to be defined well in advance. There is no single blueprint for this. The rest of this document is meant to provide a set of topics to be taken into account in defining the strategy.

In determining the urgency of this strategy, it should be noted that the central IPv4 registry (IANA) ran out of spare blocks of IPv4
addresses in February 2011 and the various regional registries are expected to exhaust their reserves over the next one to two years. After this, Internet Service Providers (ISPs) will run out at dates determined by their own customer base. No precise date can be given for when IPv6-only customers will appear in commercially significant numbers, but - particularly in the case of mobile users - it may be quite soon. Complacency about this is therefore not an option for any ICP that wishes to grow its customer base over the coming years.

The most common strategy for an ICP is to provide dual stack services - both IPv4 and IPv6 on an equal basis - to cover both existing and future customers. This is the recommended strategy in [RFC6180] for straightforward situations. Some ICPs who already have satisfactory operational experience with IPv6 might consider an IPv6-only strategy, with IPv4 clients being supported by translation or proxy at their ISP border. However, the present document is addressed to ICPs without IPv6 experience, who are likely to prefer the dual stack model to build on their existing IPv4 service.

Within the dual stack model, two approaches could be adopted, sometimes referred to as "outside in" and "inside out":

- Outside in: start by providing external users with an IPv6 public access to your services, for example by running a reverse proxy that handles IPv6 customers (see Section 7 for details). Progressively enable IPv6 internally.
- Inside out: start by enabling internal networking infrastructure, hosts, and applications to support IPv6. Progressively reveal IPv6 access to external customers.

Which of these approaches to adopt depends on the precise circumstances of the ICP concerned. "Outside in" has the benefit of giving interested customers IPv6 access at an early stage, and thereby gaining precious operational experience, before meticulously updating every piece of equipment and software. For example, if some back-office system, that is never exposed to users, only supports IPv4, it will not cause delay. "Inside out" has the benefit of completing the implementation of IPv6 as a single project. Any ICP could choose this approach, but it might be most appropriate for a small ICP without complex back-end systems.

A point that must be considered in the strategy is that some customers will remain IPv4-only for many years, others will have both IPv4 and IPv6 access, and yet others will have only IPv6. Additionally, mobile customers may find themselves switching between IPv4 and IPv6 access as they travel, even within a single session. Services and applications must be able to deal with this, just as easily as they deal today with a user whose IPv4 address changes (see
Nevertheless, the end goal is to have a network that does not need major changes when at some point in the future it becomes possible to transition to IPv6-only, even if only for some parts of the network. That is, the IPv6 deployment should be designed in such a way as to more or less assume that IPv4 is absent, so the network will function seamlessly when it is indeed no longer there.

An important first step in every strategy is to determine from every hardware and software supplier details of their planned dates for providing full IPv6 support, with performance equivalent to IPv4, in their products and services.

3. Education and Skills

Some older staff may have experience of running multiprotocol networks, which were common twenty years ago before the dominance of IPv4. However, IPv6 will be new to them, and also to younger staff brought up on TCP/IP. It is not enough to have one "IPv6 expert" in a team. On the contrary, everybody who knows about IPv4 needs to know about IPv6, from network architect to help desk responder. Therefore, an early and essential part of the strategy must be education, including practical training, so that all staff acquire a general understanding of IPv6, how it affects basic features such as the DNS, and the relevant practical skills. To take a trivial example, any staff used to dotted-decimal IPv4 addresses need to become familiar with the colon-hexadecimal format used for IPv6.

There is an anecdote of one IPv6 deployment in which prefixes including the letters A to F were avoided by design, to avoid confusing sysadmins unfamiliar with hexadecimal notation. This is not a desirable result. There is another anecdote of a help desk responder telling a customer to "disable one-Pv6" in order to solve a problem. It should be a goal to avoid having untrained staff who don’t understand hexadecimal or who can’t even spell "IPv6".

It is very useful to have a small laboratory network available for training and self-training in IPv6, where staff may experiment and make mistakes without disturbing the operational IPv4 service. This lab should run both IPv4 and IPv6, to gain experience with a dual-stack environment and new features such as having multiple addresses per interface.

A final remark about training is that it should not be given too soon, or it will be forgotten. Training has a definite need to be done "just in time" in order to properly "stick." Training, lab
experience, and actual deployment should therefore follow each other immediately. If possible, training should even be combined with actual operational experience.

4. Arranging IPv6 Connectivity

There are, in theory, two ways to obtain IPv6 connectivity to the Internet.

- Native. In this case the ISP simply provides IPv6 on exactly the same basis as IPv4 - it will appear at the ICP’s border router(s), which must then be configured in dual-stack mode to forward IPv6 packets in both directions. This is by far the better method. An ICP should contact all its ISPs to verify when they will provide native IPv6 support, whether this has any financial implications, and whether the same service level agreement will apply as for IPv4. Any ISP that has no definite plan to offer native IPv6 service should be avoided.

- Tunnel. It is possible to configure an IPv6-in-IPv4 tunnel to a remote ISP that offers such a service. A dual-stack router in the ICP’s network will act as a tunnel end-point, or this function could be included in the ICP’s border router.

A tunnel is a reasonable way to obtain IPv6 connectivity for initial testing and skills acquisition. However, it introduces an inevitable extra latency compared to native IPv6, giving users a noticeably worse response time for complex web pages. It is also likely to limit the IPv6 MTU size. In normal circumstances, native IPv6 will provide an MTU size of at least 1500 bytes, but it will almost inevitably be less for a tunnel, possibly as low as 1280 bytes (the minimum MTU allowed for IPv6). Apart from the resulting loss of efficiency, there are cases in which Path MTU Discovery fails, therefore IPv6 fragmentation fails, and in this case the lower tunnel MTU will actually cause connectivity failures for customers.

For these reasons, ICPs are strongly recommended to obtain native IPv6 service before attempting to offer a production-quality service to their users.

5. IPv6 Infrastructure

5.1. Address and subnet assignment

An ICP must first decide whether to apply for its own Provider Independent (PI) address prefix for IPv6. The default is to obtain a
Provider Aggregated (PA) prefix from each of its ISPs, and operate them in parallel. Both solutions are viable in IPv6. However, scaling properties of the wide area routing system (BGP4) limit the routing of PI prefixes, so only large content providers can justify the bother and expense of obtaining a PI prefix and convincing their ISPs to route it. Millions of enterprise networks, including smaller content providers, will use PA prefixes. In this case, a change of ISP would necessitate a change of the corresponding PA prefix, using the procedure outlined in [RFC4192].

An ICP that has multiple connections via multiple ISPs will have multiple PA prefixes. This results in multiple PA-based addresses for the servers, or for load balancers if they are in use.

An ICP may also choose to operate a Unique Local Address prefix [RFC4193] for internal traffic only, as described in [RFC4864].

Depending on its projected future size, an ICP might choose to obtain /48 PI or PA prefixes (allowing 16 bits of subnet address) or longer PA prefixes, e.g. /56 (allowing 8 bits of subnet address). Clearly the choice of /48 is more future-proof. Advice on the numbering of subnets may be found in [RFC5375].

Since IPv6 provides for operating multiple prefixes simultaneously, it is important to check that all relevant tools, such as address management packages, can deal with this. In particular, the need to allow for multiple PA prefixes with IPv6, and the possible need to renumber, means that using manually assigned static addresses for servers is problematic [I-D.carpenter-6renum-static-problem].

Theoretically, it would be possible to operate an ICP’s IPv6 network using only Stateless Address Autoconfiguration [RFC4862]. In practice, an ICP of reasonable size will probably choose to operate DHCPv6 [RFC3315] and use it to support stateful and/or on-demand address assignment.

5.2. Routing

In a dual stack network, IPv4 and IPv6 routing protocols operate quite independently and in parallel. The common routing protocols all exist in IPv6 versions, such as OSPFv3 [RFC5340], IS-IS [RFC5308], and even RIPng [RFC2080] [RFC2081]. For trained staff, there should be no particular difficulty in deploying IPv6 routing without disturbance to IPv4 services.

The performance impact of dual stack routing needs to be evaluated. In particular, what performance does the router vendor claim for IPv6? If the performance is significantly inferior compared to IPv4,
will this be an operational problem? To answer this question, the ICP will need a projected model for the amount of IPv6 traffic expected initially, and its likely rate of increase.  [[Note: further input from the WG is needed on this point.]]

If a site operates multiple PA prefixes as mentioned in Section 5.1, complexities may appear in routing configuration. In particular, source-based routing rules may be needed to ensure that outgoing packets are routed to the appropriate border router and ISP link. Normally, a packet sourced from an address assigned by ISP X should not be sent via ISP Y, to avoid ingress filtering by Y [RFC2827] [RFC3704]. Additional considerations may be found in [I-D.ietf-v6ops-ipv6-multihoming-without-ipv6nat].

Each IPv6 subnet normally has a /64 prefix, leaving another 64 bits for the interface identifiers of individual hosts. In contrast, a typical IPv4 subnet will have no more than 8 bits for the host identifier, thus limiting the subnet to 256 or fewer hosts. A dual stack design will typically use the same subnet topology for IPv4 and IPv6, and therefore the same router topology. This means that the limited subnet size of IPv4 will be imposed on IPv6. It would be theoretically possible to avoid this limitation by implementing a different subnet and router topology for IPv6, for example by ingenious use of VLANs. This is not advisable, as it would result in extremely complex fault diagnosis when something went wrong.

5.3. DNS

This is largely a case of "just do it." Each externally visible host (or virtual host) that has an A record for its IPv4 address needs an AAAA record [RFC3596] for its IPv6 address, and a reverse entry if applicable. One important detail is that some clients (especially Windows XP) can only resolve DNS names via IPv4, even if they can use IPv6 for application traffic. It is therefore advisable for all DNS servers to respond to queries via both IPv4 and IPv6.

6. Load Balancers

It is to be expected that IPv6 traffic will initially be low, i.e. a small percentage of IPv4 traffic. For this reason, updating load balancers to fully support IPv6 can perhaps be delayed; however, such an update needs to be planned in anticipation of significant growth over a period of several years. The same would apply to TLS or HTTP proxies used for load balancing purposes. It is important to obtain appropriate assurances from vendors about their IPv6 support, including performance aspects (as discussed for routers in Section 5.2).
7. Proxies

An HTTP proxy [RFC2616] can readily be configured to handle incoming connections over IPv6 and to proxy them to a server over IPv4. Therefore, a single proxy can be used as the first step in an outside-in strategy, as shown in the following diagram:

```
IPv6 Clients in the Internet
-------------
| Ingress router |
-------------

-------------
| IPv6 stack |
| HTTP proxy |
| IPv4 stack |
-------------

-------------
| IPv4 stack |
| HTTP server |
-------------
```

In this case, the AAAA record for the service would provide the IPv6 address of the proxy. This approach will work for any HTTP or HTTPS applications that operate successfully via a proxy, as long as IPv6 load remains low.

8. Servers

8.1. Network Stack

The TCP/IP network stacks in popular operating systems have supported IPv6 for many years. In most cases, it is sufficient to enable IPv6 and possibly DHCPv6; the rest will follow. Servers inside an ICP
network will not need to support any transition technologies beyond a simple dual stack, with a possible exception for 6to4 mitigation noted below in Section 9.

8.2. Application Layer

Basic HTTP servers have been able to handle an IPv6-enabled network stack for some years, so at the most it will be necessary to update to a more recent software version. The same is true of generic applications such as email protocols. No general statement can be made about other applications, especially proprietary ones, so each ASP will need to make its own determination.

One important recommendation here is that all applications should use domain names, which are IP-version-independent, rather than IP addresses. Applications based on middleware platforms which have uniform support for IPv4 and IPv6, for example Java, may be able to support both IPv4 and IPv6 naturally without additional work.

A specific issue for HTTP-based services is that IP address-based cookie authentication schemes will need to deal with dual-stack clients. Servers might create a cookie for an IPv4 connection or an IPv6 connection, depending on the setup at the client site and on the whims of the client operating system. There is no guarantee that a given client will consistently use the same address family, especially when accessing a collection of sites rather than a single site. If the client is using privacy addresses [RFC4941], the IPv6 address (but not its /64 prefix) might change quite frequently. Any cookie mechanism based on 32-bit IPv4 addresses will need significant remodelling.

Generic considerations on application transition are discussed in [RFC4038], but many of them will not apply to the dual-stack ICP scenario. An ICP that creates and maintains its own applications will need to review them for any dependency on IPv4.

8.3. Geolocation

As time goes on, it is to be assumed that geolocation methods and databases will be updated to fully support IPv6 prefixes. There is no reason they will be more or less accurate in the long term than those available for IPv4. However, we can expect many more clients to be mobile as time goes on, so geolocation based on IP addresses alone may become problematic. Initially, at least, ICPs may observe some weakness in geolocation for IPv6 clients.
9. Coping with Transition Technologies

As mentioned above, an ICP should obtain native IPv6 connectivity from its ISPs. In this way, the ICP can avoid most of the complexities of the numerous IPv4-to-IPv6 transition technologies that have been developed; they are all second-best solutions. However, some clients are sure to be using such technologies. An ICP needs to be aware of the operational issues this may cause and how to deal with them.

In some cases outside the ICP’s control, clients might reach a content server via a network-layer translator from IPv6 to IPv4. ICPs who are offering a dual stack service and providing both A and AAAA records, as recommended in this document, should not normally receive traffic from NAT64 translators [RFC6146]. Exceptionally, however, such traffic could arrive via IPv4 from an IPv6-only client whose DNS resolver failed to receive the ICP’s AAAA record for some reason. Such traffic would be indistinguishable from regular IPv4-via-NAT traffic.

Alternatively, ICPs who are offering a dual stack service might exceptionally receive IPv6 traffic translated from an IPv4-only client that somehow failed to receive the ICP’s A record. An ICP could also receive IPv6 traffic with translated prefixes [RFC6296]. These two cases would only be an issue if the ICP was offering any service that depends on the assumption of end-to-end IPv6 address transparency.

In other cases, also outside the ICP’s control, IPv6 clients may reach the IPv6 Internet via some form of IPv6-in-IPv4 tunnel. In this case a variety of problems can arise, the most acute of which affect clients connected using the Anycast 6to4 solution [RFC3068]. Advice on how ICPs may mitigate these 6to4 problems is given in Section 4.5. of [RFC6343]. For the benefit of all tunnelled clients, it is essential to verify that Path MTU Discovery works correctly (i.e., the relevant ICMPv6 packets are not blocked) and that the server-side TCP implementation correctly supports the Maximum Segment Size (MSS) negotiation mechanism [RFC2923] for IPv6 traffic.

Some ICPs have implemented an interim solution to mitigate transition problems by limiting the visibility of their AAAA records to users with validated IPv6 connectivity [I-D.ietf-v6ops-v6-aaaa-whitelisting-implications].

Another approach taken by some ICPs is to offer IPv6-only support via a specific DNS name, e.g., ipv6.example.com, if the primary service is www.example.com. In this case ipv6.example.com would have an AAAA record only. This has some value for testing purposes, but is
otherwise only of interest to hobbyist users willing to type in special URLs.

There is little an ICP can do to deal with client-side or remote ISP deficiencies in IPv6 support, but it is hoped that the "happy eyeballs" [I-D.ietf-v6ops-happy-eyeballs] approach will improve the ability for clients to deal with such problems.

10. Content Delivery Networks

DNS-based techniques for diverting users to Content Delivery Network (CDN) points of presence (POPs) will work for IPv6, if AAAA records are provided as well as A records. In general the CDN should follow the recommendations of this document, especially by operating a full dual stack service at each POP. Additionally, each POP will need to handle IPv6 routing exactly like IPv4, for example running BGP4+ [RFC4760] if appropriate.

Note that if an ICP supports IPv6 but its CDN does not, its clients will continue to use IPv4 and any IPv6-only clients will have to use a transition solution of some kind. This is not a desirable situation, since the ICP’s work to support IPv6 will be wasted. The converse is not true: if the CDN supports IPv6 but the ICP does not, dual-stack and IPv6-only clients will obtain IPv6 access.

An ICP might face a complex situation, if its CDN provider supports IPv6 at some POPs but not at others. IPv6-only clients could only be diverted to a POP supporting IPv6. There are also scenarios where a dual-stack client would be diverted to a mixture of IPv4 and IPv6 POPs for different URLs, according to the A and AAAA records provided and the availability of optimisations such as "happy eyeballs." These complications do not affect the viability of relying on a dual-stack CDN, however.

The CDN itself faces related complexity: "As IPv6 rolls out, it’s going to roll out in pockets, and that’s going to make the routing around congestion points that much more important but also that much harder," stated John Summers of Akamai in 2010.

11. Business Partners

As noted earlier, it is in an ICP’s or ASP’s best interests that their users have direct IPv6 connectivity, rather than indirect IPv4 connectivity via double NAT. If the ICP or ASP has a direct business relationship with some of their clients, or with the networks that connect them to their clients, they are advised to coordinate with
those partners to ensure that they have a plan to enable IPv6. They should also verify and test that there is first-class IPv6 connectivity end-to-end between the networks concerned. This is especially true for implementations that require IPv6 support in specialized programs or systems in order for the IPv6 support on the ICP/ASP side to be useful.

12. Operations and Management

There is no doubt that, initially, IPv6 deployment will have operational impact, as well as requiring education and training as mentioned in Section 3. Staff will have to update network elements such as routers, update configurations, provide information to end users, and diagnose new problems. However, for an enterprise network, there is plenty of experience, e.g. on numerous university campuses, showing that dual stack operation is no harder than IPv4-only in the steady state.

Whatever management, monitoring and logging is performed for IPv4 is also needed for IPv6. Therefore, all products and tools used for these purposes must be updated to fully support IPv6. Note that since an IPv6 network may operate with more than one IPv6 prefix and therefore more than one address per host, the tools must deal with this as a normal situation. This includes any address management tool in use (see Section 5.1) as well as tools used for creating DHCP and DNS configurations. There is significant overlap here with the tools involved in site renumbering [I-D.jiang-6renum-enterprise].

As far as possible, however, mutual dependency between IPv4 and IPv6 operations should be avoided. A failure of one should not cause a failure of the other. One precaution to avoid this would be for back-end systems such as network management databases to be dual stacked as soon as convenient. It should also be possible to use IPv4 connectivity to repair IPv6 configurations, and vice versa.

Dual stack, while necessary, does have management scaling and overhead considerations. As noted earlier, the long term goal is to move to single-stack IPv6, when the network and its customers can support this. This is an additional reason why mutual dependency between the address families should be avoided in the management system in particular; a hidden dependency on IPv4 that had been forgotten for many years would be highly inconvenient.

13. Security Considerations

Essentially every threat that exists for IPv4 exists or will exist
for IPv6. Therefore, it is essential to update firewalls, intrusion
detection systems, denial of service precautions, and security
auditing technology to fully support IPv6. Otherwise, IPv6 will
become an attractive target for attackers.

When multiple PA prefixes are in use as mentioned in Section 5.1,
firewall rules must allow for all valid prefixes, and must be set up
to work as intended even if packets are sent via one ISP but return
packets arrive via another.

Performance aspects of dual stack firewalls must be considered (as
discussed for routers in Section 5.2).

In a dual stack operation, there may be a risk of cross-contamination
between the two protocols. For example, a successful IPv4-based
denial of service attack might also deplete resources needed by the
IPv6 service, or vice versa. This risk strengthens the argument that
IPv6 security must be up to the same level as IPv4.

A general overview of techniques to protect an IPv6 network against
external attack is given in [RFC4864]. Assuming an ICP has native
IPv6 connectivity, it is advisable to block incoming IPv6-in-IPv4
tunnel traffic using IPv4 protocol type 41. Outgoing traffic of this
kind should be blocked except for the case noted in Section 4.5 of
[RFC6343]. ICMPv6 traffic should only be blocked in accordance with
[RFC4890]; in particular, Packet Too Big messages, which are
essential for PMTU discovery, must not be blocked.

Scanning attacks to discover the existence of hosts are much less
likely to succeed for IPv6 than for IPv4 [RFC5157]. However, this is
only true if IPv6 hosts are configured with interface identifiers
that are hard to guess; for example, it is not advisable to manually
configure servers with static interface identifiers starting from
"1".

Transport Layer Security version 1.2 [RFC5246] and its predecessors
work correctly with TCP over IPv6, meaning that HTTPS-based security
solutions are immediately applicable. The same should apply to any
other transport-layer or application-layer security techniques.

If an ASP uses IPsec [RFC4301] and IKE [RFC5996] in any way to secure
connections with clients, these too are fully applicable to IPv6, but
only if the software stack at each end has been appropriately
updated.
14. IANA Considerations

This document requests no action by IANA.

15. Acknowledgements

Valuable contributions were made by Erik Kline. Useful comments were received from Tassos Chatzithomaoglou, Wesley George, Victor Kuarsingh, Bing Liu, John Mann, and other participants in the V6OPS working group.

This document was produced using the xml2rfc tool [RFC2629].

16. Change log [RFC Editor: Please remove]

draft-carpenter-v6ops-icp-guidance-03: additional WG comments, 2012-02-23.
draft-carpenter-v6ops-icp-guidance-02: additional WG comments, 2012-01-07.
draft-carpenter-v6ops-icp-guidance-01: multiple clarifications after WG comments, 2011-12-06.
draft-carpenter-v6ops-icp-guidance-00: original version, 2011-10-22.

17. References

17.1. Normative References


17.2. Informative References


[I-D.ietf-v6ops-ipv6-multihoming-without-ipv6nat]
Troan, O., Miles, D., Matsushima, S., Okimoto, T., and D. Wing, "IPv6 Multihoming without Network Address Translation",
draft-ietf-v6ops-ipv6-multihoming-without-ipv6nat-04 (work in progress), February 2012.

[I-D.ietf-v6ops-v6-aaaa-whitelisting-implications]
Livingood, J., "Considerations for Transitioning Content to IPv6",
draft-ietf-v6ops-v6-aaaa-whitelisting-implications-09 (work in progress), February 2012.

[I-D.jiang-6renum-enterprise]
draft-jiang-6renum-enterprise-02 (work in progress), December 2011.


[RFC2923] Lahey, K., "TCP Problems with Path MTU Discovery",
RFC 2923, September 2000.


Authors’ Addresses

Brian Carpenter
Department of Computer Science
University of Auckland
PB 92019
Auckland, 1142
New Zealand

Email: brian.e.carpenter@gmail.com

Sheng Jiang
Huawei Technologies Co., Ltd
Q14, Huawei Campus
No.156 Beiqing Road
Hai-Dian District, Beijing 100095
P.R. China

Email: jiangsheng@huawei.com
Abstract

This document describes how the IPv6 flow label can be used in support of layer 3/4 load distribution and balancing for large server farms.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on September 7, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as
Table of Contents

1. Introduction ......................................................... 3
2. Role of the Flow Label ............................................. 5
3. Possible extended role .............................................. 8
4. Security Considerations ........................................... 9
5. IANA Considerations ................................................ 11
6. Acknowledgements .................................................. 11
7. Change log [RFC Editor: Please remove] ........................ 11
8. References .......................................................... 11
   8.1. Normative References ........................................ 11
   8.2. Informative References ..................................... 11
Authors’ Addresses .................................................... 12
1. Introduction

The IPv6 flow label has been redefined [RFC6437] and its use for load sharing in multipath routing has been specified [RFC6438]. Another scenario in which the flow label could be used is in load distribution for large server farms. Load distribution is a slightly more general term than load balancing, but the latter is more commonly used. This document starts with a brief introduction to load balancing techniques and then describes how the flow label might be used to enhance layer 3/4 flow balancers in particular.

Load balancing for server farms is achieved by a variety of methods, often used in combination [Tarreau]. The flow label is not relevant to all of them. The actual load balancing algorithm (the choice of server for a new client session) is irrelevant to this discussion.

- The simplest method is simply using the DNS to return different server addresses for a single name such as www.example.com to different users. Typically this is done by rotating the order in which different addresses are listed by the relevant authoritative DNS server, assuming that the client will pick the first one. Routing may be configured such that the different addresses are handled by different ingress routers. The flow label can have no impact on this method and it is not discussed further.

- Another method, for HTTP servers, is to operate a layer 7 reverse proxy in front of the server farm. The reverse proxy will present a single IP address to the world, communicated to clients by a single AAAA record. For each new client session (an incoming TCP connection and HTTP request), it will pick a particular server and proxy the session to it. Hopefully the act of proxying will be cheap compared to the act of serving the required content. The proxy must retain TCP state and proxy state for the duration of the session. This TCP state could, potentially, include the incoming flow label value.

- A component of some load balancing systems is an SSL reverse proxy farm. The individual SSL proxies handle all cryptographic aspects and exchange raw HTTP with the actual servers. Thus, from the load balancing point of view, this really looks just like a server farm, except that it’s specialised for HTTPS. Each proxy will retain SSL and TCP and maybe HTTP state for the duration of the session, and the TCP state could potentially include the flow label.

- Finally the "front end" of many load balancing systems is a layer 3/4 load balancer. While it can sometimes be a dedicated hardware, it also happens to be a standard function of some network switches or routers (eg: using ECMP, [RFC2991]). In this case, it is the layer 3/4 load balancer whose IP address is published as the primary AAAA record for the service. All client...
sessions will pass through this device. According to the precise scenario, it will spread new sessions across the actual application servers, across an SSL proxy farm, or across a set of layer 7 proxies. In all cases, the layer 3/4 load balancer has to recognize incoming packets as belonging to new or existing client sessions, and choose the target server or proxy so as to ensure persistence. ‘Persistence’ is defined as guaranteeing that a given session will run to completion on a single server. The layer 3/4 load balancer therefore needs to inspect each incoming packet to identify the session. There are two common types of layer 3/4 load balancers, the totally stateless ones which only act on packets, generally involving a per-packet hashing of easy-to-find information such as the source address and/or port into a server number, and the stateful ones which take the routing decision on the very first packets of a session and maintain the same direction for all packets belonging to the same session. Clearly, both types of layer 3/4 balancers could inspect and make use of the flow label value.

Our focus is on how the balancer identifies a particular flow. For clarity, note that two aspects of layer 3/4 load balancers are not affected at all by use of the flow label to identify sessions.

1. Balancers use various techniques to redirect traffic to a specific target server.

   - All servers are configured with the same IP address, they are all on the same LAN, and the load balancer sends directly to their individual MAC addresses.
   - Each server has its own IP address, and the balancer uses an IP-in-IP tunnel to reach it.
   - Each server has its own IP address, and the balancer performs NAPT (network address and port translation) to deliver the client’s packets to that address.

   The choice between these methods is not affected by use of the flow label.

2. A layer 3/4 balancer must correctly handle Path MTU Discovery by forwarding relevant ICMPv6 packets in both directions. This too is not affected by use of the flow label.

   The following diagram, inspired by [Tarreau], shows a maximum layout.
From the previous paragraphs, we can identify several points in this diagram where the flow label might be relevant:

1. Layer 3/4 load balancers.
2. SSL proxies.
3. HTTP proxies.

2. Role of the Flow Label

The IPv6 flow label is a 20 bit field included in every IPv6 header [RFC2460] and it is defined in [RFC6437]. According to this definition, it should be set to a constant value for a given traffic flow (such as an HTTP connection), but until the standard is widely implemented it will often be set to the default value of zero. Any device that has access to the IPv6 header has access to the flow...
label, and it is at a fixed position in every IPv6 packet. In contrast, transport layer information, such as the port numbers, is not always in a fixed position, since it follows any IPv6 extension headers that may be present. Therefore, within the lifetime of a given transport layer connection, the flow label can be a more convenient "handle" than the port number for identifying that particular connection.

According to [RFC6437], source hosts should set the flow label, but if they do not (i.e. its value is zero), forwarding nodes may do so instead. In both cases, the flow label value must be constant for a given transport session, normally identified by the IPv6 and Transport header 5-tuple. The flow label should be calculated by a stateless algorithm. The value should form part of a statistically uniform distribution, making it suitable as part of a hash function used for load distribution. Because of using a stateless algorithm to calculate the label, there is a very low (but non-zero) probability that two simultaneous flows from the same source to the same destination have the same flow label value despite having different transport protocol port numbers.

A careful reading of RFC 6437 shows that for a given source accessing a well-known TCP port at a given destination, the flow label is in effect a proxy for the source port number, found at a fixed position in the layer 3 header. Thus, the suggested model for using the flow label in a load balancing mechanism is as follows:

- It is clearly better if the original source, e.g. an HTTP client, sets the flow label. However, if the flow label of an incoming packet is zero, there are two possibilities:
  - The ingress router at the server site could implement the stateless mechanism in Section 3 of [RFC6437] to set the flow label value to an appropriate value. This relieves the subsequent load balancers of the need to fully analyse the IPv6 and Transport header 5-tuple to identify the packets belonging to the same flow.
  - Load balancers will use the flow label value as described below if it is set, but use the transport header in the traditional way otherwise.

In either case, the idea is that as the use of the flow label becomes more prevalent, load balancers will reap a growing performance benefit.

- The layer 3/4 load balancers can use the 2-tuple {source address, flow label} as the session key for whatever load distribution algorithm they support, instead of searching for the transport port number later in the header. Note that they do not need to consider the destination address as it is always the same, i.e., the server address.
Stateless layer 3/4 load balancers would simply apply a hash algorithm to the 2-tuple \{source address, flow label\} on all packets, while stateful load balancers would apply their usual load distribution algorithm to the first packet of a session, and store the \{2-tuple, server\} association in a table so that all packets belonging to the same session are forwarded to the same server. However, for all subsequent packets of the session, it can ignore all IPv6 extension headers, which should lead to a performance benefit. Whether this benefit is valuable will depend on engineering details of the specific load balancer.

Layer 3/4 balancers that redirect the incoming packets by NAPT are not expected to obtain any saving of time by using the flow label, because they must in any case follow the extension header chain in order to locate and modify the port number and transport checksum. The same would apply to balancers that perform TCP state tracking for any reason.

- Note that correct handling of ICMPv6 for Path MTU Discovery requires the layer 3/4 balancer to keep state for the client source address, independently of either the port numbers or the flow label.
- An SSL proxy should forward the flow identifier between the ciphered side and the clear side. Being able to forward data used for persistence is very important, as it’s the only way to stack multiple layers of network components without losing information.
- The HTTP proxies may do the same. However, since they have to process the transport and application layers in any case, this might not lead to any performance benefit.

Note that in the unlikely event of two simultaneous flows from the same source having the same flow label value, the two flows would end up assigned to the same server, where they would be distinguished as normal by their port numbers. Since this would be a statistically rare event, it would not damage the overall load balancing effect. Moreover, it is very likely that there will be many more servers than possible flow label values at most locations (1 million possible values), so it is already expected that many different flow label values will end up on the same server for a given IP address. In the case where many thousands of clients are hidden behind the same large-scale NAT with a single IP address, the assumption of low probability of conflicts might become incorrect unless flow label values are random enough to avoid following similar sequences for all clients. This is not expected to be a factor for IPv6 anyway, since there is no valid reason to implement NAT [RFC4864]. The statistical assumption is valid for sites that implement network prefix translation [RFC6296], since this technique provides a different address for each client.
3. Possible extended role

A particular aspect of the session persistence issue is when multiple independent transport connections from the same client need to be handled by the same server instance. This can be an extremely difficult task which often requires ugly tricks such as pattern matching within a buffered stream, cookie insertion, etc, which most load balancers have to deal with every day. If the client application has control over the outgoing flow label, then it can itself assign the same label to all transport connections related to a single application session.

A common example is FTP. For a load balancer, passive-mode FTP requires parsing the entire control stream (port 21), in order to find which incoming packet will initiate a data session on a port chosen by the server. This does not always work well due to the fact that sometimes clients don’t connect, or that the session is finally not used (e.g., because no transfer needs to be performed).

Using a flow label, the client could generate an initial random flow identifier when a file transfer is expected, and assign the same flow label to all data connections related to the same control connection. A flow label based load balancer would then by definition send the data traffic to the same server as the control traffic, and would thus guarantee that the sessions are properly associated. Such a mechanism is permitted by [RFC6437], although it is not the recommended default.

The same need is even more prominent with HTTP/HTTPS : while it is costly but not difficult to insert a cookie in an HTTP stream to identify the server the user was assigned to, it is very difficult to do that for HTTPS, because the stream must be deciphered first. Deciphering the stream requires a huge amount of centralized power, since the load balancer needs to see the clear stream; this is in fact the main reason for SSL proxies in load balancing scenarios. If a web client (browser) used the same flow label for any protocol targeting a given host (or domain), this could be used by load balancers to reach the same server for both HTTP and HTTPS, without having to open the stream payload at all nor to inspect anything beyond layer 3, which clearly is not possible today.

An additional complication that can arise is when a single client inadvertently generates sessions that appear to originate from different IP addresses. This can arise, for example, if an enterprise uses a proxy farm for outgoing traffic, or in mobile applications where several subsequent requests come from different network cells thus different IP addresses (for instance, consulting banking account in the train). When two consecutive client requests
pass through two distinct proxies, a different IP source address may be presented to the server load balancer, which then cannot rely on address-based persistence. It would be possible and desirable in principle to use the same flow label value for correlated sessions from the same client, if the proxies were transparent to the flow label value.

In some application scenarios, an inadvertent change in the client IP address may have only minor consequences, such as reloading transaction context into a new server. In other cases it may be more serious and result in a transaction failure. For this reason, a reliable solution in which the load balancer would use the flow label value on its own would be advantageous.

Using the flow label in this way would also greatly simplify the logging of user sessions. A very common task is to match logs from various equipments to follow a user’s activity and decide whether it indicates a bug, user error or attack. Logging a flow label would of course help because it’s easier to find the beginning and end of a session and decide whether it’s legitimate or not.

Such extensions to the role of the flow label in load balancing are theoretically very attractive, but would require a major refresh of client software as well as of load balancers themselves. It amounts to considering an entire application session, in a broad sense, as a single flow for the purposes of RFC 6437.

It is worth noting though that what is important to save server-side resources is wide enough adoption. Most of today's load balanced traffic is HTTP originating from a handful of browsers which are regularly upgraded for security considerations. Once a mechanism is adopted, it can quickly be deployed and become the general case.

The difficulty of the upgrade path is then on the server side. The first step would consist in having layer 7 load balancers be able to consider the flow label to avoid costly layer 7 analysis each time it is possible. This means that if a non-null flow label is seen, then the load balancer would consider it, otherwise it would fall back to its default behaviour. The second step would consist in having front layer 3/4 load balancers bypass the layer 7 load balancer farms when the flow label is found. This point would greatly offload layer 7 load balancers.

4. Security Considerations

Security aspects of the flow label are discussed in [RFC6437]. As noted there, a malicious source or man-in-the-middle could disturb...
load balancing by manipulating flow labels. This risk already exists today where the source address and port are used as hashing key in layer 3/4 load balancers, as well as where a persistence cookies is used in HTTP to designate a server. It even exists on layer 3 components which only rely on the source address to select a destination, making them more DDoS-prone, still all these methods are currently used because the benefits for load balancing and persistence hugely outweigh the risks.

Specifically, [RFC6437] states that "stateless classifiers should not use the flow label alone to control load distribution, and stateful classifiers should include explicit methods to detect and ignore suspect flow label values." The former point is answered by also using the source address. The latter point is more complex. If the risk is considered serious, the ingress router mentioned above should verify incoming flows with non-zero flow label values. If a flow from a given source address and port number does not have a constant flow label value, it is suspect and should be dropped.

The suggestion in Section 3 of using the flow label on its own as a session handle is somewhat problematic. It should never be used in applications nor where any form of resource sharing is not desired. For instance, it is not conceivable that an application would identify a user session by its flow label value due to the inevitable collisions. Using the flow label on its own should only be performed where resource sharing is inevitable and desired (for instance, load balancing) and by components explicitly designed for this task, taking into account all the risks exposed here with solid protections against mis-use, and acceptable fallbacks for the remaining situations where the flow label values will not be usable.

The flow label may be of use in protecting against distributed denial of service (DDOS) attacks against servers. As noted in RFC 6437, a source should generate flow label values that are hard to predict, most likely by including a secret nonce in the hash used to generate each label. The attacker does not know the nonce and therefore has no way to invent flow labels which will all target the same server, even with knowledge of both the hash algorithm and the load balancing algorithm. Still, it is important to understand that it is always trivial to force a load balancer to stick to the same server during an attack, so the security of the whole solution must not rely on the unpredicatability of the flow label values alone, but should include defensive measures like most load balancers already have against abnormal use of source address or session cookies.

New flows are assigned to a server according to any of the usual algorithms available on the load balancer (e.g., least connections, round robin, etc.). The association between the flow label value and
the server is stored in a table (often called stick table) so that future connections using the same flow label can be sent to the same server. This method is more robust against a loss of server and also makes it harder for an attacker to target a specific server, because the association between a flow label value and a server is not known externally.

5. IANA Considerations

This document requests no action by IANA.

6. Acknowledgements

Valuable comments and contributions were made by Fred Baker, Lorenzo Colitti, Joel Jaeggli, Gurudeep Kamat, Julius Volz, and others.

This document was produced using the xml2rfc tool [RFC2629].

7. Change log [RFC Editor: Please remove]

draft-carpenter-v6ops-label-balance-02: clarified after WG discussions, 2012-03-06.


8. References

8.1. Normative References


8.2. Informative References


[RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and


Authors' Addresses

Brian Carpenter  
Department of Computer Science  
University of Auckland  
PB 92019  
Auckland, 1142  
New Zealand  
Email: brian.e.carpenter@gmail.com

Sheng Jiang  
Huawei Technologies Co., Ltd  
Q14, Huawei Campus  
No.156 Beiqing Road  
Hai-Dian District, Beijing 100095  
P.R. China  
Email: jiangsheng@huawei.com

Willy Tarreau  
Exceliance  
R&D Produits reseau  
3 rue du petit Robinson  
78350 Jouy-en-Josas  
France  
Email: w@1wt.eu
NAT64 Operational Experiences
draft-chen-v6ops-nat64-experience-03

Abstract

This document summarizes stateful NAT64 deployment scenarios and operational experience with NAT64-CGN (NAT64 Carrier Grade NATs) and NAT64-CE (NAT64 Customer Edge).

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on January 31, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect
Table of Contents

1. Introduction .................................................. 4
2. Terminology .................................................... 5
3. NAT64-CGN Deployment Experiences ............................. 6
   3.1. NAT64-CGN Networking ................................... 6
   3.2. High Availability Considerations ......................... 7
   3.3. Traceability ............................................... 7
   3.4. Quality of Experience .................................... 8
   3.5. Load Balancer ............................................. 9
   3.6. MTU Consideration ........................................ 9
4. NAT64-CE Deployment Experiences ............................... 9
   4.1. NAT64-CE Networking .................................... 10
   4.2. Anti-DDoS/SYN Flood ..................................... 11
   4.3. User Behavior Analysis .................................. 11
   4.4. DNS Resolving ........................................... 11
   4.5. Load Balancer ............................................ 12
   4.6. MTU Consideration ....................................... 12
5. Security Considerations ......................................... 12
6. IANA Considerations ........................................... 12
7. Acknowledgements ............................................... 12
8. Additional Author List ......................................... 13
9. References ..................................................... 13
   9.1. Normative References .................................... 13
   9.2. Informative References ................................. 14
Authors’ Addresses ................................................. 15
1. Introduction

Continued development of global Internet demands IP address consumption. The IANA global IPv4 address pool was exhausted on February 3, 2011. IPv6 is the only sustainable solution for numbering nodes on the Internet. Network operators have to deploy IPv6 networks in order to meet the numbering needs of the expanding internet without available IPv4 addresses. IPv4 numbering resources and IPv4-only schemes to reduce the numbering utilization during the transitional will not be adequate to maintain connectivity and deliver Internet services.

As IPv6 deployment continues, IPv6 networks and hosts will need to coexist with IPv4 numbered resources. The Internet will include nodes that are Dual-stack, nodes that remain IPv4-only and IPv6-only nodes. It may be desirable in some cases for operators to deploy a single stack network, for reasons of simplicity, cost or performance relative to a dual stack network. As IPv4 utilization eventually declines, the appeal of single stack network deployments will likely increase. In a dual-stack architecture, operators have to maintain double management interfaces, provide operational support systems for two networks, track multiple addresses in different families per host, trouble shoot host behavior related to dual stack operation and engage in other activities that increase the overhead of operating the network.

Single stack IPv6 network deployment can simplify the network provisioning. Some justification has been described in [I-D.ietf-v6ops-464xlat]. IPv6-only networks confer some benefits to mobile operators employing them. In the mobile context, it enables the use of a single IPv6 PDP(Packet Data Protocol), which eliminates significant network cost caused by doubling the PDP count on a mass of legacy mobile terminals. In broadband networks overall, it can allow for the scaling of edge-network growth decoupled from IPv4 numbering limitations.

In a transition scenario, an existing network may rely on the IPv4 stack for a long time. There is also the troublesome trend of access network providers squatting on IPv4 address space that they do not own. Allowing for interconnection between IPv4-only nodes and IPv6-only nodes is a critical capability. Widespread dual-stack deployments have not materialized at the anticipated rate over the last 10 years on possible conclusion being that legacy networks will not make the jump quickly. A translation mechanism based on a NAT64[RFC6146] function might be a key element of the internet infrastructure supporting such legacy networks.

[RFC6036] reported at least 30% operators plan to run some kind of
translator (presumably NAT64/DNS64). Advice on NAT64 deployment and operation is therefore of some importance. [RFC6586] documented the implications for IPv6 only networks. This document intends to be specific to NAT64 network planning.

In regards to IPv4/IPv6 translation, [RFC6144] has described a framework of enabling networks to make interworking possible between IPv4-only and IPv6-only networks. Three scenarios are described, "An IPv6 Network to the IPv4 Internet", "The IPv6 Internet to an IPv4 Network" and "An IPv6 Network to an IPv4 Network" where a NAT64 function is relevant. The scenario of "The IPv6 Internet to the IPv4 Internet" seems to be the ideal case for inter-network translation technology. This document has focused on the three cases and further categorized different NAT64 location and use case. The principle distinction of location is if the NAT64 is located in a NAT64-CGN (Carrier Grade Nat) or NAT64-CE (Customer Edge). NAT64-CGN corresponds to the scenario "IPv6 Network to IPv4 Internet". The NAT64-CE location roughly corresponds to the "IPv6 Internet to IPv4 Network" and "IPv6 Network to IPv4 Network" scenarios. Based on different NAT64 modes, different considerations have been described for ISPs to facilitate NAT64 deployments.

2. Terminology

The terms of NAT-CGN/CE are understood to be a topological distinction indicating different features employed in a NAT64 deployment.

NAT64-CGN: A NAT64-CGN (Carrier Grade Nat) is placed in an ISP network and managed by an administrative entity, e.g. operator. From an administrator view, a NAT64-CGN usually forwards outbound traffic into an IPv4 network. IPv6 only subscribers leverage the NAT64-CGN to be served by existing IPv4 internet services. The ISP as an administrative entity takes full control on the IPv6 side, but has limited or no control on the IPv4 side. ISP’s should attempt to accommodate the behavior of IPv4 networks and services.

NAT64-CE: A NAT64-CE (Customer Edge) is placed at the edge of customer network, e.g. a network operated by an Enterprise or Consumer. A NAT64-CE makes IPv4 services accessible for the IPv6 only users. An upstream entity and ISP usually operates an IPv4 and potentially IPv6 network respectively. IPv6 access is the common infrastructure behind the NAT64-CE.
3. NAT64-CGN Deployment Experiences

A NAT64-CGN deployment scenario is depicted in Figure 1:

---

//
//
+-----+    +-----+    +-----+
//        \       \       \        
| XLAT |     | XLAT |     DNS |     DNS |
| IPv6 Network | IPv4 | Internet | DNS | DNS64 |
|            \     |      \    /     /   |
\            +-----+    +-----+
\            //        //
\            =====>    =====>

Figure 1: NAT64-CGN Scenario: IPv6 Network to IPv4 Internet

3.1. NAT64-CGN Networking

The NAT64-CGN use case is employed to connect IPv6-only users to the IPv4 Internet. The NAT64 gateway performs protocol translation from an IPv6 packet header to an IPv4 packet header and vice versa according to the Stateful NAT64 [RFC6146]. Address translation maps IPv6 addresses to IPv4 addresses and vice versa for return traffic.

All connections to the IPv4 Internet from IPv6-only clients must traverse the NAT64-CGN. It is advantageous from the vantage-point of troubleshooting and traffic engineering to carry the IPv6 traffic natively as long as possible within an access network and translates only at or near the network egress.

In mobile networks, various possibilities can be envisaged in which to deploy the NAT64 function. Whichever option is selected, the NAT64 function will be deployed beyond the GGSN (Gateway GPRS Support Node) or PDN-GW (Public Data Network-Gateway), i.e. first IP node in currently deployed mobile architectures.

In a given implementation, NAT64 functionality can be provided by either a dedicated GW device or an multifunction gateway with integrated NAT64 functionality. In standalone NAT64, NAT64-CGN is placed to the side of a BNG or CR. An embedded NAT64 deployment would be integrated with an existing GW. Capacities of an existing GW can be potentially limited by the inserted functionality. In a mobile context, the NAT64 function can be co-located with GGSN/PDN-GW.
or it can be embedded in an existing FW/NAT44 already deployed in support of IPv4 NAT or, the function can be collocated on a router. Whatever the solution retained for the co-location option, impact on existing services and legal obligations have to be assessed.

3.2. High Availability Considerations

High Availability (HA) is a major requirement for every service and network service.

Two mechanisms are typically used to achieve high availability, i.e. cold-standby and hot-standby. Cold-standby systems have synchronized configuration and mechanism to failover traffic between the hot and cold systems such as VRRP [RFC5798]. Unlike hot-standby, cold-standby does not synchronize NAT64 session state. This makes cold-standby less resource intensive and generally simpler, but it requires clients to re-establish sessions when a fail-over occurs. Hot-standby has all the features of cold-standby but must also synchronize the binding information base (BIB). Given that short lived sessions account for most of the bindings, hot-standby does not offer much benefit for those sessions. Consideration should be given to the importance (or lack thereof) of maintaining bindings for long lived sessions across failovers.

3.3. Traceability

Traceability is required in many cases to identify an attacker or a host that launches malicious attacks and/or for various other purposes, such as accounting requirements. NAT64 devices are required to log events like creation and deletion of translations and information about the occupied resources. There are two different demands for traceability, i.e. online or offline.

- Regarding the Online requirements, XFF (X-Forwarded-For) [I-D.ietf-appsawg-http-forwarded] would be a candidate, it appends IPv6 address of subscribers to HTTP headers which is passed on to WEB servers, and the querier server can lookup radius servers for the target subscribers based on IPv6 addresses included in XFF HTTP headers. X-Forwarded-For is specific to HTTP, requires the use of an application aware gateway, cannot in general be applied to requests made over HTTPS and cannot be assumed to be preserved end-to-end as it may be overwritten by other application-aware proxies such as load balancers.

- Some potential solutions to online traceability are explore in [I-D.ietf-intarea-nat-reveal-analysis].
A NAT64-CGN could also deliver NAT64 sessions (BIB and STE) to a Radius server by extension of the radius protocol. Such an extension is an alternative solution for online traceability, particularly high performance would be required on Radius servers on order to achieve this.

For off-line traceability, syslog might be a good choice. [RFC6269] indicates address sharing solutions generally need to record and store information for specific periods of time. A stateful NAT64 is supposed to manage one mapping per session. A large volume of logs poses a challenge for storage and processing. In order to mitigate the issue, [I-D.donley-behave-deterministic-cgn] proposed to pre-allocated a group of ports for each specific IPv6 host. A trade-off among address multiplexing efficiency, port randomization security [RFC6056] and logging storage compression should be considered during the planning. A hybrid mode combining deterministic and dynamic port assignment was recommended regarding the uncertainty of user traffic.

3.4. Quality of Experience

NAT64 is providing a translation capability between IPv6 and IPv4 end-nodes. In order to provide the reachability between two IP address families, NAT64-CGN has to implement appropriate ALGs where address translation is not itself sufficient and security mechanisms do not render it infeasible. e.g. FTP-ALG [RFC6384], RSTP-ALG, H.323-ALG, etc. It should be noted that ALGs may impact the performance on a NAT64 box to some extent. ISPs as well as content providers might choose to avoid situations where the imposition of an ALG might be required. At the same time, it is also important to remind customers that IPv6 end-to-end usage does not require ALG imposition and therefore results in a better overall user experience.

The service experience should be optimized around stateful NAT processing. Session status normally is managed by a static life-cycle. In some cases, NAT resource maybe significantly consumed by largely inactive users. The NAT translator and other customers would suffer from service degradation due to port consummation by other subscribers using the same NAT64 device. A flexible NAT session control is desirable to resolve the issues. PCP[I-D.ietf-pcp-base] could be a candidate to provide such capability. A NAT64-CGN should integrate with a PCP server, to allocate available IPv4 address/Port resources. Resources could be assigned to PCP clients through PCP MAP/PEER mode. Such an ability should also be considered to upgrade user experiences, e.g. assigning different sizes of port ranges for different subscribers. Such a mechanism is also helpful to minimize terminal battery consumption reducing the number of keepalive
messages to be sent by terminal devices.

3.5. Load Balancer

Load balancers are an essential tool to avoid the issue of single points of failure and add additional scale. It is potentially important to employ load-balancing considering that deployment of multiple NAT64 devices. Load balancers are required to achieve some service continuity and scale for customers. [I-D.zhang-behave-nat64-load-balancing] discusses several ways of achieving NAT64 load balancing, including anycast based policy and prefix64 selection based policy, either implemented via DNS64[RFC6147] or Prefix64 assignments. Since DNS64 is normally co-located with NAT64 in some scenarios, it could be leveraged to perform the load balance. For traffic which does not require a DNS resolution, prefix64 assignment based on[I-D.ietf-behave-nat64-learn-analysis] could be adopted.

3.6. MTU Consideration

IPv6 requires that every link in the internet have an MTU of 1280 octets or greater[RFC2460]. However, in case of NAT64 translation deployments, some IPv4 MTU constrained link will be used in some communication path and originating IPv6 nodes may therefore receive an ICMP Packet Too Big message, reporting a Next-Hop MTU less than 1280. The result would be that IPv6 allows packets to contain a fragmentation header, without the packet being fragmented into multiple pieces. [I-D.ietf-6man-ipv6-atomic-fragments] discusses how this situation could be exploited by an attacker to perform fragmentation-based attacks, and also proposes an improved handling of such packets. It required enhancements on protocol level, which might imply potential upgrade/modifications on behaviors to deployed nodes. Another approach that potentially avoids this issue is to configure IPv4 MTU>=1260. It would forbid the occurrence of PTB<1280. However, such an operational consideration is hard to universally apply to the legacy "IPv4 Internet".

4. NAT64-CE Deployment Experiences

The NAT64-CE Scenario is depicted in Figure 2
4.1. NAT64-CE Networking

Content providers would like to use IPv6 to serve customers since it allows for the definition of new services without having to integrate consideration of IPv4 NAT and address limitations of IPv4 networks, but they have to provide some IPv4 service continuity to their customers. In some cases, customers outside the network will have IPv6-only access provided by early adopters before the internal network. The deployment requirements could be resolved by subsiding NAT64 to a customer edge, e.g. enterprise-GW. Those cases are sure to exist for the time being. An administrator of the IPv4 network needs to be cautious and aware of the operational issues this may cause, since the native IPv6 is always more desirable than transition solution.

One potential challenge in the scenario is NAT64-CE facing IPv6 Internet, in which a significant number of IPv6 users may initiate connections. When increasingly numerous users in IPv6 Internet access an IPv4 network, scalability concerns (e.g. additional latency, a single point of failure, IPv4 pool exhaustion, etc) are apt to be applied. For a given off-the-shelf NAT64-CE, those challenges should be seriously assessed. Potential issues should be properly identified. In order to mitigate the issues, it is suggested such usage should be restrained to a relative small-scale.

For operators who seek a clear precedent for operating reliable IPv6-only services, it should be well noted that the usage is problematic at several aspects. In some sense, it’s not recommended.
4.2. Anti-DDoS/SYN Flood

For every incoming new connection from the IPv6 Internet, the NAT64-CE creates state and maps that connection to an internally-facing IPv4 address and port. An attacker can consume the resources of the NAT64-CE device by sending an excessive number of connection attempts. Without a DDOS limitation mechanism, the NAT64 is exposed to attacks from the IPv6 Internet. With service provisioning, attacks have the potential could also deteriorate service quality. One consideration in internet content providers is place a L3 load balancer with capable of line rate DDOS defense, such as the employment of SYN PROXY-COOKIE. Security domain division is necessary in this case. Load Balancers could not only serve for optimization of traffic distribution, but also serve as a DOS mitigation device.

4.3. User Behavior Analysis

IP addresses are usually used as input to geo-location services. The use of address sharing will prevent these systems from resolving the location of a host based on IP address alone. Applications that assume such geographic information may not work as intended. The possible solutions listed at section 3.3 intended to bridge the gap. However, the analysis reveals those solutions can’t be a optimal substitution to solve the problem of host identification, in particular it does not today mitigate problems with source identification through translation. That makes NAT64-CE usage becoming a unappealing approach, if customers require source address tracking.

For the operators, who already deployed NAT64-CE approach, the source address of the request is obscured without the source address mapping information previously obtained. It’s superior to present mapping information directly to applications. Some application layer proxies e.g. XFF (X-Forwarded-For), can convey this information in-band. Another approach is to ask application coordinating the information with NAT logging. But that is not sufficient, since the applications itself wants to know the original source address from an application message bus. The logging information may be used by administrators offline to inspect use behavior and preference analysis, and accurate advertisement delivery.

4.4. DNS Resolving

In the case of NAT64-CE, it is recommended to follow the recommendations in [RFC6144]. There is no need for the DNS to synthesize AAAA from A records, since static AAAA records can be registered in the authoritative DNS for a given domain to represent...
these IPv4-only hosts. How to design the FQDN for the IPv6 service is out-of-scope of this document.

4.5. Load Balancer

Load balancing on NAT64-CE has a couple of considerations. If dictated by scale or availability requirements traffic should be balanced among multiple NAT64-CE devices. One point to be noted is that synthetic AAAA records may be added directly in authoritative DNS. load balancing based on DNS64 synthetic resource records may not work in those cases. Secondly, NAT64-CE could also serve as the load balancer for IPv4 backend servers. There are also some ways of load balance for the cases, where load balancer is placed in front of NAT64(s).

4.6. MTU Consideration

As compared to the MTU consideration in NAT64-CGN, the MTU of IPv4 network are strongly recommended to set to more than 1260. Since a CE IPv4 network is normally operated by a particular administrative entity, it should take steps to prevent the risk of fragmentation discussed in [I-D.ietf-6man-ipv6-atomic-fragments].

5. Security Considerations

This document presents the deployment experiences of NAT64 in CGN and CE scenario, some security considerations are described in detail regarding to specific NAT64 mode in section 2 and 3. In general, RFC 6146[RFC6146] provides TCP-tracking, address-dependent filtering mechanisms to protect NAT64 from DDOS. In NAT64-CGN cases, ISP also could adopt uRPF and black/white-list to enhance the security by specifying access policies. for example, NAT64-CGN should forbid establish NAT64 BIB for incoming IPv6 packets if URPF (Strict or Loose mode) check does not pass or whose source IPv6 address is associated to black-lists.

6. IANA Considerations

This memo includes no request to IANA.

7. Acknowledgements

The authors would like to thank Jari Arkko, Dan Wing, Remi Despres, Fred Baker, Lee Howard and Iljitsch van Beijnum for their helpful comments. Many thanks to Wesley George and Satoru Matsushima for
their reviews.

The authors especially thank Joel Jaeggli for his efforts and contributions on editing which substantially improves the legibility of the document.

8. Additional Author List

The following are extended authors who contributed to the effort:

Qiong Sun
China Telecom
Room 708, No.118, Xizhimennei Street
Beijing 100035
P.R.China
Phone: +86-10-58552936
Email: sunqiong@ctbri.com.cn

QiBo Niu
ZTE
50,RuanJian Road.
YuHua District,
Nan Jing  210012
P.R.China
Email: niu.qibo@zte.com.cn

9. References

9.1. Normative References

[I-D.ietf-pcp-base]


9.2. Informative References


progress), July 2011.


Authors’ Addresses

Gang Chen
China Mobile
53A,Xibianmennei Ave.,
Xuanwu District,
Beijing 100053
China

Email: phdgang@gmail.com

Zhen Cao
China Mobile
53A,Xibianmennei Ave.,
Xuanwu District,
Beijing 100053
China

Email: caozhen@chinamobile.com
Cameron Byrne
T-Mobile USA
Bellevue
Washington 98105
USA

Email: cameron.byrne@t-mobile.com

Chongfeng Xie
China Telecom
Room 708 No.118, Xizhimenneidajie
Beijing 100035
P.R.China

Email: xiechf@ctbri.com.cn

David Binet
France Telecom
Rennes
35000
France

Email: david.binet@orange.com
Deterministic Address Mapping to Reduce Logging in Carrier Grade NAT Deployments
draft-donley-behave-deterministic-cgn-09

Abstract

In some instances, Service Providers have a legal logging requirement to be able to map a subscriber’s inside address with the address used on the public Internet (e.g. for abuse response). Unfortunately, many Carrier Grade NAT logging solutions require active logging of dynamic translations. Carrier Grade NAT port assignments are often per-connection, but could optionally use port ranges. Research indicates that per-connection logging is not scalable in many residential broadband services. This document suggests a way to manage Carrier Grade NAT translations in such a way as to significantly reduce the amount of logging required while providing traceability for abuse response. IPv6 is, of course, the preferred solution. While deployment is in progress, service providers are forced by business imperatives to maintain support for IPv4. This note addresses the IPv4 part of the network when a Carrier Grade NAT solution is in use.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute
working documents as Internet-Drafts. The list of current Internet-
Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on June 14, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the
document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal
Provisions Relating to IETF Documents
(http://trustee.ietf.org/license-info) in effect on the date of
publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with respect
to this document. Code Components extracted from this document must
include Simplified BSD License text as described in Section 4.e of
the Trust Legal Provisions and are provided without warranty as
described in the Simplified BSD License.

Table of Contents

1. Introduction ............................................. 3
2. Deterministic Port Ranges .............................. 4
   2.1. IPv4 Port Utilization Efficiency ..................... 7
   2.2. Planning & Dimensioning ............................. 8
   2.3. Deterministic CGN Example ........................... 8
3. Additional Logging Considerations ...................... 10
   3.1. Failover Considerations ............................. 10
4. Impact on the IPv6 Transition ........................... 11
5. Privacy Considerations ................................... 11
6. IANA Considerations ...................................... 11
7. Security Considerations .................................. 11
8. Acknowledgements ........................................ 12
9. References ............................................... 12
   9.1. Normative References ............................... 12
   9.2. Informative References ............................. 12
Authors’ Addresses ......................................... 14
1. Introduction

It is becoming increasingly difficult to obtain new IPv4 address assignments from Regional/Local Internet Registries due to depleting supplies of unallocated IPv4 address space. To meet the growing demand for Internet connectivity from new subscribers, devices, and service types, some operators will be forced to share a single public IPv4 address among multiple subscribers using techniques such as Carrier Grade Network Address Translation (CGN) [RFC6264] (e.g., NAT444 [I-D.shirasaki-nat444], DS-Lite [RFC6333], NAT64 [RFC6146] etc.). However, address sharing poses additional challenges to operators when considering how they manage service entitlement, public safety requests, or attack/abuse/fraud reports [RFC6269]. In order to identify a specific user associated with an IP address in response to such a request or for service entitlement, an operator will need to map a subscriber’s internal source IP address and source port with the global public IP address and source port provided by the CGN for every connection initiated by the user.

CGN connection logging satisfies the need to identify attackers and respond to abuse/public safety requests, but it imposes significant operational challenges to operators. In lab testing, we have observed CGN log messages to be approximately 150 bytes long for NAT444 [I-D.shirasaki-nat444], and 175 bytes for DS-Lite [RFC6333] (individual log messages vary somewhat in size). Although we are not aware of definitive studies of connection rates per subscriber, reports from several operators in the US sets the average number of connections per household at approximately 33,000 connections per day. If each connection is individually logged, this translates to a data volume of approximately 5 MB per subscriber per day, or about 150 MB per subscriber per month; however, specific data volumes may vary across different operators based on myriad factors. Based on available data, a 1-million subscriber service provider will generate approximately 150 terabytes of log data per month, or 1.8 petabytes per year. Note that many Service Providers compress log data after collection; compression factors of 2:1 or 3:1 are common.

The volume of log data poses a problem for both operators and the public safety community. On the operator side, it requires a significant infrastructure investment by operators implementing CGN. It also requires updated operational practices to maintain the logging infrastructure, and requires approximately 23 Mbps of bandwidth between the CGN devices and the logging infrastructure per 50,000 users. On the public safety side, it increases the time required for an operator to search the logs in response to an abuse report, and could delay investigations. Accordingly, an international group of operators and public safety officials
approached the authors to identify a way to reduce this impact while improving abuse response.

The volume of CGN logging can be reduced by assigning port ranges instead of individual ports. Using this method, only the assignment of a new port range is logged. This may massively reduce logging volume. The log reduction may vary depending on the length of the assigned port range, whether the port range is static or dynamic, etc. This has been acknowledged in [RFC6269], which recommends source port logging at the server and/or destination logging at the CGN and [I-D.sivakumar-behave-nat-logging], which describes information to be logged at a NAT.

However, the existing solutions still poses an impact on operators and public safety officials for logging and searching. Instead, CGNs could be designed and/or configured to deterministically map internal addresses to (external address + port range) in such a way as to be able to algorithmically calculate the mapping. Only inputs and configuration of the algorithm need to be logged. This approach reduces both logging volume and subscriber identification times. In some cases, when full deterministic allocation is used, this approach can eliminate the need for translation logging.

This document describes a method for such CGN address mapping, combined with block port reservations, that significantly reduces the burden on operators while offering the ability to map a subscriber’s inside IP address with an outside address and external port number observed on the Internet.

The activation of the proposed port range allocation scheme is compliant with BEHAVE requirements such as the support of APP.

2. Deterministic Port Ranges

While a subscriber uses thousands of connections per day, most subscribers use far fewer resources at any given time. When the compression ratio (see Appendix B of RFC6269 [RFC6269]) is low (e.g., the ratio of the number of subscribers to the number of public IPv4 addresses allocated to a CGN is closer to 10:1 than 1000:1), each subscriber could expect to have access to thousands of TCP/UDP ports at any given time. Thus, as an alternative to logging each connection, CGNs could deterministically map customer private addresses (received on the customer-facing interface of the CGN, a.k.a., internal side) to public addresses extended with port ranges (used on the Internet-facing interface of the CGN, a.k.a., external side). This algorithm allows an operator to identify a subscriber internal IP address when provided the public side IP and port number without having to examine the CGN translation logs. This prevents an
operator from having to transport and store massive amounts of
session data from the CGN and then process it to identify a
subscriber.

The algorithmic mapping can be expressed as:

(External IP Address, Port Range) = function 1 (Internal IP Address)

Internal IP Address = function 2 (External IP Address, Port Number)

The CGN SHOULD provide a method for administrators to test both
mapping functions (e.g., enter an External IP Address + Port Number
and receive the corresponding Internal IP Address).

Deterministic Port Range allocation requires configuration of the
following variables:

- Inside IPv4/IPv6 address range (I);
- Outside IPv4 address range (O);
- Compression ratio (e.g. inside IP addresses I/outside IP addresses
  O) (C);
- Dynamic address pool factor (D), to be added to the compression
  ratio in order to create an overflow address pool;
- Maximum ports per user (M);
- Address assignment algorithm (A) (see below); and
- Reserved TCP/UDP port list (R)

Note: The inside address range (I) will be an IPv4 range in NAT444
operation (NAT444 [I-D.shirasaki-nat444]) and an IPv6 range in DS-
Lite operation (DS-Lite [RFC6333]).

A subscriber is identified by an internal IPv4 address (e.g., NAT44)
or an IPv6 prefix (e.g., DS-Lite or NAT64).

The algorithm may be generalized to L2-aware NAT
[I-D.miles-behave-l2nat] but this requires the configuration of the
Internal interface identifiers (e.g., MAC addresses).

The algorithm is not designed to retrieve an internal host among
those sharing the same internal IP address (e.g., in a DS-Lite
context, only an IPv6 address/prefix can be retrieved using the
Several address assignment algorithms are possible. Using predefined algorithms, such as those that follow, simplifies the process of reversing the algorithm when needed. However, the CGN MAY support additional algorithms. Also, the CGN is not required to support all algorithms described below. Subscribers could be restricted to ports from a single IPv4 address, or could be allocated ports across all addresses in a pool, for example. The following algorithms and corresponding values of A are as follow:

0: Sequential (e.g. the first block goes to address 1, the second block to address 2, etc.)

1: Staggered (e.g. for every n between 0 and ((65536-R)/(C+D))-1, address 1 receives ports n*C+R, address 2 receives ports (1+n)*C+R, etc.)

2: Round robin (e.g. the subscriber receives the same port number across a pool of external IP addresses. If the subscriber is to be assigned more ports than there are in the external IP pool, the subscriber receives the next highest port across the IP pool, and so on. Thus, if there are 10 IP addresses in a pool and a subscriber is assigned 1000 ports, the subscriber would receive a range such as ports 2000-2099 across all 10 external IP addresses).

3: Interlaced horizontally (e.g. each address receives every Cth port spread across a pool of external IP addresses).

4: Cryptographically random port assignment (Section 2.2 of RFC6431 [RFC6431]). If this algorithm is used, the Service Provider needs to retain the keying material and specific cryptographic function to support reversibility.

5: Vendor-specific. Other vendor-specific algorithms may also be supported.

The assigned range of ports MAY also be used when translating ICMP requests (when re-writing the Identifier field).

The CGN then reserves ports as follows:

1. The CGN removes reserved ports (R) from the port candidate list (e.g., 0-1023 for TCP and UDP). At a minimum, the CGN SHOULD remove system ports (RFC6335) [RFC6335] from the port candidate list reserved for deterministic assignment.
2. The CGN calculates the total compression ratio \((C+D)\), and allocates \(1/(C+D)\) of the available ports to each internal IP address. Specific port allocation is determined by the algorithm (A) configured on the CGN. Any remaining ports are allocated to the dynamic pool.

Note: Setting \(D\) to 0 disables the dynamic pool. This option eliminates the need for per-subscriber logging at the expense of limiting the number of concurrent connections that ‘power users’ can initiate.

3. When a subscriber initiates a connection, the CGN creates a translation mapping between the subscriber’s inside local IP address/port and the CGN outside global IP address/port. The CGN MUST use one of the ports allocated in step 2 for the translation as long as such ports are available. The CGN SHOULD allocate ports randomly within the port range assigned by the deterministic algorithm. This is to increase subscriber privacy. The CGN MUST use the preallocated port range from step 2 for Port Control Protocol (PCP, [RFC6887]) reservations as long as such ports are available. While the CGN maintains its mapping table, it need not generate a log entry for translation mappings created in this step.

4. If \(D>0\), the CGN will have a pool of ports left for dynamic assignment. If a subscriber uses more than the range of ports allocated in step 2 (but fewer than the configured maximum ports \(M\)), the CGN assigns a block of ports from the dynamic assignment range for such a connection or for PCP reservations. The CGN MUST log dynamically assigned port blocks to facilitate subscriber-to-address mapping. The CGN SHOULD manage dynamic ports as described in [I-D.tsou-behave-natx4-log-reduction].

5. Configuration of reserved ports (e.g., system ports) is left to operator configuration.

Thus, the CGN will maintain translation mapping information for all connections within its internal translation tables; however, it only needs to externally log translations for dynamically-assigned ports.

2.1. IPv4 Port Utilization Efficiency

For Service Providers requiring an aggressive address sharing ratio, the use of the algorithmic mapping may impact the efficiency of the address sharing. A dynamic port range allocation assignment is more suitable in those cases.
2.2. Planning & Dimensioning

Unlike dynamic approaches, the use of the algorithmic mapping requires more effort from operational teams to tweak the algorithm (e.g., size of the port range, address sharing ratio, etc.). Dedicated alarms SHOULD be configured when some port utilization thresholds are fired so that the configuration can be refined.

The use of algorithmic mapping also affects geolocation. Changes to the inside and outside address ranges (e.g. due to growth, address allocation planning, etc.) would require external geolocation providers to recalibrate their mappings.

2.3. Deterministic CGN Example

To illustrate the use of deterministic NAT, let’s consider a simple example. The operator configures an inside address range (I) of 198.51.100.0/28 [RFC6598] and outside address (O) of 192.0.2.1. The dynamic address pool factor (D) is set to ‘2’. Thus, the total compression ratio is 1:(14+2) = 1:16. Only the system ports (e.g. ports < 1024) are reserved (R). This configuration causes the CGN to preallocate ((65536-1024)/16 =) 4032 TCP and 4032 UDP ports per inside IPv4 address. For the purposes of this example, let’s assume that they are allocated sequentially, where 198.51.100.1 maps to 192.0.2.1 ports 1024-5055, 198.51.100.2 maps to 192.0.2.1 ports 5056-9087, etc. The dynamic port range thus contains ports 57472-65535 (port allocation illustrated in the table below). Finally, the maximum ports/subscriber is set to 5040.
When subscriber 1 using 198.51.100.1 initiates a low volume of connections (e.g. < 4032 concurrent connections), the CGN maps the outgoing source address/port to the preallocated range. These translation mappings are not logged.

Subscriber 2 concurrently uses more than the allocated 4032 ports (e.g. for peer-to-peer, mapping, video streaming, or other connection-intensive traffic types), the CGN allocates up to an additional 1008 ports using bulk port reservations. In this example, subscriber 2 uses outside ports 5056-9087, and then 100-port blocks between 58000-58999. Connections using ports 5056-9087 are not logged, while 10 log entries are created for ports 58000-58099, 58100-58199, 58200-58299, ..., 58900-58999.

In order to identify a subscriber behind a CGN (regardless of port allocation method), public safety agencies need to collect source address and port information from content provider log files. Thus, content providers are advised to log source address, source port, and timestamp for all log entries, per [RFC6302]. If a public safety agency collects such information from a content provider and reports abuse from 192.0.2.1, port 2001, the operator can reverse the mapping algorithm to determine that the internal IP address subscriber 1 has been assigned generated the traffic without consulting CGN logs (by correlating the internal IP address with DHCP/PPP lease connection records). If a second abuse report comes in for 192.0.2.1, port 58204, the operator will determine that port 58204 is within the dynamic pool range, consult the log file, correlate with connection
records, and determine that subscriber 2 generated the traffic (assuming that the public safety timestamp matches the operator timestamp. As noted in RFC6292 [RFC6292], accurate time-keeping (e.g., use of NTP or Simple NTP) is vital).

In this example, there are no log entries for the majority of subscribers, who only use pre-allocated ports. Only minimal logging would be needed for those few subscribers who exceed their pre-allocated ports and obtain extra bulk port assignments from the dynamic pool. Logging data for those users will include inside address, outside address, outside port range, and timestamp.

Note that in a production environment, operators are encouraged to consider [RFC6598] for assigning inside addresses.

3. Additional Logging Considerations

In order to be able to identify a subscriber based on observed external IPv4 address, port, and timestamp, an operator needs to know how the CGN was configured with regards to internal and external IP addresses, dynamic address pool factor, maximum ports per user, and reserved port range at any given time. Therefore, the CGN MUST generate a record any time such variables are changed. The CGN SHOULD generate a log message any time such variables are changed. The CGN MAY keep such a record in the form of a router configuration file. If the CGN does not generate a log message, it would be up to the operator to maintain version control of router config changes. Also, the CGN SHOULD generate such a log message once per day to facilitate quick identification of the relevant configuration in the event of an abuse notification.

Such a log message MUST, at minimum, include the timestamp, inside prefix I, inside mask, outside prefix O, outside mask, D, M, A, and reserved port list R; for example:


3.1. Failover Considerations

Due to the deterministic nature of algorithmically-assigned translations, no additional logging is required during failover conditions provided that inside address ranges are unique within a given failover domain. Even when directed to a different CGN server, translations within the deterministic port range on either the primary or secondary server can be algorithmically reversed, provided the algorithm is known. Thus, if 198.51.100.1 port 3456 maps to 192.0.2.1 port 1000 on CGN 1 and 198.51.100.1 port 1000 on Failover
CGN 2, an operator can identify the subscriber based on outside source address and port information.

Similarly, assignments made from the dynamic overflow pool need to be logged as described above, whether translations are performed on the primary or failover CGN.

4. Impact on the IPv6 Transition

The solution described in this document is applicable to Carrier Grade NAT transition technologies (e.g. NAT444, DS-Lite, and NAT64). As discussed in [RFC7021], the authors acknowledge that native IPv6 will offer subscribers a better experience than CGN. However, many CPE devices only support IPv4. Likewise, as of October 2014, only approximately 5.2% of the top 1 million websites were available using IPv6. Accordingly, Deterministic CGN should in no way be understood as making CGN a replacement for IPv6 service; however, until such time as IPv6 content and devices are widely available, Deterministic CGN will provide operators with the ability to quickly respond to public safety requests without requiring excessive infrastructure, operations, and bandwidth to support per-connection logging.

5. Privacy Considerations

The algorithm described above makes it easier for Service Providers and public safety officials to identify the IP address of a subscriber through a CGN system. This is the equivalent level of privacy users could expect when they are assigned a public IP address and their traffic is not translated. However, this algorithm could be used by other actors on the Internet to map multiple transactions to a single subscriber, particularly if ports are distributed sequentially. While still preserving traceability, subscriber privacy can be increased by using one of the other values of the Address Assignment Algorithm (A), which would require interested parties to know more about the Service Provider’s CGN configuration to be able to tie multiple connections to a particular subscriber.

6. IANA Considerations

This document makes no request of IANA.

7. Security Considerations

The security considerations applicable to NAT operation for various protocols as documented in, for example, RFC 4787 [RFC4787] and RFC 5382 [RFC5382] also apply to this document.
Note that with the possible exception of cryptographically-based port allocations, attackers could reverse-engineer algorithmically-derived port allocations to either target a specific subscriber or to spoof traffic to make it appear to have been generated by a specific subscriber. However, this is exactly the same level of security that the subscriber would experience in the absence of CGN. CGN is not intended to provide additional security by obscurity.

8. Acknowledgements

The authors would like to thank the following people for their suggestions and feedback: Bobby Flaim, Lee Howard, Wes George, Jean-Francois Tremblay, Mohammed Boucadair, Alain Durand, David Miles, Andy Anchev, Victor Kuarsingh, Miguel Cros Cecilia, Fred Baker, Brian Carpenter, and Reinaldo Penno.

9. References

9.1. Normative References


9.2. Informative References

[I-D.shirasaki-nat444]
Yamagata, I., Shirasaki, Y., Nakagawa, A., Yamaguchi, J.,
and H. Ashida, "NAT444", draft-shirasaki-nat444-06 (work
in progress), July 2012.

[I-D.sivakumar-behave-nat-logging]
Sivakumar, S. and R. Penno, "IPFIX Information Elements
for logging NAT Events", draft-sivakumar-behave-nat-
logging-06 (work in progress), January 2013.

[I-D.tsou-behave-natx4-log-reduction]
Tsou, T., Li, W., Taylor, T., and J. Huang, "Port
Management To Reduce Logging In Large-Scale NATs", draft-
tsou-behave-natx4-log-reduction-04 (work in progress),
July 2013.

[RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful
NAT64: Network Address and Protocol Translation from IPv6
Clients to IPv4 Servers", RFC 6146, April 2011.

[RFC6292] Hoffman, P., "Requirements for a Working Group Charter
Tool", RFC 6292, June 2011.

[RFC6302] Durand, A., Gashinsky, I., Lee, D., and S. Sheppard,
"Logging Recommendations for Internet-Facing Servers", BCP
162, RFC 6302, June 2011.

[RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-
Stack Lite Broadband Deployments Following IPv4
Exhaustion", RFC 6333, August 2011.

Cheshire, "Internet Assigned Numbers Authority (IANA)
Procedures for the Management of the Service Name and
Transport Protocol Port Number Registry", BCP 165, RFC
6335, August 2011.

[RFC6431] Boucadair, M., Levis, P., Bajko, G., Savolainen, T., and
T. Tsou, "Huawei Port Range Configuration Options for PPP
IP Control Protocol (IPCP)", RFC 6431, November 2011.

[RFC6598] Weil, J., Kuarsingh, V., Donley, C., Liljenstolpe, C., and
M. Azinger, "IANA-Reserved IPv4 Prefix for Shared Address
Space", BCP 153, RFC 6598, April 2012.

[RFC6887] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P.
Selkirk, "Port Control Protocol (PCP)", RFC 6887, April
2013.

Authors’ Addresses

Chris Donley
CableLabs
858 Coal Creek Cir
Louisville, CO  80027
US

Email: c.donley@cablelabs.com

Chris Grundemann
Internet Society
Denver, CO
US

Email: cgrundemann@gmail.com

Vikas Sarawat
CableLabs
858 Coal Creek Cir
Louisville, CO  80027
US

Email: v.sarawat@cablelabs.com

Karthik Sundaresan
CableLabs
858 Coal Creek Cir
Louisville, CO  80027
US

Email: k.sundaresan@cablelabs.com

Olivier Vautrin
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA  94089
US

Email: olivier@juniper.net
Abstract

The tremendous growth in Wi-Fi technology adoption over the last decade has met the ultimate possible goal of 100% adoption rate. All most every new mobile device is now equipped with IEEE 802.11-based wireless interface and with pre-configured policy to prefer Wi-Fi to cellular access. Matching this evolution is every service provider’s desire to offer Wi-Fi based broadband services; a new business opportunity even for fixed line operators. Operators are exploring options to monetize their existing networks, most with nation-wide footprint, to build a high-speed Wi-Fi service that can be the basis for offering new wireless broadband services. This document identifies the requirements for supporting these new Wi-Fi community services and the mobility tools which have been standardized in IETF that can be used for enabling these architectures.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on October 25, 2013.

Copyright Notice
# Table of Contents

1. Introduction ............................................. 4  
2. Conventions and Terminology ............................. 5  
   2.1. Conventions ........................................ 5  
   2.2. Terminology ........................................ 5  
3. Deployment Models ....................................... 7  
4. Requirements ............................................ 8  
   4.1. IPv6 Addressing Model for SP Wi-Fi Architectures .... 9  
   4.2. Subscriber Authentication & Service Authorization .... 9  
   4.3. Location-based Services .............................. 9  
   4.4. Local Services Access & Internet Traffic Offload .... 10  
   4.5. Web-based Authentication Support .................... 10  
   4.6. Transparent Auto Login (TAL) ........................ 10  
   4.7. Multiple WLAN SSID Support ........................ 11  
   4.9. CPE Identity and Authorization ....................... 11  
   4.10. Mobility within the WLAN Access Network ............ 11  
   4.11. Mobility across WLAN and Macro Access ............. 12  
   4.12. Differentiated Services for Users behind RG ........ 12  
   4.13. Lawful Intercept (LI) ................................ 12  
   4.15. Handling the Walk-by Users .......................... 14  
   4.16. Overlapping IPv4 Address Support ................... 14  
   4.17. Service Provisioning & Monitoring ................... 14  
5. Solution Approaches & Considerations .................. 15  
   5.1. PMIPv6 MAG on the RG: Layer-3 Encapsulation between  
        CPE and Access Gateway ............................. 15  
   5.2. Ethernet-over-IP Support on the RG: Layer-2  
        Encapsulation between CPE and Access Gateway ....... 15  
   5.3. Local Aggregation for Subscriber Control and Internet  
        Offload ............................................ 15  
   5.4. Mobility Chaining: Integration with Mobile Packet Core .. 15  
6. IANA Considerations .................................... 15  
7. Security Considerations ................................. 15  
8. Acknowledgements ...................................... 16  
9. References ............................................. 16  
   9.1. Normative References ............................... 16  
   9.2. Informative References ............................. 16  
Authors’ Addresses ......................................... 17
1. Introduction

The tremendous growth in Wi-Fi technology adoption over the last decade has met the ultimate possible goal of 100% adoption rate. All most every new mobile device is now equipped with IEEE 802.11-based wireless interface and these devices are typically pre-configured with a policy to prefer Wi-Fi to cellular access. This so called, “cheap access based on unlicensed spectrum”, is no longer considered an unreliable access, but with all the available protocol tools and with maturity in technology, building a reliable broadband service that can meet the committed service-level agreements is proving to be a non-issue.

Matching this evolution is every service provider’s desire to offer Wi-Fi based broadband services; a new business opportunity even for both fixed and mobile operators. The demand for bandwidth is only growing with the availability of new smart devices, new technology applications and with all the content in the Internet. Furthermore, an increasing percentage of mobile consumption is happening in the home and so DSL/Cable operators are exploring options to monetize their existing networks, most with nation-wide footprint, to build a high-speed, nation-wide Wi-Fi service that can be the basis for offering new wireless broadband services and for building roaming agreements with traditional mobile operators, who are unable to meet the mobile subscriber growth due to the finite licensed spectrum available for macro-cell deployments. Every residential CPE device that the operator owns can now be enabled to provide Wi-Fi service and new community Wi-Fi hotspots can be built in any location where there is fixed line coverage. A wireless service based on unlicensed spectrum, and leveraging existing transport is a huge incentive for operators to enter this new market.

To support these business goals, operators are looking at mobility architectures for supporting various requirements. Not all requirements are well understood, and neither are the implications with the chosen solution approaches for each of those requirements. The choice of the architecture has an implication on the CPE evolution and on the core infrastructure feature requirements. Therefore, the sole purpose and the goal of this document is to present all the requirements, identify the protocol tools and any potential gaps. This analysis is important for enabling the network vendors and the mobile operators to make the right design choices and leverage the existing tools that the mobility groups in IETF have already developed and discourage them from adopting proprietary, non-standard mechanisms or developing redundant alternatives.
2. Conventions and Terminology

2.1. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2.2. Terminology

This document uses the following abbreviations and definitions:

Community Wi-Fi Service

It is a Wi-Fi based broadband service offered by a service provider. The Wi-Fi Access Points that are part of this service are owned and managed by the operator, and physically located in carrier premises. These operator owned CPE’s typically have a large Wi-Fi coverage area, operated on a higher signal power.
There could also be the residential Access Points that are part of this service, located in the subscriber homes, that are part of this service and allowing community access to a public SSID along with a private SSID for their personal access.

Wi-Fi Operator

A service provider that offers Community Wi-Fi services. Wi-Fi operator can be a wireline operator, mobile operator or an operator offering both wireline and mobile services.

Residential Gateway (RG)

It is a network device that is located in the Customer premises and is also referred to as Residential CPE (Customer Premises Equipment). This device is connected to service providers network and defines the demarcation point between the provider and the customer. In the context of this document this is hosting the 802.11 Access Point function.

WLAN controller (WLC)

It is an entity responsible for performing radio resource management (RRM) on the Access Points, system-wide mobility policy enforcement and centralized forwarding function for the user traffic.

Mobile Gateway

It is network entity anchoring IP traffic in the mobile core network. This entity allocates an IP address which is topologically valid in the mobile network and may act as a mobility anchor if handover between mobile and Wi-Fi is supported.

Home/Roaming User

The home user is the owner of the network where the Residential Gateway is located and is paying for the service associated with that Residential Gateway. A Roaming User is a visitor from the operator’s home network, or from a partner’s network and is allowed to access broadband services using that Residential Gateway and over a Public SSID.

Access Point Name (APN)
Its the name of a packet data network. This APN concept was first introduced in GPRS by 3GPP to enable legacy Intelligent Networking (IN) approaches to be applied to the newly deployed IP packet data services. In roaming deployments, the APN construct was visible to the visited network and allowed legacy IN charging solutions to be supported. Defining an application specific APN then allowed application charging to be supported.

Addressing Models

The term Per-MN-Prefix model [RFC5213] is used to refer to an addressing model where there is a unique network prefix or prefixes assigned for each mobile node. The term Shared-Prefix model [RFC5213] is used to refer to an addressing model where the prefix(es) are shared by more than one node.

3. Deployment Models

Figure 2 illustrates the most common residential and hotspots Wi-Fi deployment models.
4. Requirements
4.1. IPv6 Addressing Model for SP WiFi Architectures

The selection of the right IPv6 addressing model for the SP WiFi architectures is an important consideration. There are these two IPv6 addressing models:

- **Unique-Prefix Model** - As per this addressing model, home network prefix(es) assigned to a mobile node are for its exclusive use and no other node shares an address from that prefix (other than the Subnet-Router anycast address [RFC4291] that is used by the IPv6 access router hosting that prefix on that link). There could be multiple unique IPv6 prefixes assigned to each mobile node.

- **Shared-Prefix model** - The IPv6 prefix that is assigned to the mobile node is a shared prefix. There can be more than one mobile node that can be using IPv6 addresses from that prefix.

3GPP architecture supports Unique-Prefix model for the mobile node’s PDN connections. This decision was largely influenced by the IETF recommendation to 3GPP to support this specific addressing model. In the context of SP WiFi, there are clearly scenarios where a mobile node may perform an inter-technology handover from the macro network to the WLAN access network and handoff the session and is important that the addressing model is the same in both the access architectures. Even in deployment models where such handovers are not envisioned, such as an WLAN access aggregation architecture with no mobile packet core integration, there are sufficient reasons for adopting the Unique Prefix model.

4.2. Subscriber Authentication & Service Authorization

Community Wi-Fi service is designed to be available for public access. Wi-Fi operator must authenticate users before offering services to them. Once a user is authenticated, Wi-Fi operator will authorize services based on the user identity. There are many authentication mechanisms, such as 802.1x, Web-authentication, WISPr that the operator may deploy for this purpose.

4.3. Location-based Services

In many deployments, there is a need for the mobile operator to provide differentiated services and policing to the mobile nodes based on the access network to which they are attached. Policy systems in mobility architectures such as PCC and ANDSF in 3GPP system allow configuration of policy rules with conditions based on the access network information. For example, the service treatment for the mobile node’s traffic may be different when they are attached to a access network owned by the home operator than when owned by a
The service treatment can also be different based on the configured Service Set Identifiers (SSID) in case of IEEE 802.11 based access networks. Other examples of location services include the operator’s ability to display a location specific Web Page, or apply tariff based on the location.

4.4. Local Services Access & Internet Traffic Offload

In the integrated WLAN-EPC architectures, the mobile node’s IP traffic is always tunneled back from the access network to the mobile gateway in the home network. However, with the exponential growth in the mobile data traffic, mobile operators are exploring new ways to offload some of the IP traffic flows at the nearest access edge where ever there is an internet peering point, as supposed to carrying it all the way to the mobility anchor in the home network. Not all IP traffic need to be routed back to the home network, some of the non-essential traffic which does not require IP mobility support can be offloaded at the mobile access gateway in the access network. This approach provides greater leverage and efficient usage of the mobile packet core which help lowering transport cost.

4.5. Web-based Authentication Support

Most Public Wireless LAN (PWLAN) deployments today use web-based authentication for authorizing the user for network access. Web-based mode of authentication is considered a legacy mode, for its weak security properties, and there are efforts to replace it with 802.1x-based security mechanisms. However, a very high percentage of the PWLAN deployments are still using using this authentication mode and operators are not willing to move away from this mode any time soon. The reason being, lack of support for 802.1x/EAP support on the 100’s of millions of handsets that are out there, and for the lack of client software in the laptops running various operating systems versions. This is forcing the operators to support web-based authentication.

4.6. Transparent Auto Login (TAL)

In many deployments, there is a need to support Transparent Auto Login capability. This is essentially an approach for maintaining Authenticated state for a user, for a duration of time. Once an authenticated user disconnects and re-attaches to the network, the network should allows instant access without forcing the user to re-authenticate.
4.7. Multiple WLAN SSID Support

A Wi-Fi Operator may broadcast multiple SSIDs. In case Residential Wi-Fi hotspots, there can be one set of private SSIDs specific to that home user and there can be another set of public SSIDs for wider community use. In case of public hotspots, the operator can advertise the public SSID for its own subscribers and also public SSID’s belonging to other operators with whom the operator has roaming relationships.

4.8. Multiple Home Network Service (APN) Access

The 3GPP system architecture supports the concept of an Access Point Name (APN). An APN can identify a particular routing domain and can be used by 3GPP operators to segment user traffic. APNs are included in the session establishment signaling sent by 3GPP User Equipments (UEs), identifying which routing domain they want to be connected to. Furthermore, 3GPP has defined a system architecture which supports the ability of a single UE to have simultaneous connectivity to a plurality of APNs, and be allocated multiple IPv4 addresses and/or IPv6 prefixes from the network.

There is a need to ensure multiple APN access for a subscriber in the community Wi-Fi network.

4.9. CPE Identity and Authorization

There are two known models with respect to CPE roll out. The consumer may purchase a device off the shelf and plugin to the network, or the operator at the time of service creation may have shipped a new device with the pre-provisioned service configuration. In either case, the operator needs to be able to identify the device based on the IP address and associate that to a given location.

The Wi-Fi network performs access control of UEs, via the CPE acting as AAA supplicant. As a result, the mobile network does not authenticate directly the user but shall trust the CPE performing the authentication.

4.10. Mobility within the WLAN Access Network

The mobile node should have the ability to roam within the Wi-Fi domain. Depending on the deployment model, the mobile node may roam across different IP subnets. To survive to such handover, some applications (e.g. VPN, streaming) need the IP address to be preserved.

A WLAN network may include a large number of Wi-Fi base stations. In
some occasions, two or more Wi-Fi base stations may cover the same area. When a subscriber receives Wi-Fi service in this overlapped area, the device may bounce between different base stations. This is typical Proximity problem. In this scenario, it is important for the WLAN to offer mobility to the subscriber as such the subscriber can continue the services without changing its IP address.

4.11. Mobility across WLAN and Macro Access

A mobile node should have the ability to handover from macro network to the Wi-Fi network and be able to retain IP address configuration and be able to access the home operator services.

4.12. Differentiated Services for Users behind RG

A Wi-Fi operator enabling Hotspot Services on a residential gateway is required to ensure the service levels for the home user is not impacted as a result of opening up the service for public usage. The home user should always have preferred access over public users and the operator may be bound to meet the Service Level Agreements. This essentially requires the operator to be able to differentiate the service flows and apply differentiated service treatment. The operator should be able to enforce QoS policing and labeling of packets to enforce QoS differentiation.

A single operator has deployed both a fixed access network and a mobile access network. In this scenario, the operator may wish a harmonized QoS management on both accesses. However the fixed access network does not implement a QoS control framework. So, the operator may choose to rely on the mobile network, specifying the standard framework to provide a QoS control, to enforce the QoS policy from the mobile gateway to the Wi-Fi Access network.

4.13. Lawful Intercept (LI)

Lawful Intercept [RFC2119] stands for legally authorized interception and monitoring of communications to and from a subscriber under Surveillance by a Law Enforcement Agency. In most of the countries, there are legal obligations for Service Providers to facilitate the intercept of any subscriber’s communication if requested by law enforcement agencies. Communications Assistance for Law Enforcement Act (CALEA), the United States wiretapping law passed in 1994 is an example for such legal mandates. This section talks about Lawful Intercept solution requirements that are operators are required to support when offering WLAN services.

The following are the key considerations with respect to supporting Lawful Intercept capability in Wi-Fi architectures.
o The operator should have the ability to capture IP traffic from any of the mobile nodes for which the operator is offering Wi-Fi services.

o The ability to identify the Geo-location of the mobile node to the nearest WLAN access point.

o The ability to track the mobile node’s roaming within the network, even when there are no active IP flows.

o The ability to pre-provision Lawful Intercept for an inactive mobile node so that the capture of IP traffic can be initiated anytime new IP flows associated to that mobile node are detected.

o Lawful Intercept (LI) should be undetectable by the intercept subject

o Mechanisms should be in place to limit unauthorized personnel from performing or knowing about lawfully authorized intercepts

o If the information being intercepted is encrypted by the service provider and the service provider has access to the keys, then the information should be decrypted before delivery to the Law Enforcement Agency (LEA) or the encryption keys should be passed to the Law Enforcement Agency to allow them to decrypt the information.


It refers to the capability to manage network resources on a per subscriber, and eventually on a per-flow, basis. Subscriber management should be able to maintain a user context associating the user identifier with specific network resource (e.g. IP address, default router, mobility/traffic anchoring point,...), QoS profile, billing context and specific network functions (e.g. legal interception). The user context includes traffic selectors if subscriber management is on a per flow basis. Subscriber management should be done according to the user subscription, the user preferences and/or operator policies.

The ability to charge the subscriber is the fundamental business requirement before an operator can deploy the Wi-Fi service. The operator should have the ability to enforce charge the subscriber by usage and enforce quota policies. This is the basis for keeping the service operational and managing inter-operator roaming agreements.
4.15. Handling the Walk-by Users

In the case of community Wi-Fi, the network is an open network with the SSID visible to any wireless LAN device. This essentially creates a situation where any walk-by user’s mobile terminal automatically gets connected to the Wi-Fi network and results in a subscriber session creation. The user may not be having any intention in connecting to the Wi-fi network and in fact may not be using the mobile device, but the device gets attached to the network and a subscriber session and other network resources get locked up for that user session. The situation is especially worse in public hotspots such as train stations, or Airports where there is high traffic. This is important that this situation is correctly handled.

4.16. Overlapping IPv4 Address Support

The transition from IPv4 to IPv6 is a long process, and during this period of transition, the Wi-Fi operators will have to continue to offer IPv4 services. However, these operators may not have sufficient public IPv4 addresses for all the Wi-Fi devices in their network. For addressing this IPv4 exhausted issue, operators may have to leverage transitioning technologies such as NAT64, Dual-Stack Lite, 6rd or other approaches. These operators may also choose to segment the network into regions and two regions may use overlapped IPv4 address space to provide IPv4 services to users.

In a different scenario, a roaming user from a partners network, with an established mobility session with her home network, may be using a private IPv4 address and this IPv4 address may be overlapping with the address space that is being used in this access network. Furthermore, the IPv4 address space that is used for assignment to Wi-Fi subscribers should not conflict with the IPv4 addresses used on the cable/DSL transport network.

The Wi-Fi operator should be able to handle all these scenarios related to overlapping private IPv4 address usage.

4.17. Service Provisioning & Monitoring

Deployment of any community based Wi-Fi access will require additional Wi-Fi specific configuration on a per Residential Gateway basis. In order to support scalable deployment, the Service Providers should be able to provision these configuration options remotely. This remote provisioning frame work must support the following:
o Secure provisioning of the RG with community WiFi parameters to minimize the theft of service

o Ability to separate the private home subscriber traffic from the community WiFi traffic in the access network

o Privacy and protection of private Residential subscriber traffic from the community WiFi users

o Ability to remotely shut down an Residential Gateway which has been hijacked by hackers and is being used for DoS attacks.

o Ability to temporarily disable services for the community based WiFi support while maintaining service to the Residential fixed broadband subscriber

o Seamless integration of the WiFi provisioning aspects of the Residential Gateway into the existing RG provisioning infrastructure implemented by the Fixed Broadband Providers

o Dynamic Service Monitoring Capability for managing the Wi-Fi Service.

5. Solution Approaches & Considerations

The following section identifies the different mobility approaches that Wi-Fi operator can leverage for deploying this Wi-Fi services.

5.1. PMIPv6 MAG on the RG: Layer-3 Encapsulation between CPE and Access Gateway

5.2. Ethernet-over-IP Support on the RG: Layer-2 Encapsulation between CPE and Access Gateway

5.3. Local Aggregation for Subscriber Control and Internet Offload

5.4. Mobility Chaining: Integration with Mobile Packet Core

6. IANA Considerations

This document does not require any IANA actions.

7. Security Considerations

This specification identifies the requirements for enabling Community
Wi-Fi Services over Residential architectures and the potential solution approaches for addressing those requirements. The security analysis for each of those requirements are covered in those respective sections.

8. Acknowledgements

The authors would like to thank Bill Choinski, John Coppola and Sangeeta Ramakrishnan for all the discussions related to Service Provider Wi-Fi Service requirements. The authors would also like to thank Byju Pularikkal for all the discussions and text contributions related to Lawful Interception and Service Provisioning.

9. References

9.1. Normative References


9.2. Informative References

[I-D.gundavelli-netext-multiple-apn-pmipv6]

[I-D.gundavelli-netext-pmipv6-wlan-applicability]

[I-D.ietf-netext-pmip6-qos]

[I-D.ietf-netext-pmipv6-sipto-option]
Gundavelli, S., Liebsch, M., and P. Seite, "PMIPv6 interworking with WiFi access authentication", draft-liebsch-netext-pmip6-authiwk-05 (work in progress), September 2012.


Authors' Addresses

Sri Gundavelli
Cisco
170 West Tasman Drive
San Jose, CA 95134
USA

Email: sgundave@cisco.com
Abstract

This document reports and discusses issues on IPv6 only networks and IPv4/IPv6 transition technologies through our experiences on the 3rd experiment on the WIDE camp. The 3rd experiment was held from September 3rd to September 6th, 2012. As well as past two experiments, we conducted face to face interview to participants for grasping IPv6 capability on users' devices, OSes, and applications. In addition to this, we explored solutions to mitigate timeout / fallback problems of IPv4/IPv6 dual stack clients on an IPv6 only network that is composed of DHCP6 and DNS64/NAT64.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on April 6, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.
This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.
Table of Contents

1. Introduction ........................................... 4
   1.1. History of 'Live with IPv6 experiments' on the WIDE camp ........................................... 4
       1.1.1. Summary of the 1st experiment ......................... 4
       1.1.2. Summary of the 2nd experiment ......................... 4
   1.2. Abstract of the 3rd experiment .......................... 6
   1.3. Requirements Language .................................... 7

2. Technology and Terminology ............................... 7

3. Basic configuration of Network and Experiments .......... 8

4. Experiments ............................................. 11
   4.1. An Experiment in RA method .............................. 11
       4.1.1. Details of Network Configuration ....................... 11
       4.1.2. User Survey ........................................... 12
           4.1.2.1. Client Profile ................................... 12
           4.1.2.2. Behaviors of DHCP6 Clients ......................... 13
           4.1.2.3. Timeout / Fallback Problems ....................... 14
   4.2. Experiments in DHCP-PD method .......................... 14
       4.2.1. Basic Network Configuration ......................... 14
       4.2.2. Experiment 0 ........................................... 15
           4.2.2.1. Waiting timeout of DHCP4 in Windows 7 ........... 16
           4.2.2.2. Long TCP fallback in Mac OS X Lion and Mountain Lion ........................................... 16
           4.2.2.3. Incompletion of network settings in iOS 5 ........ 16
           4.2.2.4. Incapability of IPv6 DNS settings by DHCP6 ........ 16
       4.2.3. Experiment 1 ........................................... 17
           4.2.3.1. Diff of network settings .......................... 17
           4.2.3.2. Result ............................................ 18
       4.2.4. Experiment 2 ........................................... 19
           4.2.4.1. Diff of network settings .......................... 19
           4.2.4.2. Result ............................................ 21
       4.2.5. Experiment 3 ........................................... 21
           4.2.5.1. Diff of network settings .......................... 21
           4.2.5.2. Result ............................................ 22
   5. Conclusion .............................................. 23

6. Security Considerations .................................. 24

7. IANA Considerations ..................................... 24

8. References ............................................... 24
   8.1. Normative References ................................... 24
   8.2. Informative References ................................ 25

Appendix A. Acknowledgments .................................. 27
Authors' Addresses .......................................... 27
1. Introduction

This document reports and discusses issues on IPv6 only networks and IPv4/IPv6 transition technologies through our experiences on the 3rd experiment on the WIDE camp. The 3rd experiment was held from September 3rd to September 6th in Matsushiro Royal Hotel, Nagano, Japan, where is the same hotel of the 1st and 2nd experiments.

1.1. History of "Live with IPv6 experiments" on the WIDE camp

"Live with IPv6 experiment" aims to evaluate commercial IPv6 network services, the availability of IPv6 networks with several IPv4 / IPv6 translation / encapsulation technologies by actual users' experiences, and to grasp issues on IPv4 exhaustion situation or IPv4 / IPv6 transition. These experiments are based on an assumption that ISP backbone networks will be constructed on IPv6 only and end customer will have to use an IPv6 network with 64 translators or an IPv4 network with 464 translators to keep current usage of the Internet services.

1.1.1. Summary of the 1st experiment

The 1st experiment was held in Matsushiro Royal Hotel from September 6th to September 9th, 2011 with 153 participants, and the experiment result was reported in the v6ops BoF on IETF 82 Taipei. In the 1st experiment, we constructed an IPv6 only network with stateless NAT64 and DNS64 as a part of the WIDE backbone through IPv6 L2TP over a commercial IPv6 network service. The commercial IPv6 network service was provided by NTT-East as an Access Carrier, Internet MultiFeed (MFeed) as a Virtual Network Enabler (VNE) and IIJ as an IPv6 Internet Service Provider (IPv6 ISP). In addition to an IPv6 connectivity with NAT64/DNS64, we also tested a SA46T [I-D.draft-matsuhira-sa46t-spec] based IPv4 global network service and a murakami-4RD [I-D.draft-murakami-softwire-4rd] based IPv4 private network service (murakami-4RD is now merged into MAP [I-D.draft-ietf-softwire-map-02]). With referring IETF’s IPv6 only network experiences [RFC6586], we reported several new issues on an IPv6 only network with IPv4 / IPv6 transition technologies, especially on inappropriate DNS replies mentioned in [RFC4074], on MTU mismatch, on VPN protocols and applications through IPv4 / IPv6 translators.

1.1.2. Summary of the 2nd experiment

According to the experiences on the 1st experiment, the 2nd experiment was conducted from March 5th to March 8th, 2012 in Matsushiro Royal Hotel, the same hotel of the 1st experiment. 171 participants joined this 2nd experiment, most of them were engineers.
or academic people. The 2nd experiment result was reported in the v6ops BoF on IETF 83 Paris.

The settings of the core network in the 2nd experiment was same as the 1st experiment. In the 1st experiment, a commercial IPv6 network service was employed as a backbone network, in other word, we did evaluate the availability of commercial IPv6 network services from the view of home users. Therefore, the evaluation target of the 2nd experiment was planned as living in commercial IPv6 networks with IPv4 / IPv6 translation technologies or IPv4 / IPv6 translation services.

The user access networks of the 2nd experiment were achieved by two types of commercial IPv6 network services through the NTT NGNv6 access network, with four kinds of IPv4 / IPv6 translation technologies. One of the two commercial IPv6 network services was /48 prefix IPv6 network service through IPoE[RFC0894] on NTT NGNv6 (we name it "native IPoE" in this draft), the other was /56 prefix IPv6 network service through PPPoE[RFC2516] on NTT NGNv6 (we label it "native PPPoE" in this draft) [YasudaAPRICOT2011]. Both IPv6 networks were served from NTT-East, MFeed and IIJ as same as the 1st experiment.

Usually, IPv6 networks on both native IPoE and native PPPoE were provided with only DNS v6 proxy. We constructed DNS64/NAT64 service on the WIDE backbone and on the camp core network, and served it through stateless DHCP6 [RFC3736] both on native IPoE and on native PPPoE.

Along with the DNS64/NAT64 translation service, for aiming to evaluate more practical approaches on the current commercial environments, we tested three IPv4 services over IPv6 networks, murakami-4RD [I-D.draft-murakami-softwire-4rd], SA46T [I-D.draft-matsuhira-sa46t-spec] and 464XLAT [I-D.draft-ietf-v6ops-464xlat]. We mainly served seven IP networks to participants by combination of those networks and translation services, that is, native IPoE with DNS64/NAT64, native PPPoE with DNS64/NAT64, murakami-4RD on both IPoE and PPPoE, 464XLAT on both IPoE and PPPoE, SA46T on PPPoE.

Three evaluations were mainly conducted by the evaluation team, i) user survey about the availability of each network through face to face interview, ii) analysis of DNS behaviors to grasp inappropriate behaviors mentioned in [RFC4074], iii) availability test of VPN applications to analyze MTU problems For to grasp whether an unavailability of VPN applications was intentional one due to the specification of a translation technology or not. Also, Konami Digital Entertainment (KDE) joined in this experiment, and evaluated
Internet-Draft      IPv6 Experiments in the WIDE Camp       October 2012

NAT/Firewall traversal testing on each IPv6 network or each
translator service from the view of commercial (P2P) Network Game
services. KDE gave us the importance / requirements of hair-pinning
functions and of MTU / packet fragmentation handling on NAT/NAPT for
P2P based Multiplayer Online Games.

1.2. Abstract of the 3rd experiment

The 3rd experiment was conducted from September 3rd to September 6th,
2012 in Matsushiro Royal Hotel, the same hotel of the past two
experiments. 136 participants joined this 3rd experiment, most of
them were engineers or academic people.

The aims of 3rd experiments were 1) continuous user survey on IPv6
capability of devices, OSes and applications, 2) exploration of a
practical solution to mitigate timeout / fallback problems of IPv4/
IPv6 dual stack clients on an IPv6 only network.

The first aim was conducted to grasp the IPv6 capability of users’
devices, OSes, and applications and to collect users’ experiences
through face to face interview. From the 2nd experiments, several
new OSes or new devices have been released. Through this continuous
survey, we saw the current development / deployment strategy of IPv6
on commercial vendors or Telecom / Internet Service providers. This
user survey was mainly carried on September 3rd and September 4th.

The second aim was derived from our experiences of an IPv6 only
network with DHCP6/DNS64/NAT64 on past two experiments. In past two
experiments, various OSes met several timeout / fallback problems, in
the initial connection setting through Wi-Fi settings, in the name
server selection, in the establishment of a TCP connection. Most
OSes and applications, that met tedious timeout / fallback problems,
preferred IPv4 to IPv6, or required IPv4 settings to enable IPv6
settings. These timeout / fallback problems were seemed to be
derived from an assumption that there are no IPv6 only network on the
current situation.

Toward the sunset of IPv4, we have to explore and achieve a practical
solution to move from IPv4/IPv6 dual stack networks to IPv6 only
networks without giving stress or difficulties to end users. In
IPv4/IPv6 transition situation, end users will usually use IPv4/IPv6
dual stack mode, and they will leave all IPv4 / IPv6 network settings
by OSes’ auto configuration behaviors on their devices except for
selecting Wi-Fi connections.

We focused on testing an IPv6 only network that was basically
composed of DHCP6, DNS64 and NAT64. In this IPv6 only network, we
sought a current practice of timeout / fallback mitigation among
IPv4/IPv6 dual stack networks and IPv6 only networks. According to results of the user survey, we added several functions to a basic DHCP6/DNS64/NAT64 network in step by step fashion, and we analyzed or revised mitigation methods for timeout / fallback problems.

This draft is composed of following sections. We explain the overview of the network settings in the 3rd experiment at first. Next, we report the result of the user survey. Then, we describe the experiment on timeout / fallback mitigation methods. Finally, we summarize our practical timeout / fallback mitigation method. We also mention about limitations our mitigation method and our recommendations on development / deployment of IPv6 capability on end clients.

1.3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2.  Technology and Terminology

In this document, the following terms are used. "NAT44" refers to any IPv4-to-IPv4 network address translation algorithm, both "Basic NAT" and "Network Address/Port Translator (NAPT)", as defined by [RFC2663].

"Dual Stack" refers to a technique for providing complete support for both Internet protocols -- IPv4 and IPv6 -- in hosts and routers [RFC4213].

"NAT64" refers to a Network Address Translator - Protocol Translator defined in [RFC6052], [RFC6144], [RFC6145], [RFC6146], [RFC6384].

"DNS44" refers DNS extensions to use NAT44 translation from IPv6 clients to IPv4 servers with name resolution mechanisms that is defined in [RFC6147].

"DHCP4" refers Dynamic Host Configuration Protocol for IPv4 that is defined in [RFC4213].

"DHCP6" refers Dynamic Host Configuration Protocol for IPv6. So called "Stateful DHCP6" is defined in [RFC3315] and "Stateless DHCP6" is defined in [RFC3736]. "DHCP-PD" or "DHCPv6 Prefix Delegation" refers IPv6 Prefix Options for DHCP6 that is initially defined in [RFC3633] and updated in [RFC6603].
"ND" refers Neighbor Discovery for IP version 6 (IPv6) that is defined in [RFC4861] and updated in [RFC5942].

3. Basic configuration of Network and Experiments

The WIDE Camp Autumn 2012 was held at Matsushiro Royal Hotel in Nagano Prefecture of Japan, the same place of the 1st and 2nd experiment, from September 3rd to September 6th, 2012. Figure 1 shows the overview of the whole network topology on the WIDE Camp Autumn 2012.

Besides our IPv6 only experiments, the camp NOC team set up a core network (camp-net-core) for preparing a backup plan of our IPv6 only network experiments and for conducting other experiments such as OLSR emulation, SA46T-AT [I-D.draft-matsuhira-sa46t-at-00] and NAT44 double translation, and measurement of a satellite link. All server instances and routing instances of the core network were built on StarBED that is a cloud / network emulation testbed in Japan. We constructed two layer 2 tunnels between StarBED and Matsushiro Royal hotel through IPv4 PPPoE. The layer 2 tunnels over IPv4 PPPoE were constructed by NEC IX2015. The OLSR network and the satellite link were served as IPv4 / IPv6 dual stack networks. The wireless Accesses to these networks were provided by CISCO Systems Mesh Wi-Fi Access Point and WLC (Wireless LAN Controller).

As well as our 2nd experiment, a commercial IPv6 service was employed to achieve our IPv6 only network experiments. The Access Carrier (AC), the Virtual Network Enabler (VNE) and the IPv6 Internet Service Provider (v6ISP) of this 3rd experiment were same combination of past experiments, that is, NTT-East as AC, MFeed as the VNE and IIJ Mio as v6ISP. We contracted two external FTTH lines by NTT NGNv6 IPoE method. We changed the IPv6 address allocation method on NTT NGNv6 IPoE during this camp.

From September 2th (the preparation day) to September 4th, we used the RA method for the external connectivity. Figure 2 represents details of the IPv6 only network by RA method. From September 5th to September 6th, we changed the external connectivity to the DHCP-PD method.

In the RA method, we tested the DHCP6 client behaviors when two stateless DHCP6 servers exist, one is placed by the VNE or ISP to indicate AAAA name servers, the other is located in the local subnet to lead clients to a DNS64 name server. On the other hand, we explored mitigation methods for timeout / fallback problems after we changed the external connectivity to the DHCP-PD method. We explain the experiment on the RA method in Section 4.1 and the experiments on
the DHCP-PD method in Section 4.2, respectively.

We employed following implementations for key components;

- DNS64 and recursive cache server: NLNet Labs Unbound 1.4.7 with DNS64 patch
- NAT64 : OpenBSD 5.1 PF (Packet Filter)
- DHCP-PD client : WIDE DHCP client (dhcp6c)
- Stateless DHCP6 server : Alaxala 3630
Over view of the 2nd experiment topology

Figure 1
4. Experiments

4.1. An Experiment in RA method

4.1.1. Details of Network Configuration

The experiment conducted in RA method was overwriting client DNS information by a local stateless DHCP6 server. Figure 2 shows the test network topology. The RA method provided /64 prefix addresses and routing information through RA. The RA was set managed flag as zero (M flag == 0) and the other flag to one (O flag == 1) to let clients query to stateless DHCP6 servers. In this case, a stateless DHCP6 server was placed on the VNE network of MFeed and IIJ that advertised two AAAA name servers. Those two AAAA name servers returned only AAAA records to any queries.

We wanted to inform only the DNS64 IPv6 address to clients on this RA method while using address assignment and default route settings by the RA method. Of course, we could not control the DHCP6 server on the VNE network. Therefore, we tried to use the preference option of DHCP6.

The preference option of DHCP6 (section 22.8 of [RFC3315]) defines that "the Preference option is sent by a server to a client to affect the selection of a server by the client". Section 17.1.3 of [RFC3315] defines the criteria on the behavior of DHCP6 server selection by a client when the client has received two or more valid advertise messages;

- Those Advertise messages with the highest server preference value are preferred over all other Advertise messages.
- Within a group of Advertise messages with the same server preference value, a client MAY select those servers whose Advertise messages advertise information of interest to the client. For example, the client may choose a server that returned an advertisement with configuration options of interest to the client.
- The client MAY choose a less-preferred server if that server has a better set of advertised parameters, such as the available addresses advertised in IAs.

We assumed we could overwrite the name server information by sending advertise messages with highest preference value from a local stateless DHCP6 server. Thus, we placed a local stateless DHCP6 server shown in Figure 2.
This overwriting was partially succeeded as well as we assumed, however, several inconveniences were reported through face to face interview and inspection by the special observation team.

---

The Test Topology on RA method

Figure 2

4.1.2. User Survey

59 participants (42.8 %) replied our face to face interview. We show the client profile in Section 4.1.2.1 and reported troubles in Section 4.1.2.2 and Section 4.1.2.3.

4.1.2.1. Client Profile

94 unique devices were profiled. The distribution of the pair of device and OS were shown in Table 1.
<table>
<thead>
<tr>
<th>Device Type</th>
<th>OS Type</th>
<th># of devices (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC/AT Note PC</td>
<td>Windows 7</td>
<td>16 (17.0 %)</td>
</tr>
<tr>
<td>PC/AT Note PC</td>
<td>NetBSD</td>
<td>2 (2.1 %)</td>
</tr>
<tr>
<td>PC/AT Note PC</td>
<td>Linux</td>
<td>4 (4.3 %)</td>
</tr>
<tr>
<td>Apple Note PC</td>
<td>Mountain Lion</td>
<td>15 (16.0 %)</td>
</tr>
<tr>
<td>Apple Note PC</td>
<td>Lion</td>
<td>18 (19.1 %)</td>
</tr>
<tr>
<td>Apple Note PC</td>
<td>Snow Leopard</td>
<td>9 (9.6 %)</td>
</tr>
<tr>
<td>Apple Note PC</td>
<td>Windows 7 (Bootcamp)</td>
<td>3 (3.2 %)</td>
</tr>
<tr>
<td>iPhone / iPod</td>
<td>iOS 5</td>
<td>9 (9.6 %)</td>
</tr>
<tr>
<td>Android Phone</td>
<td>Android OS 4</td>
<td>3 (3.2 %)</td>
</tr>
<tr>
<td>Android Phone</td>
<td>Android OS 2</td>
<td>4 (4.3 %)</td>
</tr>
<tr>
<td>Android Phone</td>
<td>Android OS 1</td>
<td>1 (1.0 %)</td>
</tr>
<tr>
<td>iPad</td>
<td>iOS 6</td>
<td>1 (1.0 %)</td>
</tr>
<tr>
<td>iPad</td>
<td>iOS 5</td>
<td>6 (6.4 %)</td>
</tr>
<tr>
<td>Android Tablet</td>
<td>Android OS 4</td>
<td>2 (2.1 %)</td>
</tr>
<tr>
<td>Kindle</td>
<td>Kindle 3.3</td>
<td>1 (1.0 %)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>94</td>
</tr>
</tbody>
</table>

Table 1: The distributions of devices of participants

4.1.2.2. Behaviors of DHCP6 Clients

Many users reported inconveniences of DHCP6 client behaviors in the RA method. We focused on the analysis of DHCP6 client behavior of Windows 7 and of Mac OS X Lion / Mountain Lion. Both Windows 7 and Mac OS X usually stored DNS64 IPv6 address to their name server information, however, both of them sometime stored two AAAA name servers on the VNE network. Differences of their DHCP6 client behaviors were as follows;
In most cases, Windows 7 preferred to the advertise message from the local DHCP6 server that indicated the DNS64 server, however, it often preferred the advertise message from the DHCP6 server on the VNE network at the RA refresh timing.

When the DHCP6 client preferred to the DHCP6 server on the VNE network, an user had to reset the Wi-Fi device of his/her PC and to reconnect to the Wi-Fi network. "ipconfig /renew" or simply reconnecting by Wi-Fi selection icon often failed to prefer the advertise message from the local DHCP6 server.

On the other hand, Mac OS X Lion and Mountain Lion often failed to prefer the advertise message from the local DHCP6 server at the initial set up on Wi-Fi setting, however, "Renew DHCP lease" on the detail of network settings always preferred to the advertise message from the local DHCP6 server, that is, Mac OS X always changed the name server setting to only DNS64 IPv6 address by "Renew DHCP lease". At RA refresh timing, Mac OS X sometime preferred to the DHCP6 server on the VNE network, then, the user had to refresh DHCP configurations again.

4.1.2.3. Timeout / Fallback Problems

Many users reported inconveniences due to timeout / fallback problems. Root causes were roughly categorized into 1) troubles of DNS64, 2) incapability of IPv6 and of DNS64 on various servers and applications mentioned in [RFC4074] and [RFC6586], 3) incapability of DHCP6 client and / or IPv4 dependency on OSes. In Section 4.2, we explain the detail of timeout / fallback problems without effects by the selection of stateless DHCP6 servers.

4.2. Experiments in DHCP-PD method

On the contrary of the RA method mentioned in Section 4.1, the DHCP-PD method provided /56 prefix delegation by DHCP6 prefix delegation mechanism. We settled a DHCP-PD client PC router and set up static routes to two delegated /64 networks, one was labeled as "v6only-basic", the other was named as "v6only-fallback". The v6only-basic network was a basic IPv6 only network that was composed of stateless DHCP6, DNS64 and NAT64. On the other hand, we tested several timeout / fallback mitigation methods in "v6only-fallback". Figure 3 shows the basic network topology of experiments on DHCP-PD method.

4.2.1. Basic Network Configuration

Figure 3 shows the basic network topology of experiments on DHCP-PD method.
4.2.2. Experiment 0

In the experiment 0, we observed OSes behaviors again. Actually, the inconvenience on the selection of two stateless DHCP6 servers were resolved by DHCP-PD and placing one stateless DHCP6 server onto each /64 prefix subnet. However, we clearly recognized several timeout / fallback problems. In following sections, we explain timeout / fallback problems due to DHCP6 client incapability and IPv4
dependency of OSes.

4.2.2.1. Waiting timeout of DHCP4 in Windows 7

In Windows 7, timeout of DHCP4 queries spent a few minutes in the initial Wi-Fi connection setup. After fallback on the initial Wi-Fi connection, there were no problem on using IPv6 capable applications. DNS64 fallback failures due to the inappropriate authoritative servers still occurred, however, several authoritative servers, that returned inappropriate AAAA reply in past experiments, had been fixed to have appropriate fallback.

4.2.2.2. Long TCP fallback in Mac OS X Lion and Mountain Lion

Mac OS X implementations, such as Lion and Mountain Lion, had more serious timeout / fallback problems than Windows 7. After the timeout of DHCP4 queries with a few minutes as well as Windows 7, the interface that is allocated IPv4 link local address was inserted as IPv4 default route. This Mac OS X behavior may be along with IPv4 on-link assumption in Section 3.3 of [RFC3927]. Section 3.3 of RFC3927 mentions "Interaction with Hosts with Routable Addresses", which assumes all IPv4 address are on-link at Link-Local configuration.

Also, getaddrinfo implementation on Mac OS X did HappyEyeball like behavior. The getaddrinfo of Mac OS X returned an IP address list where IPv4 addresses were inserted the top of the list initially. Combining the on-link-assumption and the HappyEyeball like getaddrinfo caused long long TCP fallback from IPv4 to IPv6 in the initial TCP connection setup. Once the long long TCP fallback occurred, getaddrinfo of Mac OS X marked some flag that IPv4 is not available at the moment, then the getaddrinfo gave higher priority to IPv6 addresses than IPv4 addresses until ARP and / or ND tables were refreshed. When ARP and / or ND tables were refreshed, Mac OS X users face long long TCP fallback from IPv4 to IPv6 again.

4.2.2.3. Incompletion of network settings in iOS 5

In iOS 5, "Network Setting" were not completed, "Network Settings" will be completed only if IPv4 address, IPv4 router, and IPv4 DNS can be retrieved via DHCPv4 or manually configured all of these 3.

4.2.2.4. Incapability of IPv6 DNS settings by DHCP6

Windows XP, older Mac OS X (Snow Leopard and older) and Android OS required an IPv4 address for an DNS server even when they can use IPv6. In an IPv6 only network, DNS information should be gotten via DHCP6, these OSes did not support DHCP6 client. Also, Android cannot
be configured to use DNS over IPv6 even in manual configuration.

4.2.3. Experiment 1

4.2.3.1. Diff of network settings

In the Experiment 1, we added a DHCP4 server that provided only IPv4 private address to DHCP4 client without the default gateway IPv4 address nor IPv4 address of DNS. We employed ISC-DHCP for this DHCP4 server.
4.2.3.2. Result

As the result of Experiment 1, only timeout of DHCP4 was solved, that is, only Windows 7 was working well without any fallback problems except for DNS64 name resolving. TCP fallback problem on MacOS X still occurred. iOS applications were sometimes working, but periodically failed due to retrying Wi-Fi connection setup.
4.2.4. Experiment 2

4.2.4.1. Diff of network settings

In the Experiment 2, we put BIND9 forwarder on-link and configured DHCP4/6 to use this DNS. We configured BIND9 forwarder with: * deny-answer-addresses { 0.0.0.0/0; }; * which directed that no IPv4 address answer should be trusted. It returned SERVFAIL to resolvers.
Test Topology on Experiment 2 (v6only-fallback)

Figure 5
4.2.4.2. Result

As result of Experiment 2, Android was working well. iOS was working, but periodically failed due to retrying to Wi-Fi connection setup. MacOS X variants were working, but timeout by TCP fallback still occurred. Windows XP was not working because all DNS queries failed due to SERVFAIL.

4.2.5. Experiment 3

4.2.5.1. Diff of network settings

In the Experiment 3, we hacked AAAA filtering code on BIND9 to filter "A records" instead of "AAAA records" both on IPv4/IPv6 transport. We put BIND9 above to the local link, which was configured to forward all queries to DNS64. We also configured DHCP4/DHCP6 to use the DNS proxy.
4.2.5.2. Result

As the result of Experiment 3, Windows XP, MacOS X variants, iOS, Android were working well. Some of applications still failed on IPv6 only due to the IPv6 incapability or DNS64 fallback problem, but many
cases were fine: IE/Safari/Chrome/Firefox, Twitter, Facebook, Instagram, and so on.

Remaining issues were connection failures during a few minutes after Wi-Fi was connected. We guess the possible reason of this failures as follows: RS (Router Solicitation) was sent from kernel before Wi-Fi link was established. No IPv6 address was obtained until periodical RA (Router Advertisement) was received. The possible workaround to this connection failure is shortening RA interval to 5-10 seconds (though it disturb Wi-Fi ...) or detecting association through AP log and kicking RS or RA.

5. Conclusion

Timeout / fallback problems on IPv4/IPv6 dual stack clients in an IPv6 only network are caused by

1. timeout and fallback sequence on DHCP4 queries,
2. timeout and fallback sequence on the connectivity check to the IPv4 internet after the DHCP4 auto configuration,
3. connection retry sequence when the connectivity to the IPv4 internet was not given.
4. timeout and fallback sequence of a TCP connection on Mac OS X variants due to their HappyEyeball like behavior of getaddrinfo,
5. preference / dependency of IPv4 on name resolution,
6. connection failures during 1-2 minutes after Wi-Fi was connected.

To mitigate these timeout / fallback problems, our current practice is composed of following components;

- Configure a DNS64 and a NAT64 in somewhere.
- Configure a Dual-stack DNS proxy as follows

  * The DNS proxy forwards all queries to the DNS64 except "A" query type (IPv4 address). Since there is no IPv4 connectivity on the client, all queries to "A" should be filtered and the DNS proxy returns NO DATA, just like "AAAA" filtering.
  * This "A" filter should be enabled both on IPv4 and IPv6 transport.
* This Dual-stack "A" filter DNS proxy should be "on-link" and reachable from IPv4/IPv6 dual stack mode clients.

- Configure a DHCP4 server to reply a private IPv4 address, an IPv4 gateway router, and IPv4 address of an "A" filter DNS proxy to DHCP4 client.

- Configure a DHCP6 server to indicate the IPv6 address of "A" filter DNS Proxy to DHCP6 client.

* The IPv6 address of "A" filter DNS Proxy may be provided to IPv4/IPv6 dual stack mode clients by RDNSS [RFC6106]. However, from our experience on hot stage of Camp 1209 Autumn, Mac OS X Lion and Mountain Lion could handle RDNSS, but Windows 7 did not handle RDNSS.

* Only one DHCP6 server should be placed in each /64 prefix segment or indicated by DHCP6 relay. According to our experience, we do not recommend overwriting DNS information by a local stateless DHCP6 server with highest preference value due to the differences of handling multiple DHCP6 replies among DHCP6 client implementations.

- Configure the IPv4 gateway router not to forward any IPv4 packets.

6. Security Considerations

As well as Arkko mentioned in [RFC6586], the use of IPv6 instead of IPv4 by itself does not make a big security difference. In our experience, we only set up following security functions; the access control list on routers / servers, accounting on the wireless network access.

7. IANA Considerations

This document has no IANA implications.

8. References

8.1. Normative References


[RFC2663]  Srisuresh, P. and M. Holdrege, "IP Network Address

8.2. Informative References

[I-D.draft-ietf-softwire-map-02]

[I-D.draft-ietf-v6ops-464xlat]
Mawatari, M., Kawashima, M., and C. Byrne, "464XLAT: Combination of Stateful and Stateless Translation", September 2012, <draft-ietf-v6ops-464xlat-08 (work in progress)>.

[I-D.draft-matsuhira-sa46t-at-00]

[I-D.draft-matsuhira-sa46t-spec]

[I-D.draft-murakami-softwire-4rd]


Appendix A. Acknowledgments

Here, we thank to all the participants of WIDE camp on the experiments. We also say thank you to whom serving implementations and services in the Matsushiro Royal Hotel.

R. Nakamura of Univ. of Tokyo, Y. Ueno of Keio Univ. and R. Shouhara of Univ. of Tokyo for helping us on the base settings of the IPv6 only experiments and merging into the camp-net.

O. Onoe of Sony Corporation for his deep inspection and testing of end node devices.


Y. Atarashi of Alaxala Networks and R. Atarashi of IIJ Innovation Institute for designing the items of face to face interview and analyzing user survey data.

Authors’ Addresses

Hiroaki Hazeyama
NAIST
Takayama 8916-5
Nara,
Japan

Phone: +81 743 72 5216
Email: hiroa-ha@is.naist.jp
Ruri Hiromi  
Intec Inc.  
1-3-3 Shin-Suna, Koutou  
Tokyo,  
Japan  
Email: hiromi@inetcore.com

Tomohiro Ishihara  
Univ. of Tokyo  
3-8-1 Komaba, Meguro  
Tokyo,  
Japan  
Email: sho@c.u-tokyo.ac.jp

Osamu Nakamura  
WIDE Project  
5322 Endo  
Kanagawa,  
Japan  
Email: osamu@wide.ad.jp
464XLAT: Combination of Stateful and Stateless Translation
draft-ietf-v6ops-464xlat-10

Abstract

This document describes an architecture (464XLAT) for providing limited IPv4 connectivity across an IPv6-only network by combining existing and well-known stateful protocol translation RFC 6146 in the core and stateless protocol translation RFC 6145 at the edge. 464XLAT is a simple and scalable technique to quickly deploy limited IPv4 access service to IPv6-only edge networks without encapsulation.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on August 27, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect
to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction ............................................. 3
2. Terminology ............................................. 3
3. Motivation and Uniqueness of 464XLAT ...................... 4
4. Network Architecture ...................................... 4
   4.1. Wireline Network Architecture ....................... 4
   4.2. Wireless 3GPP Network Architecture ................. 5
5. Applicability ............................................ 6
   5.1. Wireline Network Applicability ..................... 6
   5.2. Wireless 3GPP Network Applicability ............... 7
6. Implementation Considerations ............................ 7
   6.1. IPv6 Address Format .................................. 7
   6.2. IPv4/IPv6 Address Translation Chart ................. 7
   6.3. IPv6 Prefix Handling ................................ 9
   6.4. DNS Proxy Implementation ............................ 9
   6.5. CLAT in a Gateway .................................. 9
   6.6. CLAT to CLAT communications ........................ 9
7. Deployment Considerations ................................. 10
   7.1. Traffic Engineering .................................. 10
   7.2. Traffic Treatment Scenarios ........................ 10
8. Security Considerations ................................... 11
9. IANA Considerations ....................................... 11
10. Acknowledgements ........................................ 11
11. References .............................................. 11
   11.1. Normative References ............................... 11
   11.2. Informative References ............................ 12
Appendix A. Examples of IPv4/IPv6 Address Translation ........ 13
Authors’ Addresses .......................................... 14
1. Introduction

With the exhaustion of the unallocated IPv4 address pools, it will be difficult for many networks to assign IPv4 addresses to end users.

This document describes an IPv4 over IPv6 solution as one of the techniques for IPv4 service extension and encouragement of IPv6 deployment. 464XLAT is not a one-for-one replacement of full IPv4 functionality. The 464XLAT architecture only supports IPv4 in the client server model, where the server has a global IPv4 address. This means it is not fit for IPv4 peer-to-peer communication or inbound IPv4 connections. 464XLAT builds on IPv6 transport and includes full any-to-any IPv6 communication.

The 464XLAT architecture described in this document uses IPv4/IPv6 translation standardized in [RFC6145] and [RFC6146]. It does not require DNS64 [RFC6147] since an IPv4 host may simply send IPv4 packets, including packets to an IPv4 DNS server, which will be translated on the customer side translator (CLAT) to IPv6 and back to IPv4 on the provider side translator (PLAT). 464XLAT networks may use DNS64 [RFC6147] to enable single stateful translation [RFC6146] instead of 464XLAT double translation where possible. The 464XLAT architecture encourages the IPv6 transition by making IPv4 services reachable across IPv6-only networks and providing IPv6 and IPv4 connectivity to single-stack IPv4 or IPv6 servers and peers.

2. Terminology

**PLAT:** PLAT is Provider side translator(XLAT) that complies with [RFC6146]. It translates N:1 global IPv6 addresses to global IPv4 addresses, and vice versa.

**CLAT:** CLAT is Customer side translator(XLAT) that complies with [RFC6145]. It algorithmically translates 1:1 private IPv4 addresses to global IPv6 addresses, and vice versa. The CLAT function is applicable to a router or an end-node such as a mobile phone. The CLAT should perform IP routing and forwarding to facilitate packets forwarding through the stateless translation even if it is an end-node. The CLAT as a common home router or wireless Third Generation Partnership Project (3GPP) router is expected to perform gateway functions such as DHCP server and DNS proxy for local clients. The CLAT uses different IPv6 prefixes for CLAT-side and PLAT-side IPv4 addresses and therefore does not comply with the sentence "Both IPv4-translatable IPv6 addresses and IPv4-converted IPv6 addresses should use the same prefix." in Section 3.3 of [RFC6052]. The CLAT does not facilitate
communications between a local IPv4-only node and an IPv6-only node on the Internet.

3. Motivation and Uniqueness of 464XLAT

1. Minimal IPv4 resource requirements, maximum IPv4 efficiency through statistical multiplexing.

2. No new protocols required, quick deployment.

3. IPv6-only networks are simpler and therefore less expensive to operate than dual-stack networks.

4. Consistent native IP based monitoring, traffic engineering, and capacity planning techniques can be applied without the indirection or obfuscation of a tunnel.

4. Network Architecture

Examples of 464XLAT architectures are shown in the figures in the following sections.

Wireline Network Architecture can fit in the situations where there are clients behind the CLAT in the same way regardless of the type of access service, for example FTTH, DOCSIS, or WiFi.

Wireless 3GPP Network Architecture fits in the situations where a client terminates the wireless access network and may act as a router with tethered clients.

4.1. Wireline Network Architecture

The private IPv4 host on this diagram can reach global IPv4 hosts via translation on both CLAT and PLAT. On the other hand, the IPv6 host can reach other IPv6 hosts on the Internet directly without translation. This means that the CPE/CLAT can not only have the function of a CLAT but also the function of an IPv6 native router for native IPv6 traffic. The v4p host behind the CLAT on this diagram has [RFC1918] addresses.
4.2. Wireless 3GPP Network Architecture

The CLAT function on the User Equipment (UE) provides an [RFC1918] address and IPv4 default route to the local node network stack. The applications on the UE can use the private IPv4 address for reaching global IPv4 hosts via translation on both the CLAT and the PLAT. On the other hand, reaching IPv6 hosts (including host presented via DNS64 [RFC6147]) does not require the CLAT function on the UE.

Presenting a private IPv4 network for tethering via NAT44 and stateless translation on the UE is also an application of the CLAT.
5. Applicability

5.1. Wireline Network Applicability

When an Internet Service Provider (ISP) has IPv6 access service and provides 464XLAT, the ISP can provide outgoing IPv4 service to end users across an IPv6 access network. The result is that edge network growth is no longer tightly coupled to the availability of scarce IPv4 addresses.

If another ISP operates the PLAT, the edge ISP is only required to deploy an IPv6 access network. All ISPs do not need IPv4 access networks. They can migrate their access network to a simple and
highly scalable IPv6-only environment.

5.2. Wireless 3GPP Network Applicability

At the time of writing, in February 2013, the vast majority of mobile networks are compliant to Pre-Release 9 3GPP standards. In Pre-Release 9 3GPP networks, Global System for Mobile Communications (GSM) and Universal Mobile Telecommunications System (UMTS) networks must signal and support both IPv4 and IPv6 Packet Data Protocol (PDP) attachments to access IPv4 and IPv6 network destinations [RFC6459]. Since there are two PDPs required to support two address families, this is double the number of PDPs required to support the status quo of one address family, which is IPv4.

For the cases of connecting to an IPv4 literal or IPv4 socket that require IPv4 connectivity, the CLAT function on the UE provides a private IPv4 address and IPv4 default route on the host for the applications to reference and bind to. Connections sourced from the IPv4 interface are immediately routed to the CLAT function and passed to the IPv6-only mobile network, destined for the PLAT. In summary, the UE has the CLAT function that does a stateless translation [RFC6145], but only when required by an IPv4-only scenario such as IPv4 literals or IPv4-only sockets. The mobile network has a PLAT that does stateful translation [RFC6146].

464XLAT works with today’s existing systems as much as possible. 464XLAT is compatible with existing network based deep packet inspection solutions like 3GPP standardized Policy and Charging Control (PCC) [TS.23203].

6. Implementation Considerations

6.1. IPv6 Address Format

The IPv6 address format in 464XLAT is defined in Section 2.2 of [RFC6052].

6.2. IPv4/IPv6 Address Translation Chart

This chart offers an explanation about address translation architecture using a combination of stateful translation at the PLAT and stateless translation at the CLAT. The client on this chart is delegated an IPv6 prefix from a prefix delegation mechanism such as DHCPv6-PD [RFC3633], therefore it has a dedicated IPv6 prefix for translation.
Case of enabling only stateless XLATE on CLAT
6.3. IPv6 Prefix Handling

There are two relevant IPv6 prefixes that the CLAT must be aware of.

First, CLAT must know its own IPv6 prefixes. The CLAT should acquire a /64 for the uplink interface, a /64 for all downlink interfaces, and a dedicated /64 prefix for the purpose of sending and receiving statelessly translated packets. When a dedicated /64 prefix is not available for translation from DHCPv6-PD [RFC3633], the CLAT may perform NAT44 for all IPv4 LAN packets so that all the LAN originated IPv4 packets appear from a single IPv4 address and are then statelessly translated to one interface IPv6 address that is claimed by the CLAT via NDP and defended with DAD.

Second, the CLAT must discover the PLAT-side translation IPv6 prefix used as a destination of the PLAT. The CLAT will use this prefix as the destination of all translation packets that require stateful translation to the IPv4 Internet. It may discover the PLAT-side translation prefix using [I-D.ietf-behave-nat64-discovery-heuristic]. In the future some other mechanisms, such as a new DHCPv6 option, will possibly be defined to communicate the PLAT-side translation prefix.

6.4. DNS Proxy Implementation

The CLAT should implement a DNS proxy as defined in [RFC5625]. The case of an IPv4-only node behind the CLAT querying an IPv4 DNS server is undesirable since it requires both stateful and stateless translation for each DNS lookup. The CLAT should set itself as the DNS server via DHCP or other means and proxy DNS queries for IPv4 and IPv6 LAN clients. Using the CLAT enabled home router or UE as a DNS proxy is a normal consumer gateway function and simplifies the traffic flow so that only IPv6 native queries are made across the access network. DNS queries from the client that are not sent to the DNS proxy on the CLAT must be allowed and are translated and forwarded just like any other IP traffic.

6.5. CLAT in a Gateway

The CLAT feature can be implemented in a common home router or mobile phone that has a tethering feature. Routers with a CLAT feature should also provide common router services such as DHCP of [RFC1918] addresses, DHCPv6, NDP with RA, and DNS service.

6.6. CLAT to CLAT communications

464XLAT is a hub and spoke architecture focused on enabling IPv4-only services over IPv6-only networks. ICE [RFC5245] may be used to
support peer-to-peer communication within a 464XLAT network.

7. Deployment Considerations

7.1. Traffic Engineering

Even if the ISP for end users is different from the PLAT provider (e.g. another ISP), it can implement traffic engineering independently from the PLAT provider. Detailed reasons are below:

1. The ISP for end users can figure out IPv4 destination address from translated IPv6 packet header, so it can implement traffic engineering based on IPv4 destination address (e.g. traffic monitoring for each IPv4 destination address, packet filtering for each IPv4 destination address, etc.). The tunneling methods do not have such an advantage, without any deep packet inspection for processing the inner IPv4 packet of the tunnel packet.

2. If the ISP for end users can assign an IPv6 prefix greater than /64 to each subscriber, this 464XLAT architecture can separate IPv6 prefix for native IPv6 packets and the XLAT prefixes for IPv4/IPv6 translation packets. Accordingly, it can identify the type of packets ("native IPv6 packets" and "IPv4/IPv6 translation packets"), and implement traffic engineering based on the IPv6 prefix.

7.2. Traffic Treatment Scenarios

The below table outlines how different permutations of connectivity are treated in the 464XLAT architecture.

NOTE: 464XLAT double translation treatment will be stateless when a dedicated /64 is available for translation on the CLAT. Otherwise, the CLAT will have both stateful and stateless since it requires NAT44 from the LAN to a single IPv4 address and then stateless translation to a single IPv6 address.
## Traffic Treatment Scenarios

### Traffic Treatment Scenarios

<table>
<thead>
<tr>
<th>Server</th>
<th>Application and Host</th>
<th>Traffic Treatment</th>
<th>Location of Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPv6</td>
<td>IPv6</td>
<td>End-to-end IPv6</td>
<td>None</td>
</tr>
<tr>
<td>IPv4</td>
<td>IPv6</td>
<td>Stateful Translation</td>
<td>PLAT</td>
</tr>
<tr>
<td>IPv4</td>
<td>IPv4</td>
<td>464XLAT</td>
<td>PLAT/CLAT</td>
</tr>
</tbody>
</table>

8. Security Considerations

To implement a PLAT, see security considerations presented in Section 5 of [RFC6146].

To implement a CLAT, see security considerations presented in Section 7 of [RFC6145]. The CLAT may comply with [RFC6092].

9. IANA Considerations

This document has no actions for IANA.

10. Acknowledgements

The authors would like to thank JPIX NOC members, JPIX 464XLAT trial service members, Seiichi Kawamura, Dan Drown, Brian Carpenter, Rajiv Asati, Washam Fan, Behcet Sarikaya, Jan Zorz, Tatsuya Oishi, Lorenzo Colitti, Erik Kline, Ole Troan, Macke Chen, Gang Chen, Tom Petch, Jouni Korhonen, Bjoern A. Zeeb, Hemant Singh, Vizdal Ales, Mark ZZZ Smith, Mikael Abrahamsson, Tore Anderson, Teemu Savolainen, Alexandru Petrescu, Gert Doering, Victor Kuarsingh, Ray Hunter, James Woodyatt, Tom Taylor, and Remi Despres for their helpful comments. We also would like to thank Fred Baker and Joel Jaeggli for their support.

11. References

11.1. Normative References

11.2. Informative References

[I-D.ietf-behave-nat64-discovery-heuristic]
Savolainen, T., Korhonen, J., and D. Wing, "Discovery of the IPv6 Prefix Used for IPv6 Address Synthesis",
draft-ietf-behave-nat64-discovery-heuristic-13 (work in progress), November 2012.

BCP 5, RFC 1918, February 1996.


RFC 5245, April 2010.

BCP 152, RFC 5625, August 2009.

RFC 6092, January 2011.

RFC 6147, April 2011.

RFC 6459, January 2012.
Appendix A. Examples of IPv4/IPv6 Address Translation

The following is an example of IPv4/IPv6 Address Translation on the 464XLAT architecture.

In the case that an IPv6 prefix greater than /64 is assigned to an end user by such as DHCPv6-PD [RFC3633], the CLAT can use a dedicated /64 from the assigned IPv6 prefix.
Host & configuration value
+------------------------------+                  +------------------------------+
|           IPv4 server        |               | IP packet header               |
+------------------------------+                  +------------------------------+
|         [198.51.100.1]       |               | Destination IP address        |
|                  |               | [198.51.100.1]                |
|                  |               | Source IP address              |
|                  |               | [192.0.2.1]                    |
+------------------------------+                  +------------------------------+

PLAT
IPv4 pool address
[192.0.2.1 - 192.0.2.100]
PLAT-side XLATE IPv6 prefix
[2001:db8:1234::/96]

+------------------------------+
| Destination IP address   |
+------------------------------+

CLAT
PLAT-side XLATE IPv6 prefix
[2001:db8:1234::/96]
CLAT-side XLATE IPv6 prefix
[2001:db8:aaaa::/96]

+------------------------------+
| Destination IP address   |
+------------------------------+

IPv4 client
[192.168.1.2/24]

Delegated IPv6 prefix for client: 2001:db8:aaaa::/56
Authors’ Addresses

Masataka Mawatari
Japan Internet Exchange Co., Ltd.
KDDI Otemachi Building 19F, 1-8-1 Otemachi,
Chiyoda-ku, Tokyo  100-0004
JAPAN
Phone: +81 3 3243 9579
Email: mawatari@jpix.ad.jp

Masanobu Kawashima
NEC AccessTechnica, Ltd.
800, Shimomata
Kakegawa-shi, Shizuoka  436-8501
JAPAN
Phone: +81 537 22 8274
Email: kawashimam@vx.jp.nec.com

Cameron Byrne
T-Mobile USA
Bellevue, Washington  98006
USA
Email: cameron.byrne@t-mobile.com
Basic Requirements for IPv6 Customer Edge Routers
draft-ietf-v6ops-6204bis-12

Abstract

This document specifies requirements for an IPv6 Customer Edge (CE) router. Specifically, the current version of this document focuses on the basic provisioning of an IPv6 CE router and the provisioning of IPv6 hosts attached to it. The document also covers IP transition technologies. Two transition technologies in RFC 5969’s 6rd and RFC 6333’s DS-Lite are covered in the document. The document obsoletes RFC 6204, if approved.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on May 3, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of
Table of Contents

1. Introduction .......................................................... 3
   1.1. Requirements Language ........................................... 3
2. Terminology ............................................................ 3
3. Architecture ........................................................... 4
   3.1. Current IPv4 End-User Network Architecture ............... 4
   3.2. IPv6 End-User Network Architecture ....................... 5
      3.2.1. Local Communication .................................... 6
4. Requirements .......................................................... 7
   4.1. General Requirements ........................................... 7
   4.2. WAN-Side Configuration ....................................... 8
   4.3. LAN-Side Configuration ....................................... 12
   4.4. Transition Technologies Support ............................ 14
      4.4.1. 6rd ..................................................... 14
      4.4.2. Dual-Stack Lite (DS-Lite) ............................. 15
   4.5. Security Considerations ..................................... 16
5. IANA Considerations ................................................ 16
6. Acknowledgements ................................................... 17
7. Contributors ........................................................ 17
8. References ........................................................... 17
   8.1. Normative References ......................................... 17
   8.2. Informative References ..................................... 20
Appendix A. Changes from RFC 6204 .................................. 21
Authors’ Addresses .................................................... 22
1. Introduction

This document defines basic IPv6 features for a residential or small-office router, referred to as an IPv6 CE router, in order to establish an industry baseline for features to be implemented on such a router.

These routers typically also support IPv4.

Mixed environments of dual-stack hosts and IPv6-only hosts (behind the CE router) can be more complex if the IPv6-only devices are using a translator to access IPv4 servers [RFC6144]. Support for such mixed environments is not in scope of this document.

This document specifies how an IPv6 CE router automatically provisions its WAN interface, acquires address space for provisioning of its LAN interfaces, and fetches other configuration information from the service provider network. Automatic provisioning of more complex topology than a single router with multiple LAN interfaces is out of scope for this document.

See [RFC4779] for a discussion of options available for deploying IPv6 in service provider access networks.

The document also covers the IP transition technologies that were available at the time this document was written. Two transition technologies in 6rd [RFC5969] and DS-Lite [RFC6333] are covered in the document.

1.1. Requirements Language

Take careful note: Unlike other IETF documents, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are not used as described in RFC 2119 [RFC2119]. This document uses these keyword not strictly for the purpose of interoperability, but rather for the purpose of establishing industry-common baseline functionality. As such, the document points to several other specifications (preferable in RFC or stable form) to provide additional guidance to implementers regarding any protocol implementation required to produce a successful CPE router that interoperates successfully with a particular subset of currently deploying and planned common IPv6 access networks.

2. Terminology
3. Architecture

3.1. Current IPv4 End-User Network Architecture

An end-user network will likely support both IPv4 and IPv6. It is not expected that an end-user will change their existing network topology with the introduction of IPv6. There are some differences in how IPv6 works and is provisioned; these differences have implications for the network architecture. A typical IPv4 end-user
network consists of a "plug and play" router with NAT functionality and a single link behind it, connected to the service provider network.

A typical IPv4 NAT deployment by default blocks all incoming connections. Opening of ports is typically allowed using a Universal Plug and Play Internet Gateway Device (UPnP IGD) [UPnP-IGD] or some other firewall control protocol.

Another consequence of using private address space in the end-user network is that it provides stable addressing; i.e., it never changes even when you change service providers, and the addresses are always there even when the WAN interface is down or the customer edge router has not yet been provisioned.

Many existing routers support dynamic routing (which learns routes from other routers), and advanced end-users can build arbitrary, complex networks using manual configuration of address prefixes combined with a dynamic routing protocol.

3.2. IPv6 End-User Network Architecture

The end-user network architecture for IPv6 should provide equivalent or better capabilities and functionality than the current IPv4 architecture.

The end-user network is a stub network. Figure 1 illustrates the model topology for the end-user network.
This architecture describes the:

- Basic capabilities of an IPv6 CE router
- Provisioning of the WAN interface connecting to the service provider
- Provisioning of the LAN interfaces

For IPv6 multicast traffic, the IPv6 CE router may act as a Multicast Listener Discovery (MLD) proxy [RFC4605] and may support a dynamic multicast routing protocol.

The IPv6 CE router may be manually configured in an arbitrary topology with a dynamic routing protocol. Automatic provisioning and configuration are described for a single IPv6 CE router only.

3.2.1. Local Communication

Link-local IPv6 addresses are used by hosts communicating on a single link. Unique Local IPv6 Unicast Addresses (ULAs) [RFC4193] are used by hosts communicating within the end-user network across multiple links, but without requiring the application to use a globally routable address. The IPv6 CE router defaults to acting as the
demarcation point between two networks by providing a ULA boundary, a multicast zone boundary, and ingress and egress traffic filters.

At the time of this writing, several host implementations do not handle the case where they have an IPv6 address configured and no IPv6 connectivity, either because the address itself has a limited topological reachability (e.g., ULA) or because the IPv6 CE router is not connected to the IPv6 network on its WAN interface. To support host implementations that do not handle multihoming in a multi-prefix environment [MULTIHOMING-WITHOUT-NAT], the IPv6 CE router should not, as detailed in the requirements below, advertise itself as a default router on the LAN interface(s) when it does not have IPv6 connectivity on the WAN interface or when it is not provisioned with IPv6 addresses. For local IPv6 communication, the mechanisms specified in [RFC4191] are used.

ULA addressing is useful where the IPv6 CE router has multiple LAN interfaces with hosts that need to communicate with each other. If the IPv6 CE router has only a single LAN interface (IPv6 link), then link-local addressing can be used instead.

Coexistence with IPv4 requires any IPv6 CE router(s) on the LAN to conform to these recommendations, especially requirements ULA-5 and L-4 below.

4. Requirements

4.1. General Requirements

The IPv6 CE router is responsible for implementing IPv6 routing; that is, the IPv6 CE router must look up the IPv6 destination address in its routing table to decide to which interface it should send the packet.

In this role, the IPv6 CE router is responsible for ensuring that traffic using its ULA addressing does not go out the WAN interface, and does not originate from the WAN interface.

G-1: An IPv6 CE router is an IPv6 node according to the IPv6 Node Requirements [RFC6434] specification.

G-2: The IPv6 CE router MUST implement ICMPv6 according to [RFC4443]. In particular, point-to-point links MUST be handled as described in Section 3.1 of [RFC4443].
G-3: The IPv6 CE router MUST NOT forward any IPv6 traffic between its LAN interface(s) and its WAN interface until the router has successfully completed the IPv6 address and the delegated prefix acquisition process.

G-4: By default, an IPv6 CE router that has no default router(s) on its WAN interface MUST NOT advertise itself as an IPv6 default router on its LAN interfaces. That is, the "Router Lifetime" field is set to zero in all Router Advertisement messages it originates [RFC4861].

G-5: By default, if the IPv6 CE router is an advertising router and loses its IPv6 default router(s) and/or detects loss of connectivity on the WAN interface, it MUST explicitly invalidate itself as an IPv6 default router on each of its advertising interfaces by immediately transmitting one or more Router Advertisement messages with the "Router Lifetime" field set to zero [RFC4861].

4.2. WAN-Side Configuration

The IPv6 CE router will need to support connectivity to one or more access network architectures. This document describes an IPv6 CE router that is not specific to any particular architecture or service provider and that supports all commonly used architectures.

IPv6 Neighbor Discovery and DHCPv6 protocols operate over any type of IPv6-supported link layer, and there is no need for a link-layer-specific configuration protocol for IPv6 network-layer configuration options as in, e.g., PPP IP Control Protocol (IPCP) for IPv4. This section makes the assumption that the same mechanism will work for any link layer, be it Ethernet, the Data Over Cable Service Interface Specification (DOCSIS), PPP, or others.

WAN-side requirements:

W-1: When the router is attached to the WAN interface link, it MUST act as an IPv6 host for the purposes of stateless [RFC4862] or stateful [RFC3315] interface address assignment.

W-2: The IPv6 CE router MUST generate a link-local address and finish Duplicate Address Detection according to [RFC4862] prior to sending any Router Solicitations on the interface. The source address used in the subsequent Router Solicitation MUST be the link-local address on the WAN interface.
W-3: Absent other routing information, the IPv6 CE router MUST use Router Discovery as specified in [RFC4861] to discover a default router(s) and install default route(s) in its routing table with the discovered router's address as the next hop.

W-4: The router MUST act as a requesting router for the purposes of DHCPv6 prefix delegation ([RFC3633]).

W-5: The IPv6 CE router MUST use a persistent DHCP Unique Identifier (DUID) for DHCPv6 messages. The DUID MUST NOT change between network interface resets or IPv6 CE router reboots.

W-6: The WAN interface of the CE router SHOULD support a PCP client as specified in [I-D.ietf-pcp-base] for use by applications on the CE Router. The PCP client SHOULD follow the procedure specified in Section 8.1 of [I-D.ietf-pcp-base] to discover its PCP server. This document takes no position on whether such functionality is enabled by default or mechanisms by which users would configure the functionality. Handling PCP requests from PCP clients in the LAN side of the CE Router is out of scope.

Link-layer requirements:

WLL-1: If the WAN interface supports Ethernet encapsulation, then the IPv6 CE router MUST support IPv6 over Ethernet [RFC2464].

WLL-2: If the WAN interface supports PPP encapsulation, the IPv6 CE router MUST support IPv6 over PPP [RFC5072].

WLL-3: If the WAN interface supports PPP encapsulation, in a dual-stack environment with IPCP and IPV6CP running over one PPP logical channel, the Network Control Protocols (NCPs) MUST be treated as independent of each other and start and terminate independently.

Address assignment requirements:

WAA-1: The IPv6 CE router MUST support Stateless Address Autoconfiguration (SLAAC) [RFC4862].

WAA-2: The IPv6 CE router MUST follow the recommendations in Section 4 of [RFC5942], and in particular the handling of the L flag in the Router Advertisement Prefix Information option.
WAA-3: The IPv6 CE router MUST support DHCPv6 [RFC3315] client behavior.

WAA-4: The IPv6 CE router MUST be able to support the following DHCPv6 options: IA_NA, Reconfigure Accept [RFC3315], and DNS_SERVERS [RFC3646]. The IPv6 CE router SHOULD be able to support the DNS Search List DNSSL option as specified in [RFC3646].

WAA-5: The IPv6 CE router SHOULD implement the Network Time Protocol (NTP) as specified in [RFC5905] to provide a time reference common to the service provider for other protocols, such as DHCPv6, to use. If the CE router implements NTP, it requests the NTP Server DHCPv6 option [RFC5908] and uses the received list of servers as primary time reference, unless explicitly configured otherwise. LAN side support of NTP is out of scope for this document.

WAA-6: If the IPv6 CE router receives a Router Advertisement message (described in [RFC4861]) with the M flag set to 1, the IPv6 CE router MUST do DHCPv6 address assignment (request an IA_NA option).

WAA-7: If the IPv6 CE router does not acquire global IPv6 address(es) from either SLAAC or DHCPv6, then it MUST create global IPv6 address(es) from its delegated prefix(es) and configure those on one of its internal virtual network interfaces, unless configured to require a global IPv6 address on the WAN interface.

WAA-8: The CE router must support the SOL_MAX_RT option [I-D.droms-dhc-dhcpv6-solmaxrt-update] and request the SOL_MAX_RT option in an ORO.

WAA-9: As a router, the IPv6 CE router MUST follow the weak host (Weak ES) model [RFC1122]. When originating packets from an interface, it will use a source address from another one of its interfaces if the outgoing interface does not have an address of suitable scope.

WAA-10: The IPv6 CE router SHOULD implement the Information Refresh Time option and associated client behavior as specified in [RFC4242].

Prefix delegation requirements:
WPD-1: The IPv6 CE router MUST support DHCPv6 prefix delegation requesting router behavior as specified in [RFC3633] (IA_PD option).

WPD-2: The IPv6 CE router MAY indicate as a hint to the delegating router the size of the prefix it requires. If so, it MUST ask for a prefix large enough to assign one /64 for each of its interfaces, rounded up to the nearest nibble, and SHOULD be configurable to ask for more.

WPD-3: The IPv6 CE router MUST be prepared to accept a delegated prefix size different from what is given in the hint. If the delegated prefix is too small to address all of its interfaces, the IPv6 CE router SHOULD log a system management error. [RFC6177] covers the recommendations for service providers for prefix allocation sizes.

WPD-4: By default, the IPv6 CE router MUST initiate DHCPv6 prefix delegation when either the M or O flags are set to 1 in a received Router Advertisement (RA) message. Behavior of the CE router to use DHCPv6 prefix delegation when the CE router has not received any RA or received an RA with the M and the O bits set to zero is out of scope for this document.

WPD-5: Any packet received by the CE router with a destination address in the prefix(es) delegated to the CE router but not in the set of prefixes assigned by the CE router to the LAN must be dropped. In other words, the next hop for the prefix(es) delegated to the CE router should be the null destination. This is necessary to prevent forwarding loops when some addresses covered by the aggregate are not reachable [RFC4632].

(a) The IPv6 CE router SHOULD send an ICMPv6 Destination Unreachable message in accordance with Section 3.1 of [RFC4443] back to the source of the packet, if the packet is to be dropped due to this rule.

WPD-6: If the IPv6 CE router requests both an IA_NA and an IA_PD option in DHCPv6, it MUST accept an IA_PD option in DHCPv6 Advertise/Reply messages, even if the message does not contain any addresses, unless configured to only obtain its WAN IPv6 address via DHCPv6. See [I-D.ietf-dhc-dhcpv6-stateful-issues]
WPD-7: By default, an IPv6 CE router MUST NOT initiate any dynamic routing protocol on its WAN interface.

WPD-8: The IPv6 CE Router SHOULD support the [I-D.ietf-dhc-pd-exclude] PD-Exclude option.

4.3. LAN-Side Configuration

The IPv6 CE router distributes configuration information obtained during WAN interface provisioning to IPv6 hosts and assists IPv6 hosts in obtaining IPv6 addresses. It also supports connectivity of these devices in the absence of any working WAN interface.

An IPv6 CE router is expected to support an IPv6 end-user network and IPv6 hosts that exhibit the following characteristics:

1. Link-local addresses may be insufficient for allowing IPv6 applications to communicate with each other in the end-user network. The IPv6 CE router will need to enable this communication by providing globally scoped unicast addresses or ULAs [RFC4193], whether or not WAN connectivity exists.

2. IPv6 hosts should be capable of using SLAAC and may be capable of using DHCPv6 for acquiring their addresses.

3. IPv6 hosts may use DHCPv6 for other configuration information, such as the DNS_SERVERS option for acquiring DNS information.

Unless otherwise specified, the following requirements apply to the IPv6 CE router's LAN interfaces only.

ULA requirements:

ULA-1: The IPv6 CE router SHOULD be capable of generating a ULA prefix [RFC4193].

ULA-2: An IPv6 CE router with a ULA prefix MUST maintain this prefix consistently across reboots.

ULA-3: The value of the ULA prefix SHOULD be configurable.

ULA-4: By default, the IPv6 CE router MUST act as a site border router according to Section 4.3 of [RFC4193] and filter packets with local IPv6 source or destination addresses accordingly.
ULA-5: An IPv6 CE router MUST NOT advertise itself as a default router with a Router Lifetime greater than zero whenever all of its configured and delegated prefixes are ULA prefixes.

LAN requirements:

L-1: The IPv6 CE router MUST support router behavior according to Neighbor Discovery for IPv6 [RFC4861].

L-2: The IPv6 CE router MUST assign a separate /64 from its delegated prefix(es) (and ULA prefix if configured to provide ULA addressing) for each of its LAN interfaces.

L-3: An IPv6 CE router MUST advertise itself as a router for the delegated prefix(es) (and ULA prefix if configured to provide ULA addressing) using the "Route Information Option" specified in Section 2.3 of [RFC4191]. This advertisement is independent of having or not having IPv6 connectivity on the WAN interface.

L-4: An IPv6 CE router MUST NOT advertise itself as a default router with a Router Lifetime [RFC4861] greater than zero if it has no prefixes configured or delegated to it.

L-5: The IPv6 CE router MUST make each LAN interface an advertising interface according to [RFC4861].

L-6: In Router Advertisement messages ([RFC4861]), the Prefix Information option's A and L flags MUST be set to 1 by default.

L-7: The A and L flags' ([RFC4861]) settings SHOULD be user-configurable.

L-8: The IPv6 CE router MUST support a DHCPv6 server capable of IPv6 address assignment according to [RFC3315] OR a stateless DHCPv6 server according to [RFC3736] on its LAN interfaces.

L-9: Unless the IPv6 CE router is configured to support the DHCPv6 IA_NA option, it SHOULD set the M flag to 0 and the O flag to 1 in its Router Advertisement messages [RFC4861].

L-10: The IPv6 CE router MUST support providing DNS information in the DHCPv6 DNS_SERVERS and DOMAIN_LIST options [RFC3646].
L-11: The IPv6 CE router MUST support providing DNS information in the Router Advertisement Recursive DNS Server (RDNSS) and DNS Search List options. Both options are specified in [RFC6106].

L-12: The IPv6 CE router SHOULD make available a subset of DHCPv6 options (as listed in Section 5.3 of [RFC3736]) received from the DHCPv6 client on its WAN interface to its LAN-side DHCPv6 server.

L-13: If the delegated prefix changes, i.e., the current prefix is replaced with a new prefix without any overlapping time period, then the IPv6 CE router MUST immediately advertise the old prefix with a Preferred Lifetime of zero and a Valid Lifetime of either a) zero, or b) the lower of the current Valid Lifetime and two hours (which must be decremented in real time) in a Router Advertisement message as described in Section 5.5.3, (e) of [RFC4862].

L-14: The IPv6 CE router MUST send an ICMPv6 Destination Unreachable message, code 5 (Source address failed ingress/egress policy) for packets forwarded to it that use an address from a prefix that has been invalidated.

4.4. Transition Technologies Support

4.4.1. 6rd

6rd [RFC5969] specifies an automatic tunneling mechanism tailored to advance deployment of IPv6 to end users via a service provider’s IPv4 network infrastructure. Key aspects include automatic IPv6 prefix delegation to sites, stateless operation, simple provisioning, and service that is equivalent to native IPv6 at the sites that are served by the mechanism. It is expected that such traffic is forwarded over the CE Router’s native IPv4 WAN interface, and not encapsulated in another tunnel.

The CE Router SHOULD support 6rd functionality. If 6rd is supported, it MUST be implemented according to [RFC5969]. The following CE Requirements also apply:

6rd requirements:

6RD-1: The IPv6 CE router MUST support 6rd configuration via the 6rd DHCPv4 Option (212). If the CE router has obtained an IPv4 network address through some other means such as PPP, it SHOULD use the DHCPINFORM request message [RFC2131] to request the 6rd DHCPv4 Option. The IPv6 CE router MAY use other mechanisms to configure 6rd parameters. Such
mechanisms are outside the scope of this document.

6RD-2: If the IPv6 CE router is capable of automated configuration of IPv4 through IPCP (i.e., over a PPP connection), it MUST support user-entered configuration of 6rd.

6RD-3: If the CE router supports configuration mechanisms other than the 6rd DHCPv4 Option 212 (user-entered, TR-69, etc.), the CE router MUST support 6rd in "hub and spoke" mode. 6rd in "hub and spoke" requires all IPv6 traffic to go to the 6rd Border Relay. In effect, this requirement removes the "direct connect to 6rd" route defined in Section 7.1.1 of [RFC5969].

6RD-4: A CE router MUST allow 6rd and native IPv6 WAN interfaces to be active alone as well as simultaneously in order to support coexistence of the two technologies during an incremental migration period such as a migration from 6rd to native IPv6.

6RD-5: Each packet sent on a 6rd or native WAN interface MUST be directed such that its source IP address is derived from the delegated prefix associated with the particular interface from which the packet is being sent [Section 4.3 [RFC3704]].

6RD-6: The CE router MUST allow different as well as identical delegated prefixes to be configured via each (6rd or native) WAN interface.

6RD-7: In the event that forwarding rules produce a tie between 6rd and native IPv6, by default, the IPv6 CE Router MUST prefer native IPv6.

4.4.2. Dual-Stack Lite (DS-Lite)

Dual-Stack Lite [RFC6333] enables both continued support for IPv4 services and incentives for the deployment of IPv6. It also decouples IPv6 deployment in the Service Provider network from the rest of the Internet, making incremental deployment easier. Dual-Stack Lite enables a broadband service provider to share IPv4 addresses among customers by combining two well-known technologies: IP in IP (IPv4-in-IPv6) and Network Address Translation (NAT). It is expected that DS-Lite traffic is forwarded over the CE Router's native IPv6 WAN interface, and not encapsulated in another tunnel.

The IPv6 CE Router SHOULD implement DS-Lite functionality. If DS-Lite is supported, it MUST be implemented according to [RFC6333]. This document takes no position on simultaneous operation of Dual-Stack Lite and native IPv4. The following CE Router requirements also apply:
WAN requirements:

DLW-1: The CE Router MUST support configuration of DS-Lite via the DS-Lite DHCPv6 option [RFC6334]. The IPv6 CE Router MAY use other mechanisms to configure DS-Lite parameters. Such mechanisms are outside the scope of this document.

DLW-2: IPv6 CE Router MUST NOT perform IPv4 Network Address Translation (NAT) on IPv4 traffic encapsulated using DS-Lite.

DLW-3: If the IPv6 CE Router is configured with an IPv4 address on its WAN interface then the IPv6 CE Router SHOULD disable the DS-Lite B4 element.

4.5. Security Considerations

It is considered a best practice to filter obviously malicious traffic (e.g., spoofed packets, "Martian" addresses, etc.). Thus, the IPv6 CE router ought to support basic stateless egress and ingress filters. The CE router is also expected to offer mechanisms to filter traffic entering the customer network; however, the method by which vendors implement configurable packet filtering is beyond the scope of this document.

Security requirements:

S-1: The IPv6 CE router SHOULD support [RFC6092]. In particular, the IPv6 CE router SHOULD support functionality sufficient for implementing the set of recommendations in [RFC6092], Section 4. This document takes no position on whether such functionality is enabled by default or mechanisms by which users would configure it.

S-2: The IPv6 CE router SHOULD support ingress filtering in accordance with BCP 38 [RFC2827]. Note that this requirement was downgraded from a MUST from RFC 6204 due to the difficulty of implementation in the CE router and the feature's redundancy with upstream router ingress filtering.

S-3: If the IPv6 CE router firewall is configured to filter incoming tunneled data, the firewall SHOULD provide the capability to filter decapsulated packets from a tunnel.

5. IANA Considerations

This document has no actions for IANA.
6. Acknowledgements

Thanks to the following people (in alphabetical order) for their guidance and feedback:


This document is based in part on CableLabs’ eRouter specification. The authors wish to acknowledge the additional contributors from the eRouter team:

Ben Bekele, Amol Bhagwat, Ralph Brown, Eduardo Cardona, Margo Dolas, Toerless Eckert, Doc Evans, Roger Fish, Michelle Kuska, Diego Mazzola, John McQueen, Harsh Parandekar, Michael Patrick, Saifur Rahman, Lakshmi Raman, Ryan Ross, Ron da Silva, Madhu Sudan, Dan Torbet, and Greg White.

7. Contributors

The following people have participated as co-authors or provided substantial contributions to this document: Ralph Droms, Kirk Erichsen, Fred Baker, Jason Weil, Lee Howard, Jean-Francois Tremblay, Yiu Lee, John Jason Brzozowski, and Heather Kirksey. Thanks to Ole Troan for editorship in the original RFC 6204 document.

8. References

8.1. Normative References

[I-D.droms-dhc-dhcpcv6-solmaxrt-update]
Droms, R., "Modification to Default Value of SOL_MAX_RT",
draft-droms-dhc-dhcpcv6-solmaxrt-update-03 (work in progress), August 2012.

[I-D.ietf-dhc-pd-exclude]
Korhonen, J., Savolainen, T., Krishnan, S., and O. Troan,


[I-D.ietf-pcp-base]


[RFC4632] Fuller, V. and T. Li, "Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan", BCP 122, RFC 4632, August 2006.


RFC 5969, August 2010.


8.2. Informative References


Appendix A. Changes from RFC 6204

1. Added IP transition technologies available in RFC form.

2. Changed requirement G-5 to augment the condition of losing IPv6 default router(s) with loss of connectivity.

3. Removed requirement WAA-7 due to not reaching consensus by various service provider standards bodies. The removal of text does not remove any critical functionality from the CE specification.

4. Changed requirement WAA-8 to qualify WAN behavior only if not configured to perform DHCPv6. This way a deployment specific profile can mandate DHCPv6 numbered WAN without conflicting with this document.

5. Changed the WPD-2 requirement from MUST be configurable to SHOULD be configurable.

6. Changed requirement WPD-4 for a default behavior without compromising any prior specification of the CE device. The change was needed by a specific layer 2 deployment which wanted to specify a MUST for DHCPv6 in their layer 2 profile and not conflict with this document.

7. Changed requirement WPD-7 to qualify text for DHCPv6. Removed W-5 and WPD-5 because the text does not have consensus from the IETF DHC Working Group for what the final solution related to the removed requirements will be.

8. Added a new WAN DHCPv6 requirement for SOL_MAX_RT of DHCPv6 so that if an service provider does not have DHCPv6 service enabled CE routers do not send too frequent DHCPv6 requests to the service provider DHCPv6 server.

9. Changed requirement L-11 from SHOULD provide DNS options in the RA to MUST provide DNS option in the RA.

10. New requirement added to the Security Considerations section due to addition of transition technology. The CE router filters decapsulated 6rd data.

11. Minor change involved changing ICMP to ICMPv6.

12. Added PCP client requirement for the WAN.
13. Added a requirement for the DHCPv6 pd-exclude option.

Authors’ Addresses

Hemant Singh
Cisco Systems, Inc.
1414 Massachusetts Ave.
Boxborough, MA 01719
USA

Phone: +1 978 936 1622
EMail: shemant@cisco.com
URI: http://www.cisco.com/

Wes Beebee
Cisco Systems, Inc.
1414 Massachusetts Ave.
Boxborough, MA 01719
USA

Phone: +1 978 936 2030
EMail: wbeebee@cisco.com
URI: http://www.cisco.com/

Chris Donley
CableLabs
858 Coal Creek Circle
Louisville, CO 80027
USA

EMail: c.donley@cablelabs.com

Barbara Stark
AT&T
725 W Peachtree St.
Atlanta, GA 30308
USA

EMail: barbara.stark@att.com
Abstract

A stateless IPv4/IPv6 translator may receive ICMPv6 packets containing non IPv4-translatable addresses as the source. These packets should be passed across the translator as ICMP packets directed to the IPv4 destination. This document presents recommendations for source address translation in ICMPv6 headers to handle such cases.
Table of Contents

1. Introduction ............................................. 3
2. Notational Conventions .................................... 3
3. Problem Statement and Considerations ....................... 3
   3.1. Considerations ...................................... 3
   3.2. Recommendations .................................... 4
4. ICMP Extension ............................................. 4
5. Stateless Address Mapping Algorithm ......................... 4
6. Security Considerations .................................... 5
7. IANA Considerations ....................................... 5
8. Acknowledgments .......................................... 5
9. References ................................................... 5
   9.1. Normative References .................................. 5
   9.2. Informative References ............................... 6
Authors’ Addresses ............................................ 6
1. Introduction

[RFC6145] section 5.2 of the "IP/ICMP Translation Algorithm" document states that "the IPv6 addresses in the ICMPv6 header may not be IPv4-translatable addresses and there will be no corresponding IPv4 addresses representing this IPv6 address. In this case, the translator can do stateful translation. A mechanism by which the translator can instead do stateless translation is left for future work." This document, Stateless Source Address Mapping for ICMPv6 Packets, provides recommendations for this case.

For the purposes of this document, the term IPv4-translatable address" is as defined in Section 2.2 of [RFC6052].

2. Notational Conventions

The key words MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL, when they appear in this document, are to be interpreted as described in [RFC2119].

3. Problem Statement and Considerations

When a stateless IPv4/IPv6 translator receives an ICMPv6 message [RFC4443] (for example "Packet Too Big") sourced from an non-IPv4-translatable IPv6 address, bound for to an IPv4-translatable IPv6 address, the translator needs to pick a source address with which to generate an ICMP message. For the reasons discussed below, this choice is problematic.

3.1. Considerations

The source address used, SHOULD NOT cause the ICMP packet to be discarded. It SHOULD NOT be drawn from [RFC1918] address space, because [RFC1918] sourced packets are likely to be subject to uRPF [RFC3704] filtering.

IPv4/IPv6 translation is intended for use in contexts where IPv4 addresses may not be readily available, so it is not considered appropriate to assign IPv4-translatable IPv6 addresses for all internal points in the IPv6 network that may originate ICMPv6 messages.

Another consideration for source selection is that it should be possible for the IPv4 recipients of the ICMP message to be able to distinguish between different IPv6 network origination of ICMPv6 messages, (for example, to support a traceroute diagnostic utility
that provides some limited network level visibility across the IPv4/IPv6 translator). This consideration implies that an IPv4/IPv6 translator needs to have a pool of IPv4 addresses for mapping the source address of ICMPv6 packets generated from different origins, or to include the IPv6 source address information for mapping the source address by others means. Currently, the TRACEROUTE and MTR [MTR] are the only consumers of translated ICMPv6 messages that care about the ICMPv6 source address.

3.2. Recommendations

The recommended approach to source selection is to use the a single (or small pool) of public IPv4 address as the source address of the translated ICMP message and leverage ICMP extension [RFC5837] to include IPv6 address as an Interface IP Address Sub-Object.

4. ICMP Extension

In the case of either a single public IPv4 address (the IPv4 interface address or loopback address of the translator) or a pool of public IPv4 addresses, the translator SHOULD implement ICMP extension defined by [RFC5837]. The ICMP message SHOULD include the Interface IP Address Sub-Object, and specify the source IPv6 addresses of the original ICMPv6. When an enhanced traceroute application is used, it can derive the real IPv6 source addresses which generated the ICMPv6 messages. Therefore, it would be able improve on visibility towards the origin rather than simply blackholing at or beyond the translator. In the future, a new ICMP extension whose presence indicates that the packet has been translated and that the source address belongs to the translator, not the originating node can also be considered.

5. Stateless Address Mapping Algorithm

If a pool of public IPv4 addresses is configured on the translator, it is RECOMMENDED to randomly select the IPv4 source address from the pool. Random selection reduces the probability that two ICMP messages elicited by the same TRACEROUTE might specify the same source address and, therefore, erroneously present the appearance of a routing loop.

[RFC5837] extensions and an enhanced traceroute application, if used, will reveal the IPv6 source addresses which generated the original ICMPv6 messages.
6. Security Considerations

This document recommends the generation of IPv4 ICMP messages from IPv6 ICMP messages. These messages would otherwise have been discarded. It is not expected that new considerations result from this change. As with a number of ICMP messages, a spoofed source address may result in replies arriving at hosts that did not expect them using the facility of the translator.

7. IANA Considerations

There is no consideration requested of IANA.

8. Acknowledgments

The authors would like to acknowledge the following contributors of this document: Kevin Yin, Chris Metz, Neeraj Gupta and Joel Jaeggli. The authors would also like to thank Ronald Bonica, Ray Hunter, George Wes, Yu Guanghui, Sowmini Varadhan, David Farmer, Fred Baker, Leo Vegoda, Joel Jaeggli, Henrik Levkowitz, Henrik Levkowitz, Randy Bush and Warren Kumari for their comments and suggestions.

9. References

9.1. Normative References


9.2. Informative References

[MTR] "http://www.bitwizard.nl/mtr/".

Authors’ Addresses

Xing Li
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing 100084
CN
Phone: +86 10-62785983
Email: xing@cernet.edu.cn

Congxiao Bao
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing 100084
CN
Phone: +86 10-62785983
Email: congxiao@cernet.edu.cn

Dan Wing
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA
Email: dwing@cisco.com
Ramji Vaithianathan
Cisco Systems, Inc.
A 5-2, BGL 12-4, SEZ Unit,
Cessna Business Park, Varthur Hobli
Sarjapur Outer Ring Road
BANGALORE   KARNATAKA 560 103
INDIA

Phone: +91 80 4426 0895
Email: rvaithia@cisco.com

Geoff Huston
APNIC

Email: gih@apnic.net
Implementation Advice for IPv6 Router Advertisement Guard (RA-Guard)
draft-ietf-v6ops-ra-guard-implementation-07

Abstract

The IPv6 Router Advertisement Guard (RA-Guard) mechanism is commonly employed to mitigate attack vectors based on forged ICMPv6 Router Advertisement messages. Many existing IPv6 deployments rely on RA-Guard as the first line of defense against the aforementioned attack vectors. However, some implementations of RA-Guard have been found to be prone to circumvention by employing IPv6 Extension Headers. This document describes the evasion techniques that affect the aforementioned implementations, and formally updates RFC 6105, such that the aforementioned RA-Guard evasion vectors are eliminated.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on May 18, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents
carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction .................................................. 3
2. Evasion techniques for some Router Advertisement Guard (RA Guard) implementations ................................. 4
   2.1. Attack Vector based on IPv6 Extension Headers ........ 4
   2.2. Attack vector based on IPv6 fragmentation .............. 4
3. RA-Guard implementation advice ................................. 8
4. Other Implications ............................................... 11
5. IANA Considerations ............................................ 12
6. Security Considerations ......................................... 13
7. Acknowledgements ................................................ 14
8. References ....................................................... 15
   8.1. Normative References ....................................... 15
   8.2. Informative References ..................................... 15
Appendix A. Assessment tools ....................................... 17
Author’s Address ................................................... 18
1. Introduction

IPv6 Router Advertisement Guard (RA-Guard) is a mitigation technique for attack vectors based on ICMPv6 Router Advertisement messages. [RFC6104] describes the problem statement of "Rogue IPv6 Router Advertisements", and [RFC6105] specifies the "IPv6 Router Advertisement Guard" functionality.

The concept behind RA-Guard is that a layer-2 device filters ICMPv6 Router Advertisement messages, according to a number of different criteria. The most basic filtering criterion is that Router Advertisement messages are discarded by the layer-2 device unless they are received on a specified port of the layer-2 device. Clearly, the effectiveness of the RA Guard mitigation relies on the ability of the layer-2 device to identify ICMPv6 Router Advertisement messages.

Some popular RA-Guard implementations have been found to be easy to circumvent by employing IPv6 extension headers [CPNI-IPv6]. This document describes such evasion techniques, and provides advice to RA-Guard implementers such that the aforementioned evasion vectors can be eliminated.

It should be noted that the aforementioned techniques could also be exploited to evade network monitoring tools such as NDPMon [NDPMon], ramond [ramond], and rafixd [rafixd], and could probably be exploited to perform stealth DHCPv6 attacks.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].
2. Evasion techniques for some Router Advertisement Guard (RA Guard) implementations

The following subsections describe two different vectors that have been found to be effective for the evasion of popular implementations of the RA-Guard protection. Section 2.1 describes an attack vector based on the use of IPv6 Extension Headers with the ICMPv6 Router Advertisement messages, which may be used to circumvent the RA-Guard protection of those implementations that fail to process an entire IPv6 header chain when trying to identify the ICMPv6 Router Advertisement messages. Section 2.2 describes an attack method based on the use of IPv6 fragmentation, possibly in conjunction with the use of IPv6 Extension Headers. This later vector has been found to be effective with all existing implementations of the RA-Guard mechanism.

2.1. Attack Vector based on IPv6 Extension Headers

While there is currently no legitimate use for IPv6 Extension Headers in ICMPv6 Router Advertisement messages, Neighbor Discovery implementations allow the use of Extension Headers with these messages, by simply ignoring the received options. Some RA-Guard implementations try to identify ICMPv6 Router Advertisement messages by simply looking at the "Next Header" field of the fixed IPv6 header, rather than following the entire header chain. As a result, such implementations fail to identify any ICMPv6 Router Advertisement messages that include any Extension Headers (for example, a Hop by Hop Options header, a Destination Options Header, etc.), and can be easily circumvented.

The following figure illustrates the structure of ICMPv6 Router Advertisement messages that implement this evasion technique:

```
+-----------------+-----------------+-----------------+-----------------+-----------------+
| NH=60 | NH=58 | | | |
|++++++ | +++-- ++ | | |
| | IPv6 header | Dst Opt Hdr | ICMPv6 Router Advertisement |
| | | + + + | + |
`-----------------+-----------------+-----------------+-----------------+-----------------+
```

2.2. Attack vector based on IPv6 fragmentation

This section presents a different attack vector, which has been found to be effective against all implementations of RA-Guard. The basic idea behind this attack vector is that if the forged ICMPv6 Router Advertisement is fragmented into at least two fragments, the layer-2
A device implementing "RA-Guard" would be unable to identify the attack packet, and would thus fail to block it.

A first variant of this attack vector would be an original ICMPv6 Router Advertisement message preceded with a Destination Options Header, that results in two fragments. The following figure illustrates the "original" attack packet, prior to fragmentation, and the two resulting fragments which are actually sent as part of the attack.

Original packet:

```
+------------------------------------------+
| NH=60 | NH=58 |                             |
|       +       +                  |
| IPv6 header | Dst Opt Hdr | ICMPv6 RA |
| +     +     +     +             |
+------------------------------------------+
```

First fragment:

```
+------------------------------------------+
| NH=44 | NH=60 | NH=58 |
|       +       +       +               |
| IPv6 Header | Frag Hdr | Dst Opt Hdr |
| +     +     +     +               |
+------------------------------------------+
```

Second fragment:

```
+------------------------------------------+
| NH=44 | NH=60 |                         |
|       +       +               |
| IPv6 header | Frag Hdr | Dst Opt Hdr | ICMPv6 RA |
| +     +     +     +               |
+------------------------------------------+
```

It should be noted that the "Hdr Ext Len" field of the Destination Options Header is present in the first fragment (rather than the second). Therefore, it is impossible for a device processing only the second fragment to locate the ICMPv6 header contained in that fragment, since it is unknown how many bytes should be "skipped" to get to the next header following the Destination Options Header.
Thus, by leveraging the use of the Fragment Header together with the use of the Destination Options header, the attacker is able to conceal the type and contents of the ICMPv6 message he is sending (an ICMPv6 Router Advertisement in this example). Unless the layer-2 device were to implement IPv6 fragment reassembly, it would be impossible for the device to identify the ICMPv6 type of the message.

A layer-2 device could, however, at least detect that an ICMPv6 message (or some type) is being sent, since the "Next Header" field of the Destination Options header contained in the first fragment is set to "58" (ICMPv6).

This idea can be taken further, such that it is also impossible for the layer-2 device to detect that the attacker is sending an ICMPv6 message in the first place. This can be achieved with an original ICMPv6 Router Advertisement message preceded with two Destination Options Headers, that results in two fragments. The following figure illustrates the "original" attack packet, prior to fragmentation, and the two resulting packets which are actually sent as part of the attack.
Original packet:

```
++++++-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|NH=60|       |NH=60|       |NH=58|       |
++++++        +++-+        +++-+        +         +
| IPv6 header | Dst Opt Hdr | Dst Opt Hdr | ICMPv6 RA |
|             |             |             |           |
++++++-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

First fragment:

```
++++++-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|NH=44|       |NH=60|       |NH=60|       |
++++++        +++-+        +++-+        +         +
| IPv6 header | Frag Hdr  |       Dst Opt Hdr       |
|             |             |             |           |
++++++-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

Second fragment:

```
++++++-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
|NH=44|       |NH=60|       |       |NH=58|       |
++++++        +++-+        +        +++-+        +
| IPv6 header | Frag Hdr  | Dst O Hdr | Dst Opt Hdr | ICMPv6 RA |
|             |             |           |             |           |
++++++-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-
```

In this variant, the "Next Header" field of the Destination Options header contained in the first fragment is set "60" (Destination Options header), and thus it is impossible for a device processing only the first fragment to detect that an ICMPv6 message is being sent in the first place.

The second fragment presents the same challenges as the second fragment of the previous variant. That is, it would be impossible for a device processing only the second fragment to locate the second Destination Options header (and hence the ICMPv6 header), since the "Hdr Ext Len" field of the first Destination Options header is present in the first fragment (rather than the second).
3. RA-Guard implementation advice

The following filtering rules must be implemented as part of an "RA-Guard" implementation on ports that face interfaces that are not allowed to send ICMPv6 Router Advertisement messages, such that the vulnerabilities discussed in this document are eliminated:

1. If the IPv6 Source Address of the packet is not a link-local address (fe80::/10), RA-Guard must pass the packet.

   RATIONALE: This prevents "RA-Guard" from dedicating compute cycles to filtering packets that originate off-net and, if they are RA’s, would not be accepted by the host. Section 6.1.2 of [RFC4861] requires nodes to discard Router Advertisement messages if their IPv6 Source Address is not a link-local address.

2. If the Hop Limit is not 255, RA-Guard must pass the packet.

   RATIONALE: This prevents "RA-Guard" from dedicating compute cycles to filtering packets that originate off-net and, if they are RA’s, would not be accepted by the host. Section 6.1.2 of [RFC4861] requires nodes to discard Router Advertisement messages if their Hop Limit is not 255.

3. RA-Guard must parse the IPv6 entire header chain present in the packet, to identify whether the packet is a Router Advertisement message.

   RATIONALE: [RFC6564] specifies a uniform format for IPv6 Extension Header, thus meaning that an IPv6 node can parse an IPv6 header chain even if it contains Extension Headers that are not currently supported by that node. Additionally, [I-D.ietf-6man-oversized-header-chain] requires that if a packet is fragmented, the first fragment contains the entire IPv6 header chain.

   RA-Guard implementations must not enforce a limit on the number of bytes they can inspect (starting from the beginning of the IPv6 packet), since this could introduce false-positives: legitimate packets could be dropped simply because the RA-Guard device does not parse the entire IPv6 header chain present in the packet. An implementation that has such an implementation-specific limit must not claim compliance with this specification, and must pass the packet when such implementation-specific limit is reached.
4. When parsing the IPv6 header chain, if the packet is a first-fragment (i.e., a packet containing a Fragment Header with the Fragment Offset set to 0) and it fails to contain the entire IPv6 header chain (i.e., all the headers starting from the IPv6 header up to, and including, the upper-layer header), RA-Guard must drop the packet, and should log the packet drop event in an implementation-specific manner as a security fault.

RATIONALE: [I-D.ietf-6man-oversized-header-chain] specifies that the first-fragment (i.e., the fragment with the Fragment Offset set to 0) MUST contain the entire IPv6 header chain, and allows intermediate systems such as routers to drop those packets that fail to comply with this requirement.

NOTE: This rule should only be applied to IPv6 fragments with a Fragment Offset of 0 (non-first fragments can be safely passed, since they will never reassemble into a complete datagram if they are part of a Router Advertisement received on a port where such packets are not allowed).

5. When parsing the IPv6 header chain, if the packet is identified to be an ICMPv6 Router Advertisement message, RA-Guard must drop the packet, and should log the packet drop event in an implementation-specific manner as a security fault.

RATIONALE: By definition, Router Advertisement messages MUST originate on-link, MUST have a link-local IPv6 Source Address, and MUST have a Hop Limit value of 255. [RFC4861].

6. In all other cases, RA-Guard must pass the packet as usual.

NOTE: For the purpose of enforcing the RA-Guard filtering policy, an ESP header [RFC4303] should be considered to be an "upper-layer protocol" (that is, it should be considered the last header in the IPv6 header chain). This means that packets employing ESP would be passed by the RA-Guard device to the intended destination. If the destination host does not have a security association with the sender of the aforementioned IPv6 packet, the packet would be dropped. Otherwise, if the packet is considered valid by the IPsec implementation at the receiving host and encapsulates a Router Advertisement message, it is up to the receiving host what to do with such packet.

If a packet is dropped due to this filtering policy, then the packet drop event should be logged in an implementation-specific manner as a security fault. The logging mechanism should include a drop counter dedicated to RA-Guard packet drops.
In order to protect current end-node IPv6 implementations, Rule #4 has been defined as a default rule to drop packets that cannot be positively identified as not being Router Advertisement (RA) messages (because the packet is a fragment that fails to include the entire IPv6 header chain). This means that, at least in theory, RA-Guard could result in false-positive blocking of some legitimate non-RA packets that could not be positively identified as being non-RA. In order to reduce the likelihood of false positives, Rule #1 and Rule #2 require that packets that would not pass the required validation checks for RA messages (Section 6.1.2 of [RFC4861]) be passed without further inspection. In any case, as noted in [I-D.ietf-6man-oversized-header-chain], IPv6 packets that fail to include the entire IPv6 header chain are virtually impossible to police with state-less filters and firewalls, and hence are unlikely to survive in real networks. [I-D.ietf-6man-oversized-header-chain] requires that hosts employing fragmentation include the entire IPv6 header chain in the first fragment (the fragment with the Fragment Offset set to 0), thus eliminating the aforementioned false positives.

This filtering policy assumes that host implementations require that the IPv6 Source Address of ICMPv6 Router Advertisement messages be a link-local address, and that they discard the packet if this check fails, as required by the current IETF specifications [RFC4861]. Additionally, it assumes that hosts require the Hop Limit of Neighbor Discovery messages to be 255, and discard those packets otherwise.

The aforementioned filtering rules implicitly handle the case of fragmented packets: if the RA-Guard device fails to identify the upper-layer protocol as a result of the use of fragmentation, the corresponding packets would be dropped.

Finally, we note that IPv6 implementations that allow overlapping fragments (i.e. that do not comply with [RFC5722]) might still be subject of RA-based attacks. However, a recent assessment of IPv6 implementations [SI6-FRAG] with respect to their fragment reassembly policy seems to indicate that most current implementations comply with [RFC5722].
4. Other Implications

A similar concept to that of "RA-Guard" has been implemented for protecting against forged DHCPv6 messages. Such protection can be circumvented with the same techniques discussed in this document, and the counter-measures for such evasion attack are analogous to those described in Section 3 of this document.

[DHCPv6-Shield] specifies a mechanism to protect against rogue DHCPv6 servers, while taking into consideration the evasion techniques discussed in this document.
5. IANA Considerations

This document has no actions for IANA.
6. Security Considerations

This document describes a number of techniques that have been found to be effective to circumvent popular RA-Guard implementations, and provides advice to RA-Guard implementations such that those evasion vulnerabilities are eliminated.

As noted in Section 3, IPv6 implementations that allow overlapping fragments (i.e. that do not comply with [RFC5722]) might still be subject of RA-based attacks. However, most current implementations seem to comply with [RFC5722].

We note that if an attacker sends a fragmented Router Advertisement message on a port not allowed to send such packets, the first-fragment would be dropped, and the rest of the fragments would be passed. This means that the victim node would tie memory buffers for the aforementioned fragments, which would never reassemble into a complete datagram. If a large number of such packets were sent by an attacker, and the victim node failed to implement proper resource management for the fragment reassembly buffer, this could lead to a Denial of Service (DoS). However, this does not really introduce a new attack vector, since an attacker could always perform the same attack by sending forged fragmented datagram in which at least one of the fragments is missing. [CPNI-IPv6] discusses some resource management strategies that could be implemented for the fragment reassembly buffer.

We note that most effective and efficient mitigation for these attacks would be to prohibit the use of IPv6 fragmentation with Router Advertisement messages (as proposed by [I-D.ietf-6man-nd-extension-headers]), such that the RA-Guard functionality is easier to implement. However, since such mitigation would require an update to existing implementations, it cannot be relied upon in the short or near term.

Finally, we note that RA-Guard only mitigates attack vectors based on ICMPv6 Router advertisement messages. Protection against similar attacks based on other messages (such as DHCPv6) is considered out of the scope of this document, and left for other documents(e.g. [DHCPv6-Shield]).
7. Acknowledgements

The author would like to thank Ran Atkinson, who provided very detailed comments and suggested text that was incorporated into this document.

The author would like to thank Ran Atkinson, Karl Auer, Robert Downie, Washam Fan, David Farmer, Marc Heuse, Nick Hilliard, Ray Hunter, Joel Jaeggli, Simon Perreault, Arturo Servin, Gunter van de Velde, James Woodyatt, and Bjoern A. Zeeb, for providing valuable comments on earlier versions of this document.

The author would like to thank Arturo Servin, who presented this document at IETF 81.

This document resulted from the project "Security Assessment of the Internet Protocol version 6 (IPv6)" [CPNI-IPv6], carried out by Fernando Gont on behalf of the UK Centre for the Protection of National Infrastructure (CPNI). The author would like to thank the UK CPNI, for their continued support.
8. References

8.1. Normative References


8.2. Informative References


[CPNI-IPv6]


[S16-FRAG]


Appendix A. Assessment tools

[SI6-IPv6] is a publicly-available set of tools (for Linux, *BSD, and Mac OS) that implements the techniques described in this document.

[THC-IPV6] is a publicly-available set of tools (for Linux) that implements some of the techniques described in this document.
Author’s Address

Fernando Gont
Centre for the Protection of National Infrastructure

Email: fgont@si6networks.com
URI: http://www.cpni.gov.uk
Wireline Incremental IPv6
draft-ietf-v6ops-wireline-incremental-ipv6-06

Abstract

Operators worldwide are in various stages of preparing for, or deploying IPv6 into their networks. The operators often face difficult challenges related to both IPv6 introduction along with those related to IPv4 run out. Operators will need to meet the simultaneous needs of IPv6 connectivity and continue support for IPv4 connectivity for legacy devices with a stagnant supply of IPv4 addresses. The IPv6 transition will take most networks from an IPv4-only environment to an IPv6 dominant environment with long transition period varying by operator. This document helps provide a framework for wireline providers who are faced with the challenges of introducing IPv6 along with meeting the legacy needs of IPv4 connectivity utilizing well defined and commercially available IPv6 transition technologies.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on March 14, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.
Table of Contents

1.  Introduction ...........................................  4
2.  Operator Assumptions ...................................  4
3.  Reasons and Considerations for a Phased Approach ............  5
   3.1.  Relevance of IPv6 and IPv4 ............................  6
   3.2.  IPv4 Resource Challenges ...............................  6
   3.3.  IPv6 Introduction and Operational Maturity ...............  7
   3.4.  Service Management ....................................  8
   3.5.  Sub-Optimal Operation of Transition Technologies ..........  8
   3.6.  Future IPv6 Network ...................................  9
4.  IPv6 Transition Technology Analysis ............................  9
   4.1.  Automatic Tunneling using 6to4 and Teredo ................  9
   4.2.  Carrier Grade NAT (NAT444) ............................ 10
   4.3.  6RD .................................................. 10
   4.4.  Native Dual Stack ...................................... 11
   4.5.  DS-Lite .............................................. 11
   4.6.  NAT64 ................................................ 12
5.  IPv6 Transition Phases ....................................... 12
   5.1.  Phase 0 - Foundation ..................................... 13
      5.1.1.  Phase 0 - Foundation: Training ....................... 13
      5.1.2.  Phase 0 - Foundation: System Capabilities .......... 14
      5.1.3.  Phase 0 - Foundation: Routing ....................... 14
      5.1.4.  Phase 0 - Foundation: Network Policy and Security .... 14
      5.1.5.  Phase 0 - Foundation: Transition Architecture ....... 14
      5.1.6.  Phase 0- Foundation: Tools and Management .......... 15
   5.2.  Phase 1 - Tunneled IPv6 ............................ 15
      5.2.1.  6RD Deployment Considerations ....................... 17
   5.3.  Phase 2: Native Dual Stack ............................ 19
      5.3.1.  Native Dual Stack Deployment Considerations .......... 19
   5.4.  Intermediate Phase for CGN ............................ 20
      5.4.1.  CGN Deployment Considerations ....................... 21
   5.5.  Phase 3 - IPv6-Only .................................... 22
      5.5.1.  DS-Lite ............................................ 23
      5.5.2.  DS-Lite Deployment Considerations ................... 23
      5.5.3.  NAT64 Deployment Considerations ..................... 24
6.  IANA Considerations .......................................... 25
7.  Security Considerations ..................................... 25
8.  Acknowledgements ........................................... 26
9.  References .................................................. 26
   9.1.  Normative References .................................... 26
   9.2.  Informative References .................................. 26
Authors’ Addresses ............................................. 29
1. Introduction

This draft sets out to help wireline operators in planning their IPv6 deployments while ensuring continued support for IPv6-incapable consumer devices and applications. This document identifies which technologies can be used incrementally to transition from IPv4-only to an IPv6 dominant environment with support for dual stack operation. The end state goal for most operators will be IPv6-only, but the path to this final state will heavily depend on the amount of legacy equipment resident in end networks and management of long tail IPv4-only content. Although no single plan will work for all operators, options listed herein provide a baseline which can be included in many plans.

This draft is intended for wireline environments which include Cable, DSL and/or fiber as the access method to the end consumer. This document attempts to follow the principles laid out in [RFC6180] which provides guidance on using IPv6 transition mechanisms. This document will focus on technologies which enable and mature IPv6 within the operator’s network, but will also include a cursory view of IPv4 connectivity continuance. The focal transition technologies include 6RD [RFC5969], DS-Lite [RFC6333], NAT64 [RFC6146] and Dual Stack operation which may also include a CGN/NAT444 deployment. Focus on these technologies is based on their inclusion in many off-the-shelf CPEs and availability in commercially available equipment.

2. Operator Assumptions

For the purposes of this document, the authors assume:

- The operator is considering deploying IPv6 or is in the progress in deploying IPv6
- The operator has a legacy IPv4 subscriber base that will continue to exist for a period of time
- The operator will want to minimize the level of disruption to the existing and new subscribers
- The operator may also want to minimize the number of technologies and functions that are needed to mediate any given set of subscribers flows (overall preference for Native IP flows)
- The operator is able to run Dual Stack on their own core network and is able to transition their own services to support IPv6

Based on these assumptions, an operator will want to utilize
technologies that minimize the need to tunnel, translate or mediate flows to help optimize traffic flow and lower the cost impacts of transition technologies. Transition technology selections should be made to mediate the non-dominant IP family flows and allow native routing (IPv4 and/or IPv6) to forward the dominant traffic whenever possible. This allows the operator to minimize the cost of IPv6 transition technologies by minimizing the transition technology deployment size.

An operator may also choose to prefer more IPv6 focused models where the use of transition technologies are based on an effort to enable IPv6 at the base layer as soon as possible. Some operators may want to promote IPv6 early on in the deployment and have IPv6 traffic perform optimally from the outset. This desire would need to be weighed against the cost and support impacts of such a choice and the quality of experience offered to subscribers.

3. Reasons and Considerations for a Phased Approach

When faced with the challenges described in the introduction, operators may want to consider a phased approach when adding IPv6 to an existing subscriber base. A phased approach allows the operator to add in IPv6 while not ignoring legacy IPv4 connection requirements. Some of the main challenges the operator will face include:

- IPv4 exhaustion may occur long before all traffic is able to be delivered over IPv6, necessitating IPv4 address sharing

- IPv6 will pose operational challenges since some of the software is quite new and has had short run time in large production environments and organizations are also not acclimatized to supporting IPv6 as a service

- Connectivity modes will move from IPv4-only to Dual Stack in the home, changing functional behaviors in the consumer network and increasing support requirements for the operator

- Although IPv6 support on CPEs is a newer phenomenon, there is a strong push by operators and the industry as a whole to enable IPv6 on devices. As demand grows, IPv6 enablement will no longer be optional, but necessary on CPEs. Documents like [RFC6540] provide useful tools in the short term to help vendors and implementors understand what "IPv6 support" means.

These challenges will occur over a period of time, which means that the operator’s plans need to address the ever changing requirements
of the network and subscriber demand. Although phases will be presented in this document, not all operators may need to enable each discrete phase. It is possible that characteristics in individual networks may allow certain operators to skip or jump to various phases.

3.1. Relevance of IPv6 and IPv4

The delivery of high-quality unencumbered Internet service should be the primary goal for operators. With the imminent exhaustion of IPv4, IPv6 will offer the highest quality of experience in the long term. Even though the operator may be focused on IPv6 delivery, they should be aware that both IPv4 and IPv6 will play a role in the Internet experience during transition. The Internet is made of many interconnecting systems, networks, hardware, software and content sources - all of which will move to IPv6 at different rates.

Many subscribers use older operating systems and hardware which support IPv4-only operation. Internet subscribers don’t buy IPv4 or IPv6 connections; they buy Internet connections, which demands the need to support both IPv4 and IPv6 for as long as the subscriber’s home network demands such support. The operator may be able to leverage one or the other protocol to help bridge connectivity on the operator’s network, but the home network will likely demand both IPv4 and IPv6 for some time.

3.2. IPv4 Resource Challenges

Since connectivity to IPv4-only endpoints and/or content will remain common, IPv4 resource challenges are of key concern to operators. The lack of new IPv4 addresses for additional devices means that meeting the growth in demand of IPv4 connections in some networks will require address sharing.

Networks are growing at different rates including those in emerging markets and established networks based on the proliferation of Internet based services and devices. IPv4 address constraints will likely affect many if not most operators at some point, increasing the benefits of IPv6. IPv4 address exhaustion is a consideration when selecting technologies which rely on IPv4 to supply IPv6 services, such as 6RD. Additionally, if native Dual Stack is considered by the operator, challenges related to IPv4 address exhaustion remain a concern.

Some operators may be able to reclaim small amounts IPv4 addresses through addressing efficiencies in the network, although this will have little lasting benefits to the network and not meet longer term connectivity needs. Secondary markets for IPv4 addresses have also
begun to arise, but it’s not well understood how this will complement overall demand for Internet growth. Address transfers will also be subject to market prices and transfer rules governed by the Regional Registries.

The lack of new global IPv4 address allocations will therefore force operators to support some form of IPv4 address sharing and may impact technological options for transition once the operator runs out of new IPv4 addresses for assignment.

3.3. IPv6 Introduction and Operational Maturity

The introduction of IPv6 will require new operational practices. The IPv4 environment we have today was built over many years and matured by experience. Although many of these experiences are transferable from IPv4 to IPv6, new experience and practices specific to IPv6 will be needed.

Engineering and Operational staff will need to develop experience with IPv6. Inexperience may lead to early IPv6 deployment instability, and operators should consider this when selecting technologies for initial transition. Operators may not want to subject their mature IPv4 service to a "new IPv6" path initially while it may be going through growing pains. DS-Lite [RFC6333] and NAT64 [RFC6146] are both technologies which requires IPv6 to support connectivity to IPv4 endpoints or content over an IPv6-only access network.

Further, some of these transition technologies are new and require refinement within running code. Deployment experience is required to expose bugs and stabilize software in production environments. Many supporting systems are also under development and have newly developed IPv6 functionality including vendor implementations of DHCPv6, management tools, monitoring systems, diagnostic systems, logging, along with other elements.

Although the base technological capabilities exist to enable and run IPv6 in most environments, organizational experience is low. Until such time as each key technical member of an operator’s organization can identify IPv6, understand its relevance to the IP service offering, how it operates and how to troubleshoot it, the deployment needs to mature, and may be subject to subscriber-impacting events. This fact should not incline operators to delay their IPv6 deployment, but should drive them to deploy IPv6 sooner to gain the much needed experience before IPv6 is the only viable way to connect new hosts to the network.

It should also be noted that although many transition technologies
may be new, and some code related to access environments may be new, there is a large segment of the networking fabric which has had IPv6 available for a long period of time and has had extended exposure in production. Operators may use this to their advantage by first enabling IPv6 in the core of their network then work outward towards the subscriber edge.

3.4. Service Management

Services are managed within most networks and are often based on the gleaning and monitoring of IPv4 addresses assigned to endpoints. Operators will need to address such management tools, troubleshooting methods and storage facilities (such as databases) to deal with not just a new address type containing a 128-bit IPv6 address [RFC2460], but often both IPv4 and IPv6 at the same time. Examination of address type, and recording delegated prefixes along with single address assignments, will likely require additional development.

With any Dual Stack service – whether Native, 6RD-based, DS-Lite, NAT64 or otherwise – two address families may need to be managed simultaneously to help provide for the full Internet experience. This would indicate that IPv6 management is not just a simple add in, but needs to be well integrated into the service management infrastructure. In the early transition phases, it’s quite likely that many systems will be missed and that IPv6 services will go un-monitored and impairments undetected.

These issues may be of consideration when selecting technologies that require IPv6 as the base protocol to deliver IPv4 connectivity. Instability on the IPv6 service in such a case would impact IPv4 services.

3.5. Sub-Optimal Operation of Transition Technologies

Native delivery of IPv4 and IPv6 provides a solid foundation for delivery of Internet services to subscribers since native IP paths are well understood and networks are often optimized to send such traffic efficiently. Transition technologies however, may alter the normal path of traffic or reduce the path MTU, removing many network efficiencies built for native IP flows. Tunneling and translation devices may not be located on the most optimal path in line with the natural traffic flow (based on route computation) and therefore may increase latency. These paths may also add additional points of failure.

Generally, the operator will want to deliver native IPv6 as soon as possible and utilize transition technologies only when required. Transition technologies may be used to provide continued access to
IPv6 via tunneling and/or translation or may be used to deliver IPv6 connectivity. The delivery of Internet or internal services should be considered by the operator, since supplying connections using a transition technology will reduce the overall performance for the subscriber.

When choosing between various transition technologies, operators should consider the benefits and drawbacks of each option. Some technologies like CGN/NAT444 utilize many existing addressing and management practices. Other options such as DS-Lite and NAT64 remove the IPv4 addressing requirement to the subscriber premise device but require IPv6 to be operational and well supported.

3.6. Future IPv6 Network

An operator should also be aware that longer-term plans may include IPv6-only operation in all or much of the network. The IPv6-only operation may be complemented by technologies such as NAT64 for long-tail IPv4 content reach. This longer term view may be distant to some, but should be considered when planning out networks, addressing and services. The needs and costs of maintaining two IP stacks will eventually become burdensome and simplification will be desirable. The operators should plan for this state and not make IPv6 inherently dependent on IPv4 as this would unnecessarily constrain the network.

Other design considerations and guidelines for running an IPv6 network should also be considered by the operator. Guidance on designing an IPv6 network can be found in [draft-matthews-v6ops-design-guidelines] and [draft-ietf-v6ops-icp-guidance].

4. IPv6 Transition Technology Analysis

Operators should understand the main transition technologies for IPv6 deployment and IPv4 run out. This draft provides a brief description of some of the mainstream and commercially available options. This analysis is focused on the applicability of technologies to deliver residential services and less focused on commercial access, wireless, or infrastructure support.

The technologies in focus for this document are targeted on those commercially available and in deployment.

4.1. Automatic Tunneling using 6to4 and Teredo

Even when operators may not be actively deploying IPv6, automatic mechanisms exist on subscriber operating systems and CPE hardware.
Such technologies include 6to4 [RFC3056], which is most commonly used with anycast relays [RFC3068]. Teredo [RFC4380] is also used widely by many Internet hosts.

Documents such as [RFC6343] have been written to help operators understand observed problems with 6to4 deployments and provides guidelines on how to improve its performance. An operator may want to provide local relays for 6to4 and/or Teredo to help improve the protocol’s performance for ambient traffic utilizing these IPv6 connectivity methods. Experiences such as those described in [I-D.jjmb-v6ops-comcast-ipv6-experiences] show that local relays have proved beneficial to 6to4 protocol performance.

Operators should also be aware of breakage cases for 6to4 if non-RFC1918 addresses are used within CGN/NAT444 zones. Many off-the-shelf CPEs and operating systems may turn on 6to4 without a valid return path to the originating (local) host. This particular use case can occur if any space other than [RFC1918] is used, including Shared Address Space [RFC6598] or space registered to another organization (squat space). The operator can use 6to4-PMT [I-D.kuarsingh-v6ops-6to4-provider-managed-tunnel] or attempt to block 6to4 operation entirely by blocking the anycast ranges associated with [RFC3068].

4.2. Carrier Grade NAT (NAT444)

Carrier Grade NAT (CGN), specifically as deployed in a NAT444 scenario [I-D.ietf-behave-lsn-requirements], may prove beneficial for those operators who offer Dual Stack services to subscriber endpoints once they exhaust their pools of IPv4 addresses. CGNs, and address sharing overall, are known to cause certain challenges for the IPv4 service [RFC6269][I-D.donley-nat444-impacts], but may be necessary depending on how an operator has chosen to deal with IPv6 transition and legacy IPv4 connectivity requirements.

In a network where IPv4 address availability is low, CGN/NAT444, may provide continued access to IPv4 endpoints. Some of the advantages of using CGN/NAT444 include the similarities in provisioning and activation models. IPv4 hosts in a CGN/NAT444 deployment will likely inherit the same addressing and management procedures as legacy IPv4, globally addressed hosts (i.e. DHCPv4, DNSv4, TFTP, TR-069 etc).

4.3. 6RD

6RD [RFC5969] provides a way of offering IPv6 connectivity to subscriber endpoints when native IPv6 addressing on the access network is not yet possible. 6RD provides tunneled connectivity for IPv6 over the existing IPv4 path. As the access edge is upgraded and
subscriber premise equipment is replaced, 6RD can be replaced by native IPv6 connectivity. 6RD can be delivered over top a CGN/NAT444 deployment, but this would cause all traffic to be subject to some type of transition technology.

6RD may also be advantageous during the early transition while IPv6 traffic volumes are low. During this period, the operator can gain experience with IPv6 on the core and improve their peering framework to match those of the IPv4 service. 6RD scales by adding relays to the operator’s network. Another advantage for 6RD is that the operator does not need a DHCPv6 address assignment infrastructure and does not need to support IPv6 routing to the CPE to support a delegated prefix (as it’s derived from the IPv4 address and other configuration parameters).

Client support is required for 6RD operation and may not be available on deployed hardware. 6RD deployments may require the subscriber or operator to replace the CPE. 6RD will also require parameter configuration which can be powered by the operator through DHCPv4, manually provisioned on the CPE or automatically through some other means. Manual provisioning would likely limit deployment scale.

4.4. Native Dual Stack

Native Dual Stack is often referred to as the "gold standard" of IPv6 and IPv4 delivery. It is a method of service delivery that is already used in many existing IPv6 deployments. Native Dual Stack does, however, require that Native IPv6 be delivered through the access network to the subscriber premise. This technology option is desirable in many cases and can be used immediately if the access network and subscriber premise equipment supports native IPv6.

An operator who runs out of IPv4 addresses to assign to subscribers will not be able to provide traditional native Dual Stack connectivity for new subscribers. In Dual Stack deployments where sufficient IPv4 addresses are not available, CGN/NAT444 can be used on the IPv4 path.

Delivering native Dual Stack would require the operator’s core and access network to support IPv6. Other systems like DHCP, DNS, and diagnostic/management facilities need to be upgraded to support IPv6 as well. The upgrade of such systems may often be non-trivial and costly.

4.5. DS-Lite

Dual-Stack Lite (DS-Lite, [RFC6333]) is based on a native IPv6 connection model where IPv4 services are supported. DS-Lite provides
tunneled connectivity for IPv4 over the IPv6 path between the subscriber’s network device and a provider managed gateway (AFTR).

DS-Lite can only be used where there is a native IPv6 connection between the AFTR and the CPE. This may mean that the technology’s use may not be viable during early transition if the core or access network lacks IPv6 support. During the early transition period, a significant amount of content and services may be by IPv4-only. Operators may be sensitive to this and may not want the newer IPv6 path to be the only bridge to IPv4 at that time given the potential impact. The operator may also want to make sure that most of its internal services and a significant amount of external content is available over IPv6 before deploying DS-Lite. The availability of services on IPv6 would help lower the demand on the AFTRs.

By sharing IPv4 addresses among multiple endpoints, like CGN/NAT444, DS-Lite can facilitate continued support of legacy IPv4 services even after IPv4 address run out. There are some functional considerations to take into account with DS-Lite, such as those described in [I-D.donley-nat444-impacts] and in [I-D.ietf-softwire-dslite-deployment].

DS-Lite requires client support on the CPE to function. The ability to utilize DS-Lite will be dependent on the operator providing DS-Lite capable CPEs or retail availability of the supported client for subscriber-acquired endpoints.

4.6. NAT64

NAT64 [RFC6146] provides the ability to connect IPv6-only connected clients and hosts to IPv4 servers without any tunneling. NAT64 requires that the host and home network support IPv6-only modes of operation. Home networks do not commonly contain equipment that is 100% IPv6-capable. It is also not anticipated that common home networks will be ready for IPv6-only operation for a number of years. However, IPv6-only networking can be deployed by early adopters or highly controlled networks [RFC6586].

Viability of NAT64 will increase in wireline networks as consumer equipment is replaced by IPv6 capable versions. There are incentives for operators to move to IPv6-only operation, when feasible, which includes the simplicity of a single stack access network.

5. IPv6 Transition Phases

The Phases described in this document are not provided as a rigid set of steps, but are considered a guideline which should be analyzed by
operators planning their IPv6 transition. Operators may choose to skip steps based on technological capabilities within their specific networks, (residential/corporate, fixed/mobile), their business development perspectives (which may affect the pace of migration towards full IPv6), or a combination thereof.

The phases are based on the expectation that IPv6 traffic volume may initially be low, and operator staff will gain experience with IPv6 over time. As traffic volumes of IPv6 increase, IPv4 traffic volumes will decline (in percentage relative to IPv6), until IPv6 is the dominant address family used. Operators may want to keep the traffic flow for the dominant traffic class (IPv4 vs. IPv6) native to help manage costs related to transition technologies. The cost of using multiple technologies in succession to optimize each stage of the transition should also be compared against the cost of changing and upgrading subscriber connections.

Additional guidance and information on utilizing IPv6 transition mechanisms can be found in [RFC6180]. Also, guidance on incremental CGN for IPv6 transition can also be found in [RFC6264].

5.1. Phase 0 - Foundation

5.1.1. Phase 0 - Foundation: Training

Training is one of the most important steps in preparing an organization to support IPv6. Most people have little experience with IPv6, and many do not even have a solid grounding in IPv4. The implementation of IPv6 will likely produce many challenges due to immature code on hardware, and the evolution of many applications and systems to support IPv6. To properly deal with these impending or current challenges, organizations must train their staff on IPv6.

Training should also be provided within reasonable timelines from the actual IPv6 deployment. This means the operator needs to plan in advance as it trains the various parts of its organization. New Technology and Engineering staff often receive little training because of their depth of knowledge, but must at least be provided opportunities to read documentation, architectural white papers, and RFCs. Operations personnel who support the network and other systems need to be trained closer to the deployment timeframes, so they immediately use their new-found knowledge before forgetting.

Subscriber support staff would require much more basic but large scale training since many organizations have massive call centers to support the subscriber base. Tailored training will also be required for marketing and sales staff to help them understand IPv6 and build it into the product development and sales process.
5.1.2. Phase 0 - Foundation: System Capabilities

An important component with any IPv6 network architecture and implementation is the assessment of the hardware and operating capabilities of the deployed equipment (and software). The assessment needs to be conducted irrespective of how the operator plans to transition their network. The capabilities of the install base will however impact what technologies and modes of operation may be supported and therefore what technologies can be considered for the transition. If some systems do not meet the needs of the operator’s IPv6 deployment and/or transition plan, the operator may need to plan for replacement and/or upgrade.

5.1.3. Phase 0 - Foundation: Routing

The network infrastructure will need to be in place to support IPv6. This includes the routed infrastructure along with addressing principles, routing principles, peering policy and related network functions. Since IPv6 is quite different from IPv4 in several ways including the number of addresses which are made available, careful attention to a scalable and manageable architecture needs to be made. One such change is the notion of a delegated prefix, which deviates from the common single address phenomenon in IPv4-only deployments. Deploying prefixes per CPE can load the routing tables and require a routing protocol or route gleaning to manage connectivity to the subscriber’s network. Delegating prefixes can be of specific importance in access network environments where downstream subscribers often move between access nodes, raising the concern of frequent renumbering and/or managing movement of routed prefixes within the network (common in cable based networks).

5.1.4. Phase 0 - Foundation: Network Policy and Security

Many, but not all, security policies will map easily from IPv4 to IPv6. Some new policies may be required for issues specific to IPv6 operation. This document does not highlight these specific issues, but raises the awareness they are of consideration and should be addressed when delivering IPv6 services. Other IETF documents such as [RFC4942], [RFC6092], and [RFC6169] are excellent resources.

5.1.5. Phase 0 - Foundation: Transition Architecture

The operators should plan out their transition architecture in advance (with room for flexibility) to help optimize how they will build out and scale their networks. Should the operator consider multiple technologies like CGN/NAT444, DS-Lite, NAT64 and 6RD, they may want to plan out where network resident equipment may be located and potentially choose locations which can be used for all functional
roles (i.e. Placement of NAT44 translator, AFTR, NAT64 gateway and 6RD relays). Although these functions are not inherently connected, additional management, diagnostic and monitoring functions can be deployed along side the transition hardware without the need to distribute these to an excessive or divergent number of locations.

This approach may also prove beneficial if traffic patterns change rapidly in the future as the operators may need to evolve their transition infrastructure faster than originally anticipated. One such example may be the movement from a CGN/NAT44 model (dual stack) to DS-Lite. Since both traffic sets require a translation function (NAT44), synchronized pool management, routing and management system positioning can allow rapid movement (notwithstanding the technological means to re-provision the subscriber).

Operators should inform their vendors of what technologies they plan to support over the course of the transition to make sure the equipment is suited to support those modes of operation. This is important for both network gear and subscriber premise equipment.

The operator should also plan their overall strategy to meet the target needs of an IPv6-only deployment. As traffic moves to IPv6, the benefits of only a single stack on the access network may eventually justify the removal of IPv4 for simplicity. Planning for this eventual model, no matter how far off this may be, will help the operator embrace this end state when needed.

5.1.6. Phase 0- Foundation: Tools and Management

The operator should thoroughly analyze all provisioning and management systems to develop requirements for each phase. This will include concepts related to the 128-bit IPv6 address, the notation of an assigned IPv6 prefix (Prefix Delegation) and the ability to detect either or both address families when determining if a subscriber has full Internet service.

If an operator stores usage information, this would need to be aggregated to include both the IPv4 and IPv6 as both address families are assigned to the same subscriber. Tools that verify connectivity may need to query the IPv4 and IPv6 addresses.

5.2. Phase 1 - Tunneled IPv6

Tunneled access to IPv6 can be regarded as an early stage transition option by operators. Many network operators can deploy native IPv6 from the access edge to the peering edge fairly quickly but may not be able to offer IPv6 natively to the subscriber edge device. During this period of time, tunneled access to IPv6 is a viable alternative.
to native IPv6. It is also possible that operators may be rolling out IPv6 natively to the subscriber edge but the time involved may be long due to logistics and other factors. Even while carefully rolling out native IPv6, operators can deploy relays for automatic tunneling technologies like 6to4 and Teredo. Where native IPv6 to the access edge is a longer-term project, operators can consider 6RD [RFC5969] as an option to offer in-home IPv6 access. Note that 6to4 and Teredo have different address selection behaviors than 6RD [RFC3484]. Additional guidelines on deploying and supporting 6to4 can be found in [RFC6343].

The operator can deploy 6RD relays into the network and scale them as needed to meet the early subscriber needs of IPv6. Since 6RD requires the upgrade or replacement of CPE devices, the operator may want to ensure that the CPE devices support not just 6RD but native Dual Stack and other tunneling technologies if possible such as DS-Lite [I-D.ietf-v6ops-6204bis]. 6RD clients are becoming available in some retail channel products and within the OEM market. Retail availability of 6RD is important since not all operators control or have influence over what equipment is deployed in the consumer home network. The operator can support 6RD access with unmanaged devices using DHCPv4 option 212 (OPTION_6RD) [RFC5969].

6RD used as an initial transition technology also provides the added benefit of a deterministic IPv6 prefix based on the IPv4 assigned address. Many operational tools are available or have been built to identify what IPv4 (often dynamic) address was assigned to a subscriber CPE. So, a simple tool and/or method can be built to help identify the IPv6 prefix using the knowledge of the assigned IPv4 address.
An operator may choose to not offer internal services over IPv6 if tunneled access to IPv6 is used since some services generate a large amount of traffic. Such traffic may include Video content like IPTV. By limiting how much traffic is delivered over the 6RD connection (if possible), the operator can avoid costly and complex scaling of the relay infrastructure.

5.2.1. 6RD Deployment Considerations

Deploying 6RD can greatly speed up an operator’s ability to support IPv6 to the subscriber network if native IPv6 connectivity cannot be supplied. The speed at which 6RD can be deployed is highlighted in [RFC5569].

The first core consideration is deployment models. 6RD requires the CPE (6RD client) to send traffic to a 6RD relay. These relays can share a common anycast address, or can use unique addresses. Using an anycast model, the operator can deploy all the 6RD relays using the same IPv4 interior service address. As the load increases on the deployed relays, the operator can deploy more relays into the network. The one drawback is that it may be difficult to manage the traffic volume among additional relays, since all 6RD traffic will find the nearest (in terms of IGP cost) relay. Use of multiple relay addresses can help provide more control but has the disadvantage of being more complex to provision. Subsets of CPEs across the network will require and contain different relay information. An alternative approach is to use a hybrid model using multiple anycast service IP Addresses for clusters of 6RD relays, should the operator anticipate massive scaling of the environment. Thus, the operator has multiple vectors by which to scale the service.

```
+--------+
|        |
IPv4 Addr.X   |  6RD   |
|<-- - - -->  |   BR   |
|        |
+--------+
| Client A |          |        |
|--------+
|        |
Separate IPv4 Service Addresses
```

```
+--------+
|        |
IPv4 Addr.X   |  6RD   |
|<-- - - -->  |   BR   |
|        |
+--------+
| Client B |          |        |
|--------+
|        |
|        |
+--------+
|        |
IPv4 Addr.Y   |  6RD   |
|<-- - - -->  |   BR   |
|        |
+--------+
```
Provisioning of the 6RD endpoints is affected by the deployment model chosen (i.e. anycast vs. specific service IP Addresses). Using multiple IP Addresses may require more planning and management, as subscriber equipment will have different sets of data to be provisioned into the devices. The operator may use DHCPv4, manual provisioning or other mechanisms to provide parameters to subscriber equipment.

If the operator manages the CPE, support personnel will need tools able to report the status of the 6RD tunnel. Usage information can be counted on the operator edge, but if it requires source/destination flow details, data must be collected after the 6RD relay (IPv6 side of connection).

6RD [RFC5969], as any tunneling option, is subject to a reduced MTU so operators need to plan to manage this environment.
5.3. Phase 2: Native Dual Stack

Either as a follow-up phase to "Tunneled IPv6" or as an initial step, the operator may deploy native IPv6 down to the CPEs. This phase would then allow for both IPv6 and IPv4 to be natively accessed by the subscriber home network without translation or tunneling. The native Dual Stack phase can be rolled out across the network while the tunneled IPv6 service remains operational, if used. As areas begin to support native IPv6, subscriber home equipment will generally prefer using the IPv6 addresses derived from the delegated IPv6 prefix versus tunneling options such as 6to4 and Teredo as defined in [RFC3484]. Specific care is needed when moving to native Dual Stack from 6RD as documented in [I-D.townsley-v6ops-6rd-sunsetting].

Native Dual Stack is the best option at this point in the transition, and should be sought as soon as possible. During this phase, the operator can confidently move both internal and external services to IPv6. Since there are no translation devices needed for this mode of operation, it transports both protocols (IPv6 and IPv4) efficiently within the network.

5.3.1. Native Dual Stack Deployment Considerations

Native Dual Stack is a very desirable option for the IPv6 transition, if feasible. The operator must enable IPv6 on the network core and peering edge before they attempt to turn on native IPv6 services. Additionally, provisioning and support systems such as DHCPv6, DNS and other functions that support the subscriber’s IPv6 Internet connection need to be in place.

The operator must treat IPv6 connectivity with the same operational importance as IPv4. A poor IPv6 service may be worse than not offering an IPv6 service at all as it will negatively impact the
subscriber’s Internet experience. This may cause users or support personnel to disable IPv6, limiting the subscriber from the benefits of IPv6 connectivity as the network performance improves. New code and IPv6 functionality may cause instability at first. The operator will need to monitor, troubleshoot and resolve issues promptly.

Prefix assignment and routing are new for common residential services. Prefix assignment is straightforward (DHCPv6 using IA_PD), but installation and propagation of routing information for the prefix, especially during access layer instability, is often poorly understood. The operator should develop processes for renumbering subscribers who move to new access nodes.

Operators need to keep track of both the dynamically assigned IPv4 address along with the IPv6 address and prefix. Any additional dynamic elements, such as auto-generated host names, need to be considered and planned for.

5.4. Intermediate Phase for CGN

Acquiring more IPv4 addresses is already challenging, if not impossible; therefore address sharing may be required on the IPv4 path of a Dual Stack deployment. The operator may have a preference to move directly to a transition technology such as DS-Lite [RFC6333] or may use Dual Stack with CGN/NAT444 to facilitate IPv4 connections.

CGN/NAT444 requires IPv4 addressing between the subscriber premise equipment and the operator’s translator which may be facilitated by shared address [RFC6598], private address [RFC1918] or other address space. CGN/NAT444 is only recommended to be used along side IPv6 in a Dual Stack deployment and not on its own. Figure 5 provides a comparative view of a traditional IPv4 path versus one which uses CGN/NAT444.
In the case of native Dual Stack, CGN/NAT444 can be used to assist in extending connectivity for the IPv4 path while the IPv6 path remains native. For endpoints operating in a IPv6+CGN/NAT444 model, the native IPv6 path is available for higher quality connectivity, helping host operation over the network. At the same time, the CGN path may offer a less than optimal performance. These points are also true for DS-Lite.

CGN/NAT444 deployments may make use of a number of address options, which include [RFC1918] or Shared Address Space [RFC6598]. It is also possible that operators may use part of their own RIR assigned address space for CGN zone addressing if [RFC1918] addresses pose technical challenges in their network. It is not recommended that operators use 'squat space', as it may pose additional challenges with filtering and policy control [RFC6598].

5.4.1. CGN Deployment Considerations

CGN is often considered undesirable by operators but required in many cases. An operator who needs to deploy CGN capabilities should consider the impacts of the function to the network. CGN is often deployed in addition to running IPv4 services and should not negatively impact the already working Native IPv4 service. CGNs will be needed at low scale at first and grow to meet the demands based on traffic and connection dynamics of the subscriber, content and network peers.

The operator may want to deploy CGNs more centrally at first and then scale the system as needed. This approach can help conserve costs of
the system limiting the deployed base and scaling it based on actual traffic demand. The operator should use a deployment model and architecture which allows the system to scale as needed.

![Diagram showing CGN Deployment: Centralized vs. Distributed]

Figure 7: CGN Deployment: Centralized vs. Distributed

The operator may be required to log translation information [I-D.ietf-behave-lsn-requirements]. This logging may require significant investment in external systems which ingest, aggregate and report on such information [I-D.donley-behave-deterministic-cgn].

Since CGN has noticeable impacts on certain applications [I-D.donley-nat444-impacts], operators may deploy CGN only for those subscribers who may be less affected by CGN (if possible).

5.5. Phase 3 - IPv6-Only

Once Native IPv6 is widely deployed in the network and well-supported by tools, staff, and processes, an operator may consider supporting only IPv6 to all or some subscriber endpoints. During this final phase, IPv4 connectivity may or may not need to be supported, depending on the conditions of the network, subscriber demand and legacy device requirements. If legacy IPv4 connectivity is still demanded (e.g. for older nodes), DS-Lite [RFC6333] may be used to tunnel the traffic. If IPv4 connectivity is not required, but access to legacy IPv4 content is, then NAT64 [RFC6144][RFC6146] can be used.
5.5.1. DS-Lite

DS-Lite allows continued access for the IPv4 subscriber base using address sharing for IPv4 Internet connectivity, but with only a single layer of translation, compared to CGN/NAT444. This mode of operation also removes the need to directly supply subscriber endpoints with an IPv4 address, potentially simplifying the connectivity to the customer (single address family) and supporting IPv6 only addressing to the CPE.

The operator can also move Dual Stack endpoints to DS-Lite retroactively to help optimize the IPv4 address sharing deployment by removing the IPv4 address assignment and routing component. To minimize traffic needing translation, the operator should have already moved most content to IPv6 before the IPv6-only phase is implemented.

![Diagram of DS-Lite Basic Model]

Figure 8: DS-Lite Basic Model

If the operator had previously decided to enable a CGN/NAT444 deployment, it may be able to co-locate the AFTR and CGN/NAT444 processing functions within a common network location to simplify capacity management and the engineering of flows. This case may be evident in a later transition stages when an operator chooses to optimize its network and IPv6-only operation is feasible.

5.5.2. DS-Lite Deployment Considerations

The same deployment considerations associated with Native IPv6 deployments apply to DS-Lite and NAT64. IPv4 will now be dependent on IPv6 service quality, so the IPv6 network and services must be running well to ensure a quality experience for the end subscriber. Tools and processes will be needed to manage the encapsulated IPv4 service. If flow analysis is required for IPv4 traffic, this may be enabled at a point beyond the AFTR (after de-capsulation) or DS-Lite [RFC6333] aware equipment is used to process traffic midstream.
DS-Lite [RFC6333] also requires client support on the subscriber premises device. The operator must clearly articulate to vendors which technologies will be used at which points, how they interact with each other at the CPE, and how they will be provisioned. As an example, an operator may use 6RD in the outset of the transition, then move to Native Dual Stack followed by DS-Lite.

DS-Lite [RFC6333], as any tunneling option, is subject to a reduced MTU so operators need to plan to manage this environment. Additional considerations for DS-Lite deployments can be found in [I-D.ietf-softwire-dslite-deployment].

5.5.3. NAT64 Deployment Considerations

The deployment of NAT64 assumes the network assigns an IPv6 address to a network endpoint that is translated to an IPv4 address to provide connectivity to IPv4 Internet services and content. Experiments such as the one described in [RFC6586] highlight issues related to IPv6-only deployments due to legacy IPv4 APIs and IPv4 literals. Many of these issues will be resolved by the eventual removal of this undesired legacy behavior. Operational deployment models, considerations and experiences related to NAT64 have been documented in [I-D.chen-v6ops-nat64-experience].
To navigate around some of the limitations of NAT64 when dealing with legacy IPv4 applications, the operator may choose to implement 464XLAT [I-D.ietf-v6ops-464xlat] if possible. As support for IPv6 on subscriber equipment and content increases, the efficiency of NAT64 increases by reducing the need to translate traffic. The NAT64 deployment would see an overall decline in usage as more traffic is promoted to IPv6-to-IPv6 native communication. NAT64 may play an important part of an operator’s late stage transition, as it removes the need to support IPv4 on the access network and provides a solid go-forward networking model.

It should be noted, as with any technology which utilizes address sharing, that the IPv4 public pool sizes (IPv4 transport addresses per [RFC6146]) can pose limits to IPv4 server connectivity for the subscriber base. Operators should be aware that some IPv4 growth in the near term is possible, so IPv4 translation pools need to be monitored.

6. IANA Considerations

No IANA considerations are defined at this time.

7. Security Considerations

Operators should review the documentation related to the technologies selected for IPv6 transition. In those reviews, operators should understand what security considerations are applicable to the chosen technologies. As an example, [RFC6169] should be reviewed to understand security considerations around tunnelling technologies.
8. Acknowledgements

Special thanks to Wes George, Chris Donley, Christian Jacquenet and John Brzozowski for their extensive review and comments.

Thanks to the following people for their textual contributions, guidance and comments: Jason Weil, Gang Chen, Nik Lavorato, John Cianfarani, Chris Donley, Tina TSOU, Fred Baker and Randy Bush.

9. References

9.1. Normative References


9.2. Informative References

[I-D.chen-v6ops-nat64-experience]
Chen, G., Cao, Z., Byrne, C., Xie, C., and D. Binet, "NAT64 Operational Experiences", draft-chen-v6ops-nat64-experience-03 (work in progress), July 2012.

[I-D.donley-behave-deterministic-cgn]

[I-D.donley-nat444-impacts]

[I-D.ietf-behave-lsn-requirements]
Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common requirements for Carrier Grade NATs (CGNs)", draft-ietf-behave-lsn-requirements-09 (work in progress), June 2012.

[I-D.ietf-softwire-dslite-deployment]
Lee, Y., Maglione, R., Williams, C., Jacquenet, C., and M. Boucadair, "Deployment Considerations for Dual-Stack
Internet-Draft          Wireline Incremental IPv6         September 2012

[81x692]Internet-Draft          Wireline Incremental IPv6         September 2012


[I-D.ietf-v6ops-464xlat]
Mawatari, M., Kawashima, M., and C. Byrne, "464XLAT: Combination of Stateful and Stateless Translation",
draft-ietf-v6ops-464xlat-07 (work in progress), July 2012.

[I-D.ietf-v6ops-6204bis]
Singh, H., Beebee, W., Donley, C., and B. Stark, "Basic Requirements for IPv6 Customer Edge Routers",
draft-ietf-v6ops-6204bis-10 (work in progress), May 2012.

[I-D.jjmb-v6ops-comcast-ipv6-experiences]
Brzozowski, J. and C. Griffiths, "Comcast IPv6 Trial/Deployment Experiences",
draft-jjmb-v6ops-comcast-ipv6-experiences-02 (work in progress), October 2011.

[I-D.kuarsingh-v6ops-6to4-provider-managed-tunnel]
Kuarsingh, V., Lee, Y., and O. Vautrin, "6to4 Provider Managed Tunnels",
draft-kuarsingh-v6ops-6to4-provider-managed-tunnel-07 (work in progress), July 2012.

[I-D.townsley-v6ops-6rd-sunsetting]
Cassen, A. and M. Townsley, "6rd Sunsetting",
draft-townsley-v6ops-6rd-sunsetting-00 (work in progress), November 2011.

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


[RFC3484]  Draves, R., "Default Address Selection for Internet

[RFC4380]  Huitema, C., "Teredo: Tunneling IPv6 over UDP through
Network Address Translations (NATs)", RFC 4380,
February 2006.


[RFC6586] Arkko, J. and A. Keranen, "Experiences from an IPv6-Only
Network", RFC 6586, April 2012.

[RFC6598]  Weil, J., Kuarsingh, V., Donley, C., Liljenstolpe, C., and
          M. Azinger, "IANA-Reserved IPv4 Prefix for Shared Address
          Space", BCP 153, RFC 6598, April 2012.

Authors’ Addresses

Victor Kuarsingh (editor)
Rogers Communications
8200 Dixie Road
Brampton, Ontario  L6T 0C1
Canada

Email: victor.kuarsingh@gmail.com
URI:   http://www.rogers.com

Lee Howard
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA  20171
US

Email: lee.howard@twcable.com
URI:   http://www.timewarnercable.com
Basic Requirements for Customer Edge Routers - multihoming and
transition
draft-townsley-troan-ipv6-ce-transitioning-02

Abstract

This document specifies general IPv6 multihoming and specific 6rd transitioning requirements for an IPv6 Customer Edge (CE) router. It also provides an illustrative implementation model for IPv4 multihoming in order to support shared IPv4 transition mechanisms that utilize IPv6 as a transport for IPv4.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

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

This Internet-Draft will expire on June 18, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as
described in the Simplified BSD License.

Table of Contents

1. Introduction ........................................... 3
   1.1. Requirements Language ........................... 4
2. Terminology ............................................ 4
3. IPv6 MultiPrefix Multihoming .......................... 4
   3.1. Multihoming requirements ....................... 5
   3.2. 6rd Sunsetting Requirements ..................... 5
4. IPv4 NAT Multihoming ................................. 6
   4.1. IPv4 Multihoming Data Structures ................. 6
   4.2. IPv4 Packet flow ................................ 7
   4.3. Example RIB Policy for IPv4 to IPv6 Transition .... 7
5. Security Considerations ............................... 8
6. Acknowledgements ..................................... 8
7. IANA Considerations ................................... 9
8. References ............................................. 9
   8.1. Normative References ............................ 9
   8.2. Informative References .......................... 9
Appendix A. Configuration-oriented multihoming .......... 10
Authors’ Addresses .................................. 10
1. Introduction

The CE requirements specified in RFC 6204 are based on a fundamental assumption that a CE router has a single active WAN interface for forwarding IPv4 and IPv6 traffic towards an ISP. The operation of IPv6 via 6rd, IPv6 via Native, IPv4 via DS-lite and IPv4 via Native together, forces us to reconsider this basic assumption.

There are three possible steady-state combinations of "native" and "virtual" (tunneled) dual-stack service models (an IPv6-only service model, while imperative for finalizing the transition to IPv6, is currently out of scope in [RFC6204] and [I-D.ietf-v6ops-6204bis]) that do not break the fundamental assumption of having no more than one WAN interface per IP version.

1. One Native IPv4 and IPv6 interface (Classic Dual-Stack)
2. One Native IPv4 and one Virtual IPv6 interface (6rd, Softwires Hub and Spoke via L2TP, TSP, etc)
3. One Virtual IPv4 and one Native IPv6 interface (DS-Lite, 4rd, etc)

For (1), IPv4 and IPv6 each share a single WAN interface, so there is no problem when enabling one vs. the other.

For (2), when enabling tunneled IPv6 on an existing IPv4-only network there is no significant change in the basic model as each IP version still has its own distinct single WAN interface. Multihoming issues arise when enabling native IPv6 alongside tunneled IPv6 (needed for "6rd sunsetting") as IPv6 may be enabled on two distinct interfaces at the same time.

For (3) there similarly is no problem when enabling tunneled IPv4 on an existing IPv6-only network, and we understand that there are greenfield deployments just like this happening. Multihoming issues arise when enabling tunneled IPv4 on a network that has native IPv4 available at the same time.

The multihoming model described in this document assumes the operator or user can configure any WAN interface type at any time. The CE follows forwarding rules defined in this document in order to ensure packets make it out the right interface on WAN egress, and liberally accepts packets on WAN ingress. This is "classic multihoming" and should work for any order of planned incremental transition steps, as well as failover and/or transient situations.

While this authors of this document believe that the forward-oriented
model, there is another approach which we identify as "Configuration-oriented" and describe briefly in Appendix A.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Terminology

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRIB</td>
<td>A Source Address Routing Information Base containing an entry per delegated prefix. Each entry points to one or more Destination Address Routing Tables (DRIB).</td>
</tr>
<tr>
<td>DRIB</td>
<td>A Destination Address Routing Information Base used for destination address longest matching lookups. Each entry points to one or more next-hops.</td>
</tr>
<tr>
<td>NPIB</td>
<td>Network Address and Port Translation (NAPT) Information Base used binding &quot;flows&quot; to an egress WAN interface.</td>
</tr>
<tr>
<td>NPIB entry</td>
<td>The address and port mapping state [RFC4787] at the NAT necessary for network address and port translation.</td>
</tr>
</tbody>
</table>

3. IPv6 MultiPrefix Multihoming

A multihomed, multiprefix, IPv6 CE router has multiple WAN interfaces connecting it to one or more Service Providers. The interfaces may be "real" or "virtual" in the case of tunneling technology such as 6rd [RFC5969]. The CE router receives one or more delegated prefixes, each associated with one or more WAN interfaces. The CE router has a single SRIB, and one DRIB associated with each WAN Interface.

WAN interfaces are used to send Ingress traffic from the Internet to the End-User, and Egress traffic from the End-User network to the Internet. Ingress traffic may be received on any active interface at any time. Egress traffic follows a set of rules within the CE in order to choose the proper WAN interface. This is important not only in order to choose the best path, but also because the networks that the CE are connected to typically employ source address verification
Packets arriving at the CE have an IPv6 source address chosen by the host [RFC3484]. The SRIB contains an entry for each delegated prefix with a pointer to one or more DRIBs. A longest matching lookup based upon the source address of each arriving packet is performed within the SRIB to determine the DRIB(s). The egress WAN interface to use for sending a packet is then chosen by performing a longest matching lookup within the resulting DRIB(s).

3.1. Multihoming requirements

MH-1: An IPv6 CE router MUST create a separate DRIB for each WAN interface (real or virtual) and installs a route for the associated delegated prefix, default route and more specific routes.

MH-2: An IPv6 CE router MUST create an SRIB containing entries for associated delegated prefixes. Each entry points to one or more DRIBs. An entry points to multiple DRIBs only in the case where an identical delegated prefix is associated with multiple WAN interfaces.

MH-3: When forwarding a packet from a LAN interface, the CE router MUST do a longest matching lookup based on the packet’s Source Address in the SRIB. A Destination Address lookup is then performed in the corresponding DRIB or DRIBs. When there are multiple equal matches, the route with the lowest cost is chosen.

3.2. 6rd Sunsetting Requirements

6RDS-1: Multihoming as defined in section Section 3 MUST be supported, allowing 6rd and native packets to be sent and received as long as 6rd configuration is provided by the ISP.

6RDS-2: By default, the 6rd virtual interface MUST be assigned a higher routing cost than a native IPv6 interface.

6RDS-3: The IPv6 CE router MUST support that 6rd and native IPv6 delegated prefixes are identical or different, and operate as defined in the multihoming section.
4. IPv4 NAT Multihoming

This section describes a general implementation model used to illustrate CE IPv4 multihoming, alongside specifics for using the model to support IPv6 transition technologies aimed at delivering IPv4 within an IPv6 transport (e.g., DS-Lite).

A multihomed IPv4 CE router has multiple "physical" or "virtual" WAN interfaces connecting it to one or more Service Providers. WAN interfaces are used to send Ingress traffic from the Internet to the End-User, and Egress traffic from the End-User network to the Internet. Each WAN interface may be configured with a single public IPv4 address, private IPv4 address [RFC1918], a shared IPv4 address [A+P], or no IPv4 address at all.

An IPv4 CE router WAN interface is often configured in the same manner as a single host, i.e., with a single 32-bit IPv4 address. A CE router NAPT function, in turn, allows more than one device within the end-user network to appear as a single host with a single IPv4 address facing the ISP network. The CE router NAPT function is responsible for rewriting IPv4 headers with the address assigned to the associated WAN interface with the expectation that return traffic will be sent back to that address.

IPv4 WAN interfaces which are not configured with an IPv4 address by the ISP (e.g., DS-Lite, L2-aware NAT), bypass the translation function of the NAPT when forwarding traffic. In this case, the ISP’s centralized NAPT function is responsible for rewriting IPv4 headers with a source address which ensures return traffic will reach the proper NAPT binding within the ISP.

4.1. IPv4 Multihoming Data Structures

The CE router has a single NAPT Information Base (NPIB) which consists of Dynamic and Configured entries. A WAN interface provisioned with an IPv4 address has an associated NAPT function which examines packet flow and programs the NPIB with Dynamic entries as they are identified. For example, an active TCP session on the CE correlates to a single Dynamic entry within the NPIB. A Dynamic entry in the NPIB table unambiguously identifies packets that are associated with it when an NPIB Lookup is performed.

Configured NPIB entries allow for rules to be specified which direct traffic in a specific manner. Unlike Dynamic entries, Configured entries allow for "wildcards", and may be end-user configured or configured by a protocol such as NAT-PMP, UPnP, PCP, etc. Packets directed by configured entries may or may not instantiate Dynamic NPIB entries for specific flows.
Finally, the CE router also has a single IPv4 Routing Information
Base (RIB), containing IPv4 routes learned from active LAN and WAN
interfaces. The RIB is only consulted for packets that fail to match
an NPIB entry. The RIB supports a classic IPv4 routing function,
including use of the longest matching algorithm and preferences that
may be assigned to each interface. A RIB lookup ultimately resolves
to an output interface which, in turn may have an NAPT function
enabled. If the output interface has an NAPT function enabled, it is
responsible for programming the NPIB with a new Dynamic entry and
translating the packet before sending.

4.2. IPv4 Packet flow

Ingress (from the Internet to the End-User) traffic may be received
on any active interface at any time. Egress (from the End-User to
the Internet) traffic follows a set of rules within the CE in order
to choose the proper WAN interface and associated NAPT function on
that interface if present.

Egress packets arriving at the CE trigger a lookup in the NPIB. A
successful match on a Dynamic entry in the NPIB provides all
information necessary to translate and send a packet out the proper
interface (e.g., no additional lookup in the RIB or elsewhere is
required). Packets which do not have a matching NPIB entry but match
a Configured entry are treated similarly. Packets which fail the
NPIB lookup entirely are sent to the RIB.

The RIB performs a classic IPv4 longest-matching routing lookup based
on the destination address of the IPv4 packet. If more than one
interface is selected (which will be the case when more than one
active WAN interface programs a default route in the RIB), then
packets are sent via the interface with the highest configured
preference. If the preference is the same, packets may be load-
balanced. After selecting the proper output interface for the
packet, the packet is either sent immediately or translated and sent
if a NAPT function is enabled. If NAPT function is enabled on that
interface, it is responsible for programming the NPIB with a new
Dynamic entry and translating the packet before sending.

4.3. Example RIB Policy for IPv4 to IPv6 Transition

Interface preferences in the RIB allow policy definition for choosing
one type of interface over another. Well-defined defaults which
encourage transition to IPv6, less use of NAPT, and more distributed
state may be defined. The policy may be represented as a simple
table which may be altered by the operator or user without any change
to the CE packet forwarding implementation itself.
In this example, an interface with a higher preference value is preferred over an interface with a lower preference value. Weighted values are assigned according to the following basic principles for IPv4 interface selection:

1. IPv6 transport is preferred over any other.
2. Less address translation occurrences is preferred over more. [RFC5864][I-D.donley-nat444-impacts]
3. The closer the state is to the edge, the better.[RFC1958]

In the case of DS-Lite and Native IPv4 configuration being present at the same time, DS-Lite would be preferred as it uses IPv6 transport and Native IPv4 does not. When transitioning an active dual-stack network to DS-Lite, this means that when the DS-Lite IPv4 interface is made active, traffic that does not match an active entry in the NPIB table would be directed over the DS-Lite tunnel. As entries in the NPIB table naturally time out, or if the Native interface is deactivated, the CGN within the DS-Lite AFTR takes over the NAPT state of the CE router. In the event the DS-Lite tunnel fails, the Native IPv4 interface and local NAPT will naturally begin taking over.

The above principles may be applied to other methods of IPv4 transport as well. The following table indicates a basic ordering (least to most preferred) for some of the known IPv4 extension and IPv6 transition mechanisms under development today.

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>#1 (100)</th>
<th>#2 (10)</th>
<th>#3 (1)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGN</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>111</td>
</tr>
<tr>
<td>SD-NAT</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>112</td>
</tr>
<tr>
<td>Native IPv4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>122</td>
</tr>
<tr>
<td>MAP-T</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>212</td>
</tr>
<tr>
<td>DS-lite</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>221</td>
</tr>
<tr>
<td>MAP-E</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>222</td>
</tr>
</tbody>
</table>

Table 1: IPv4 Preference Table

5. Security Considerations

6. Acknowledgements
7. IANA Considerations

This memo includes no request to IANA.

8. References

8.1. Normative References


8.2. Informative References


[RFC5864] Allbery, R., "DNS SRV Resource Records for AFS", RFC 5864,
Appendix A. Configuration-oriented multihoming

This method attempts to actively avoid multihoming by forcing only a single configured WAN egress interface to be active at any time for a given IP version. For this to work, one of the following assumptions must hold.

a. From the perspective of the CE router, the network supports only one type of interface for a given IP version, or

b. The CE router is configured in advance of any IP configuration to support only one type of interface for a given IP version, or

c. The CE router goes through an ordered set of configuration attempts in series, each requiring a timeout before moving to the next. Transition-oriented changes after steady-state is reached will require "reboot" to go through the ordered process from scratch.

d. The CE router chooses one type of interface and shuts down all others based on a predetermined priority when more than one interface with the same IP version is configured. This allows parallel configuration attempts and changes after reaching steady-state, but requires the CE router and network to manage a "flash cut" from one configured interface to the other and may be prone to tricky race-conditions.

Authors’ Addresses

Mark Townsley
cisco

Email: mark@townsley.net