V6OPS
Internet-Draft
Intended status: Informational
Expires: August 26, 2012

B. Carpenter
Univ. of Auckland
S. Jiang
Huawei Technologies Co., Ltd

February 23, 2012

IPv6 Guidance for Internet Content and Application Service Providers
draft-carpenter-v6ops-icp-guidance-03

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.

Table of Contents

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 importance 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

the discussion of cookies in Section 8.2).

Neverthless, 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

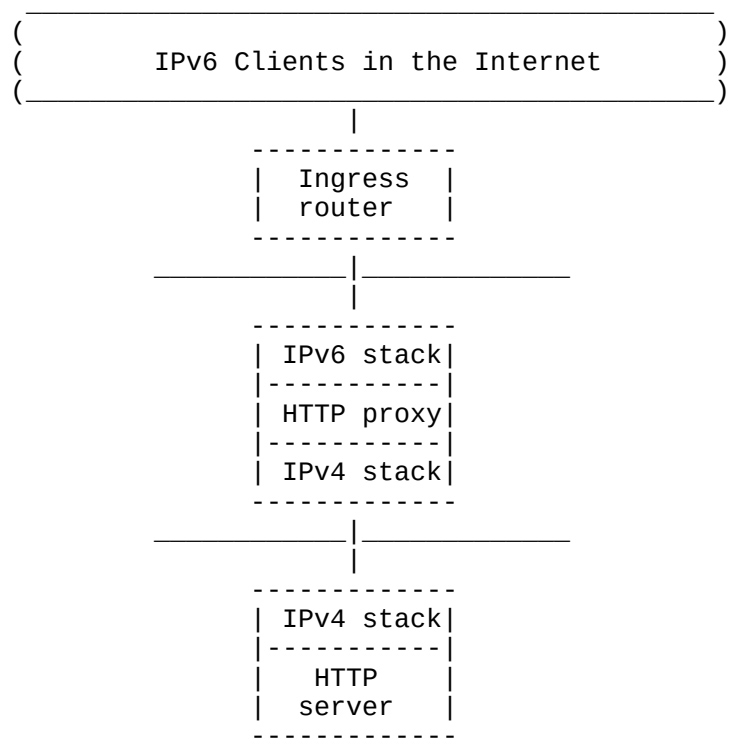
This is largely a case of "just do it." Each externally visible host (or virtual host) that has an A record for its IPv4 address needs an AAAA record [RFC3596] for its IPv6 address, and a reverse entry if applicable. One important detail is that some clients (especially Windows XP) can only resolve DNS 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:



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 middlware 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 guarentee 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

[RFC2080] Malkin, G. and R. Minnear, "RIPng for IPv6", RFC 2080, January 1997.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C.,

and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.

[RFC3596] Thomson, S., Huitema, C., Ksinant, V., and M. Souissi, "DNS Extensions to Support IP Version 6", RFC 3596, October 2003.

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.

[RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.

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

[RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, October 2008.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

[RFC5996] Kaufman, C., Hoffman, P., Nir, Y., and P. Eronen, "Internet Key Exchange Protocol Version 2 (IKEv2)", RFC 5996, September 2010.

17.2. Informative References

[I-D.carpenter-6renum-static-problem]
Carpenter, B. and S. Jiang, "Problem Statement for Renumbering IPv6 Hosts with Static Addresses", draft-carpenter-6renum-static-problem-01 (work in progress), December 2011.

[I-D.ietf-v6ops-happy-eyeballs]
Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", draft-ietf-v6ops-happy-eyeballs-07 (work in progress), December 2011.

[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]
Jiang, S., Liu, B., and B. Carpenter, "IPv6 Enterprise Network Renumbering Scenarios and Guidelines", draft-jiang-6renum-enterprise-02 (work in progress), December 2011.

[RFC2081] Malkin, G., "RIPng Protocol Applicability Statement", RFC 2081, January 1997.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

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

[RFC3068] Huitema, C., "An Anycast Prefix for 6to4 Relay Routers", RFC 3068, June 2001.

[RFC4038] Shin, M-K., Hong, Y-G., Hagino, J., Savola, P., and E. Castro, "Application Aspects of IPv6 Transition", RFC 4038, March 2005.

[RFC4192] Baker, F., Lear, E., and R. Droms, "Procedures for Renumbering an IPv6 Network without a Flag Day", RFC 4192, September 2005.

[RFC4864] Van de Velde, G., Hain, T., Droms, R., Carpenter, B., and E. Klein, "Local Network Protection for IPv6", RFC 4864, May 2007.

[RFC4890] Davies, E. and J. Mohacsi, "Recommendations for Filtering ICMPv6 Messages in Firewalls", RFC 4890, May 2007.

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

IPv6", RFC 4941, September 2007.

[RFC5157] Chown, T., "IPv6 Implications for Network Scanning", RFC 5157, March 2008.

[RFC5375] Van de Velde, G., Popoviciu, C., Chown, T., Bonness, O., and C. Hahn, "IPv6 Unicast Address Assignment Considerations", RFC 5375, December 2008.

[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.

[RFC6180] Arkko, J. and F. Baker, "Guidelines for Using IPv6 Transition Mechanisms during IPv6 Deployment", RFC 6180, May 2011.

[RFC6296] Wasserman, M. and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, June 2011.

[RFC6343] Carpenter, B., "Advisory Guidelines for 6to4 Deployment", RFC 6343, August 2011.

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

V6OPS
Internet-Draft
Intended status: Informational
Expires: September 7, 2012

B. Carpenter
Univ. of Auckland
S. Jiang
Huawei Technologies Co., Ltd
W. Tarreau
Exceliance

March 6, 2012

Using the IPv6 Flow Label for Server Load Balancing
draft-carpenter-v6ops-label-balance-02

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

described in the Simplified BSD License.

Table of Contents

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.

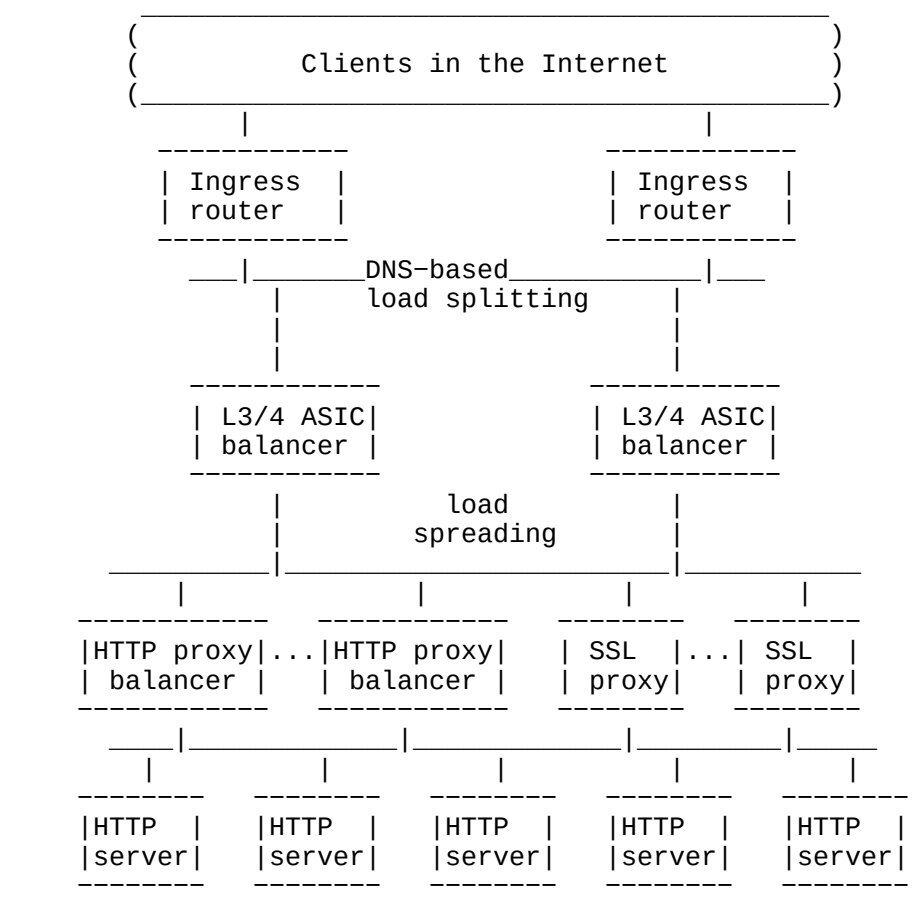
1. Balancers use various techniques to redirect traffic to a specific target server.

   - 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:
  1. 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.
  2. 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 targetting 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 nothing though that what is important to save server-side resources is wide enough adoption. Most of todays 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 outweight 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 explicitely 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.

draft-carpenter-v6ops-label-balance-01: updated with community comments, additional author, 2012-01-17.

draft-carpenter-v6ops-label-balance-00: original version, 2011-10-13.

8. References

8.1. Normative References

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, November 2011.

8.2. Informative References

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

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

Multicast Next-Hop Selection", RFC 2991, November 2000.

[RFC4864] Van de Velde, G., Hain, T., Droms, R., Carpenter, B., and E. Klein, "Local Network Protection for IPv6", RFC 4864, May 2007.

[RFC6296] Wasserman, M. and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, June 2011.

[RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, November 2011.

[Tarreau] Tarreau, W., "Making applications scalable with load balancing", 2006, <http://1wt.eu/articles/2006_lb/>.

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

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 3, 2012

G. Chen
China Mobile

Oct 2011

NAT64 Operational Considerations
draft-chen-v6ops-nat64-cpe-03

Abstract

The document has summarized NAT64 usages on different modes, in which NAT64 may serve for a large-scale network or would give enterprise or residential service opportunities to be accessed by IPv6 remote subscribers. The document has described different operations for each usage and proposed operational considerations for each particular NAT64-mode.

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 3, 2012.

Copyright Notice

Table of Contents

1. Introduction

With fast developments of global Internet, the demands for IP address are rapidly increasing at present.This year, IANA announced that the global free pool of IPv4 depleted on 3 February. IPv6 is the only real option on the table. Operators have to accelerate the process of deploying IPv6 networks in order to address IP address strains.IPv6 deployment normally involves a step-wise approach where parts of the network should properly updated gradually. As IPv6 deployment progresses it may be simpler for operators and ICP/ISP to employ NAT64[RFC6146] functionalities at edge of IPv4 and IPv6 networks, since a significant part of network will still stay in IPv4 for long time. Especially, NAT64 could facilitate large ICP/ISP IPv6 transition process by eliminating upgrations of tremendous legacy IPv4 servers. Therefore, it’s quite popular to deploy NAT64 at the front of IDC to shift the entire service to be IPv6-enable.

Depending on different usage, NAT64 could be deployed on different places. The document has summarized NAT64 usages on different modes. Considering the existing deployment approaches, the memo has proposed different operational consideration for each particular NAT64-mode.

2. NAT64-CGN Deployment

2.1. Deployment in IDC

NAT has widely used in data center environments whenever IDC have to make your IPv4-only content available to IPv6 clients.

Figure 1 illustrates the usage where an IPv6-only host would like to initiate communications with IDC in IPv4 domain through NAT64. The NAT64 would accept IPv6 incoming session and distribute them to multiple IPv4 servers.

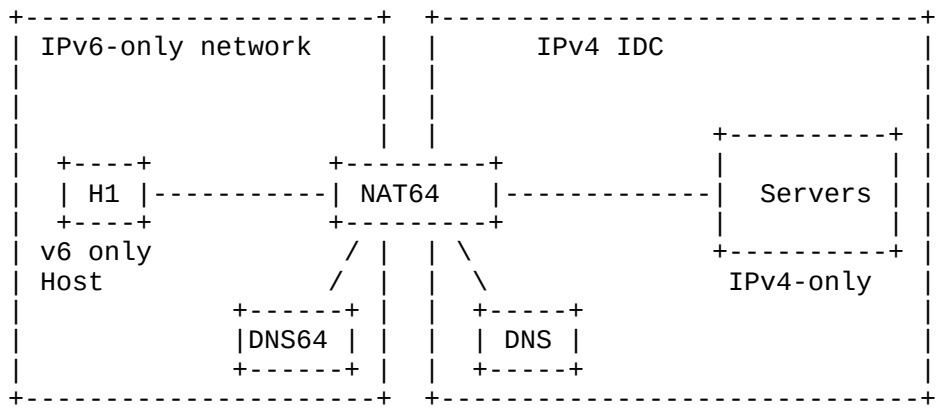


Figure 1: NAT64-CGN Mode Usage

NAT64 device in IDC may also take responsibilities of load balancer, which can accept incoming TCP/UDP sessions on a single virtual IPv6 interface or multiple IPv6 interfaces. Afterwards, it distributes them according to a specific algorithm it uses to multiple IPv4 servers. Ideally you could have a mix of IPv4 and IPv6 servers sitting behind the virtual IPv6 address.

Therein, NAT64 has to pick a new source IPv4 address and associated port number from local IPv4 address pool. DNS64 is a logical function that synthesizes DNS resource records(e.g., AAAA records containing IPv6 addresses) from DNS resource records actually contained in the DNS (e.g., A records containing IPv4 addresses).

2.2. Connecting with IPv4 Internet

NAT64 may also be used to connecting IPv6 users with IPv4 Internet. In this cases, NAT64 could colocated with BNG or Core Router to map legacy IPv4 servers into a NAT64 prefix and performs 6-to-4 address.

Therein, NAT64 would perform protocol translation mechanism and address translation mechanism. Protocol translation from an IPv4 packet header to an IPv6 packet header and vice versa is performed according to the IP/ICMP Translation Algorithm [RFC6145]. Address translation maps IPv6 transport addresses to IPv4 transport addresses and vice versa.

Following illustrates normal process for this usage.

- Step1: IPv6-only host performs an AAAA DNS query to DNS64 for the IPv6 address of the Pv4-only sever.
- Step2: DNS64 could not find the IPv6 address of the IPv4-only sever. So it tries to get the IPv4 address of the Pv4-only sever by sending A DNS query to DNS4.
- Step3: DNS4 return the A record to the DNS64.
- Step4: DNS64 map the IPv4 address to IPv6 address and send a synthetic AAAA record which is translated from A record to IPv6-only host.
- Step5: IPv6-only host send the IPv6 packet to the NAT64. NAT64 translates the IPv6 packet to IPv4 packet and send it to IPv4-only server.

2.3. NAT64-CGN Mode Requirements

According to above description for NAT64-CGN, the NAT64-CGN requirements are listed as following.

NAT64-CGN-R1: Each NAT64 device MUST have at least one unicast IPv6 prefix assigned to it, denoted Pref64::/n.

NAT64-CGN-R2:A NAT64 MUST have one or more unicast IPv4 addresses assigned to it.

NAT64-CGN-R3:Irrespective of the transport protocol used, the NAT64 MUST silently discard all incoming IPv6 packets containing a source address that contains the Pref64::/n.

NAT64-CGN-R4:The NAT64 MUST only process incoming IPv6 packets that contain a destination address that contains Pref64::/n. Likewise, the NAT64 MUST only process incoming IPv4 packets that contain a destination address that belongs to the IPv4 pool assigned to the NAT64.

NAT64-CGN-R5:NAT64 MUST support the algorithm for generating IPv6 representations of IPv4 addresses defined in RFC6052 as Address Translation Algorithms.

NAT64-CGN-R6:For incoming packets carrying TCP or UDP fragments with a non-zero checksum, NAT64 MAY elect to queue the fragments as they arrive and translate all fragments at the same time.

NAT64-CGN-R7: For incoming IPv4 packets carrying UDP packets with a zero checksum, if the NAT64 has enough resources, the NAT64 MUST

reassemble the packets and MUST calculate the checksum. If the NAT64 does not have enough resources, then it MUST silently discard the packets.

NAT64-CGN-R8: The NAT64 MAY require that the UDP, TCP, or ICMP header be completely contained within the fragment that contains fragment offset equal to zero.

NAT64-CGN-R9:The NAT64 MUST limit the amount of resources devoted to the storage of fragmented packets in order to protect from DoS attacks.

NAT64-CGN-R10:The NAT64 MUST make fragmentation process when MTU of incoming IPv4 traffic exceed maximum MTU on IPv6 side.

NAT64-CGN-R11:The NAT64 MAY let hosts and applications know IPv6 prefix used by the NAT64 and DNS64 so as to hosts have knowledge whether synthetic IPv6 address is targeted.

NAT64-CGN-R12:The NAT64 MAY decouple with DNS64 in order to establish communication with IPv4-only servers.

NAT64-CGN-R13:The NAT64 MAY take load-balancing functionalities incorporating with DNS64.

3. NAT64-CE Mode

NAT64-CE mode represents usages where there NAT64 is closed to customer edges, like enterprise network edge or residential network edge.

3.1. NAT64 at Enterprise Network Edge

Some enterprise would like to offers their employees with IPv6 access. However, the service may still stay in IPv4 domain. NAT64 useges in enterprise network could help shift all enterprise service to be IPv6 enable.

Figure 2 illustrates a network usage where an IPv6-only client attached to a dual-stack network, but the destination server is running on a private site where there is NAT64-CE numbered with public IPv6 addresses and private IPv4 addresses.

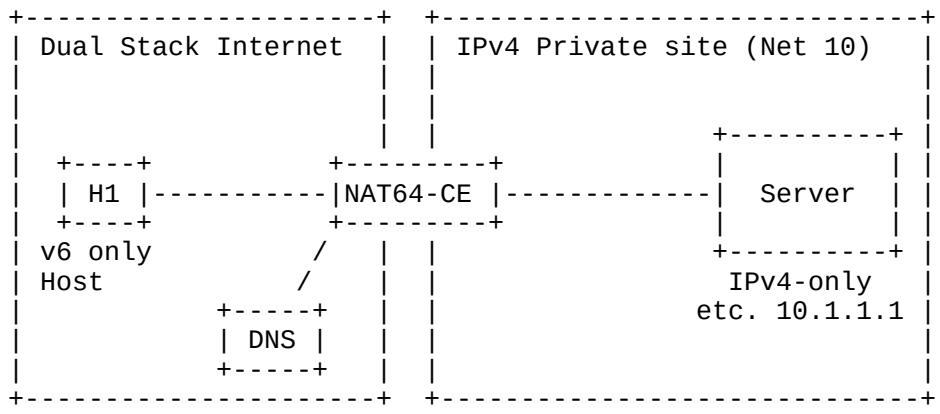


Figure 2: NAT64-CPE Mode Usage

3.2. NAT64 at Residential Network Edge

Residential servers are usually going beyond the operator's management. They may not able to IPv6-enable due to limitations of application supporting. In this case, ISP is still assigning private IPv4 address to servers. However, The nature of private IPv4 would block the end-to-end bi-directional communications. On the other hand, IPv6 will bring end-to-end benefits to operators. NAT64-CPE mode could let IPv6 users to access such IPv6-disable services in residential areas.

This scenario may appears in ISP network for several cases. As the instances, visitors go through distant network to take care of family affairs, like monitoring house security via residential camera, manipulating household appliances remotely prior to comeback home.

4. Security Considerations

Essentially, there are strong demands to have thorough security mechanism to prevent privacy invasion in NAT64-CPE scenario. The detailed considerations need to be further identified.

5. IANA Considerations

This memo includes no request to IANA.

6. Normative References

[RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.

[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.

[RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 2011.

[RFC6204] Singh, H., Beebee, W., Donley, C., Stark, B., and O. Troan, "Basic Requirements for IPv6 Customer Edge Routers", RFC 6204, April 2011.

Author's Address

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

Email: chengang@chinamobile.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 6, 2013

H. Hazeyama
NAIST
R. Hiromi
Intec Inc.
T. Ishihara
Univ. of Tokyo
O. Nakamura
WIDE Project

October 3, 2012

Experiences from IPv6-Only Networks with Transition Technologies in the WIDE Camp Autumn 2012

draft-hazeyama-widecamp-ipv6-only-experience-02

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.

Table of Contents

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

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].

"DNS64" refers DNS extensions to use NAT64 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 [RFC2131].

"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.

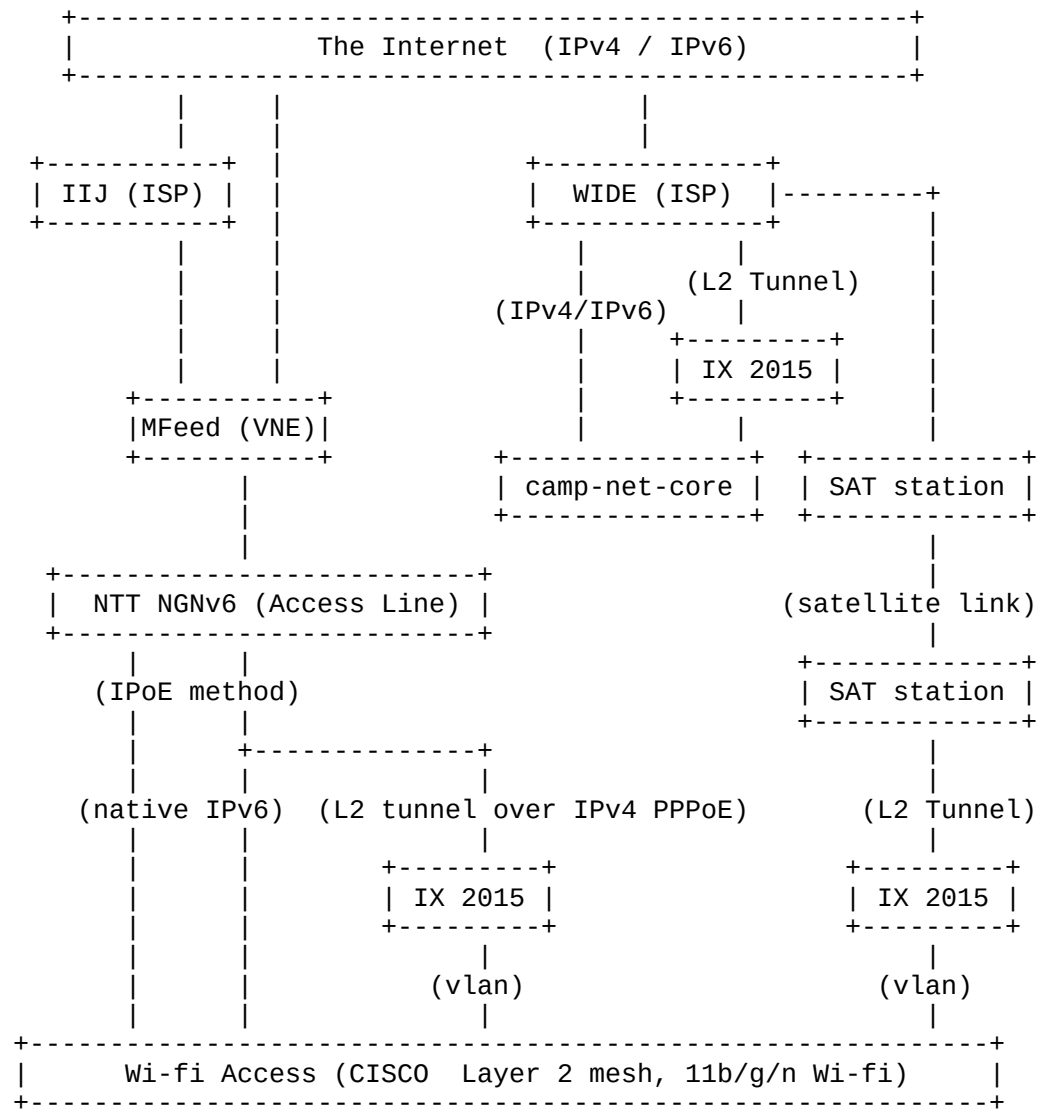
From September 2th (the preparation day) to September 4th, we used the RA method for the external connectivity. Figure 2 represents details of the IPv6 only network by RA method. From September 5th to September 6th, we changed the external connectivity to the DHCP-PD method.

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;

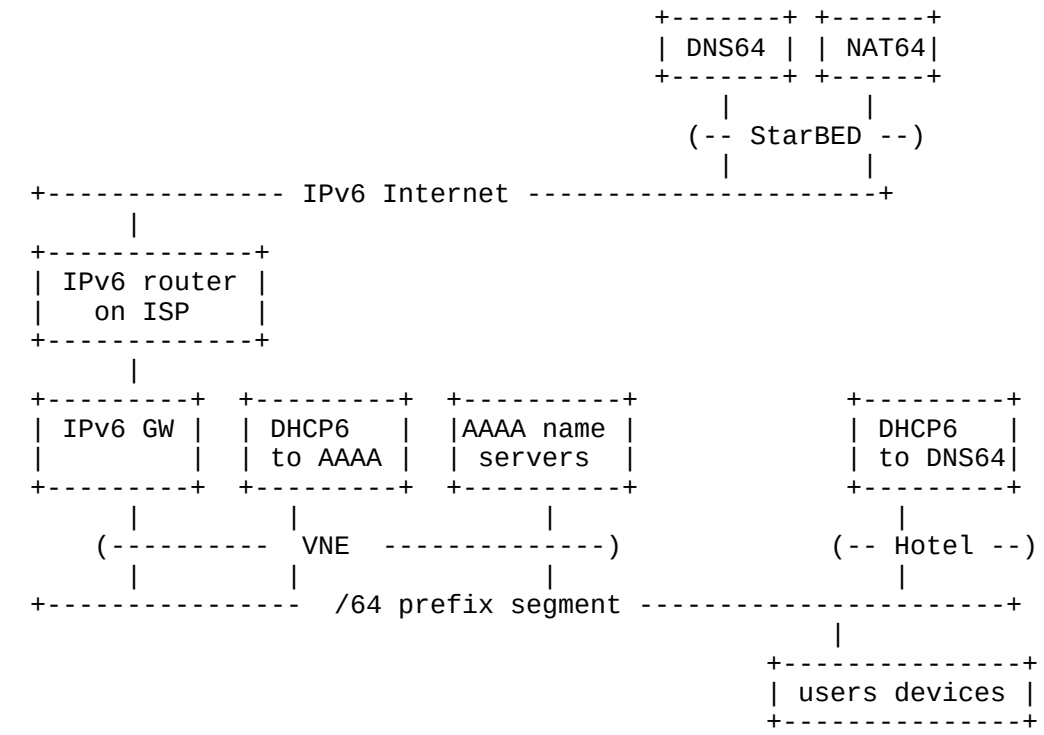
- Those Advertise messages with the highest server preference value are preferred over all other Advertise messages.

- 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.

Device Type	OS Type	# of devices (%)
PC/AT Note PC	Windows 7	16 (17.0 %)
PC/AT Note PC	NetBSD	2 (2.1 %)
PC/AT Note PC	Linux	4 (4.3 %)
Apple Note PC	Mountain Lion	15 (16.0 %)
Apple Note PC	Lion	18 (19.1%)
Apple Note PC	Snow Leopard	9 (9.6 %)
Apple Note PC	Windows 7 (Bootcamp)	3 (3.2 %)
iPhone / iPod	iOS 5	9 (9.6 %)
Android Phone	Android OS 4	3 (3.2 %)
Android Phone	Android OS 2	4 (4.3 %)
Android Phone	Android OS 1	1 (1.0 %)
iPad	iOS 6	1 (1.0 %)
iPad	iOS 5	6 (6.4 %)
Android Tablet	Android OS 4	2 (2.1 %)
Kindle	Kindle 3.3	1 (1.0 %)
Total		94

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

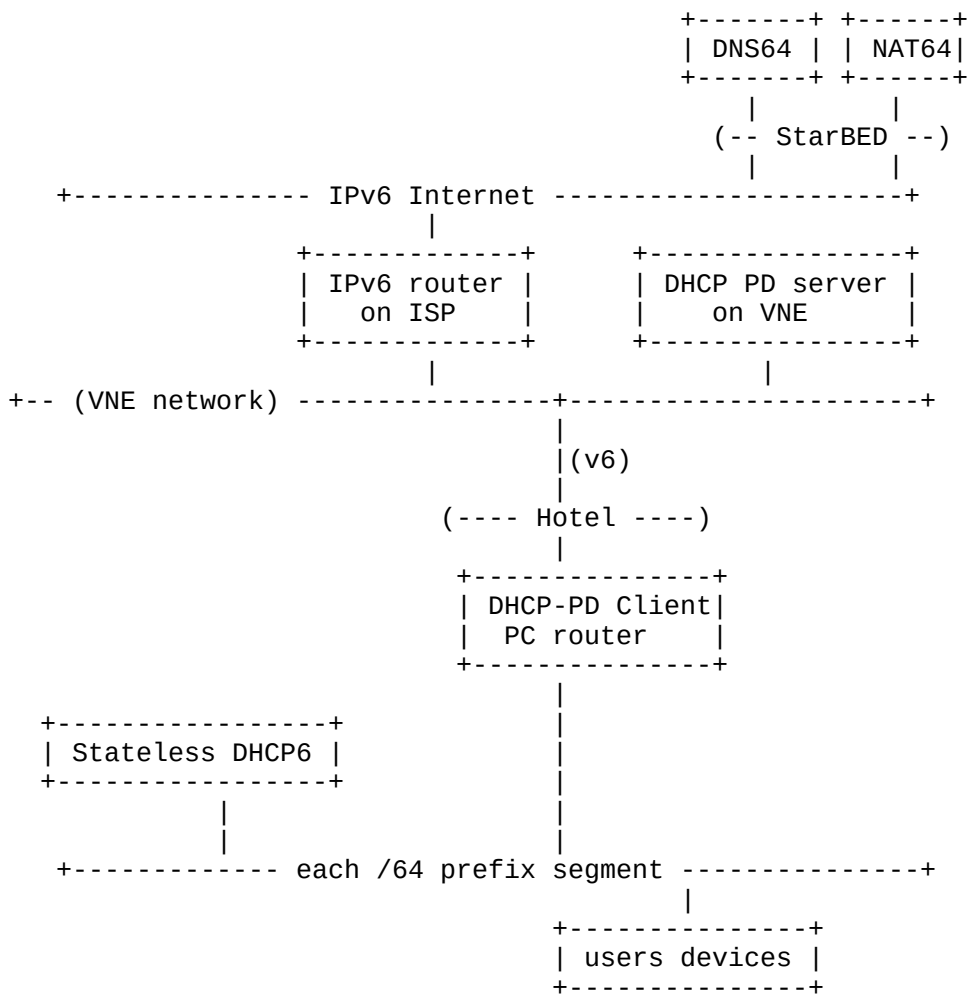
Many users reported inconveniences due to timeout / fallback problems. Root causes were roughly categorized into 1) troubles of DNS64, 2) incapability of IPv6 and of DNS64 on various servers and applications mentioned in [RFC4074] and [RFC6586], 3) incapability of DHCP6 client and / or IPv4 dependency 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.



Basic Network Topology on DHCP-PD method (v6only-basic)

Figure 3

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

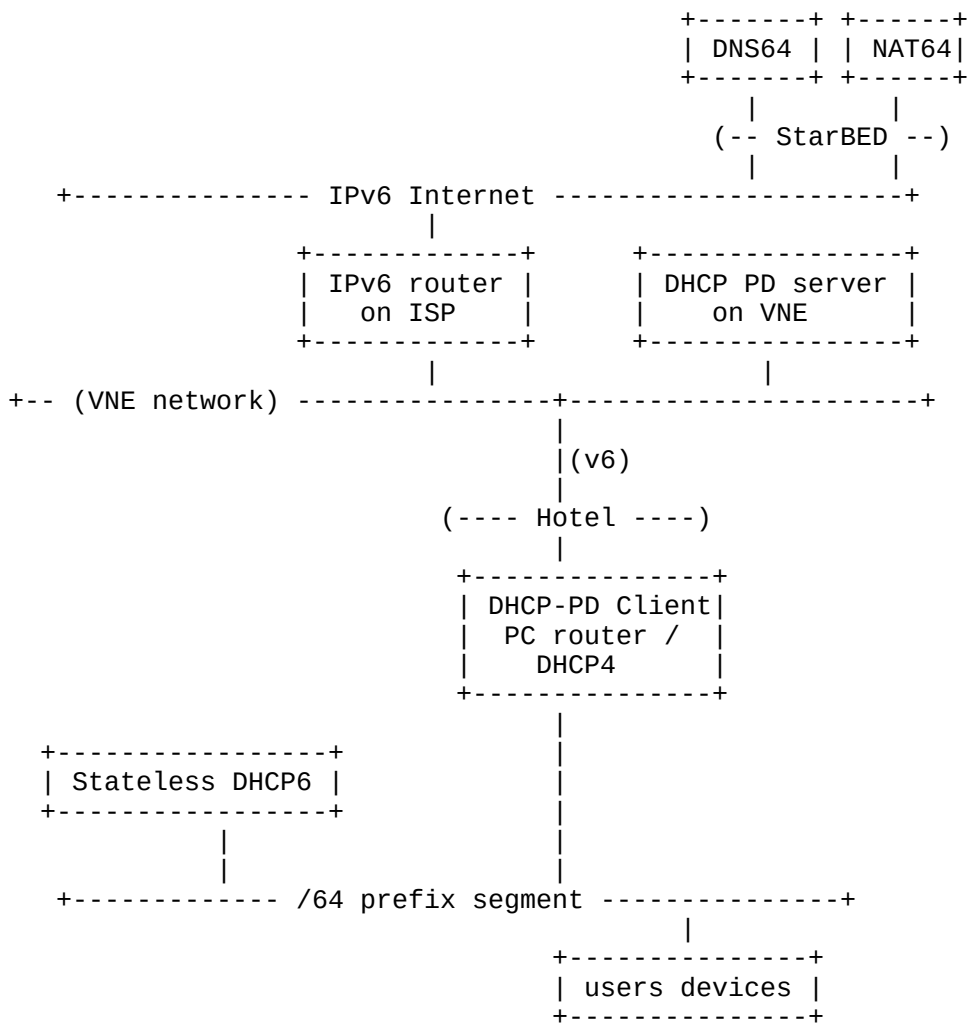
Windows XP, older Mac OS X (Snow Leopard and older) and Android OS required an IPv4 address for an DNS server even when they can use IPv6. In an IPv6 only network, DNS information should be gotten via DHCP6, these OSes did not support DHCP6 client. Also, Android cannot

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.



Test Topology on Experiment 1 (v6only-fallback)

Figure 4

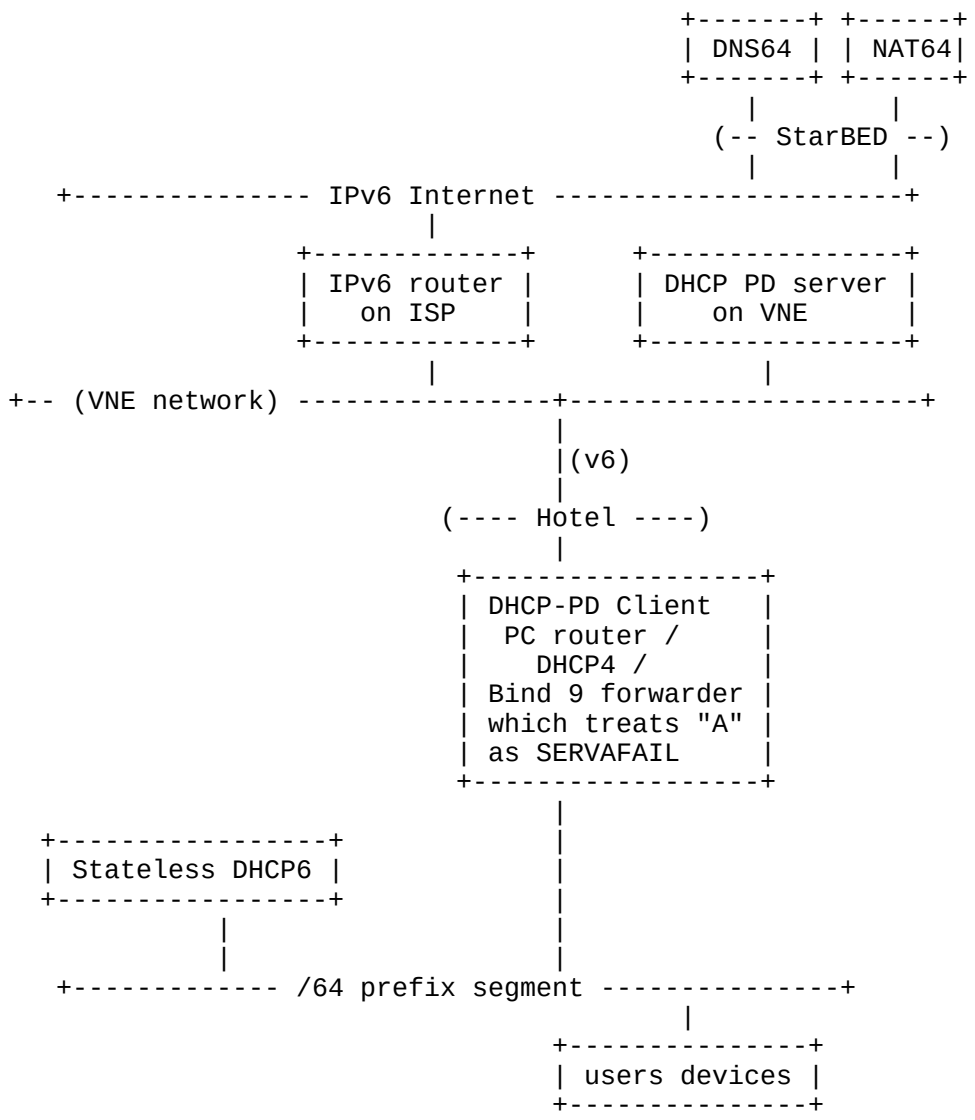
4.2.3.2. Result

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

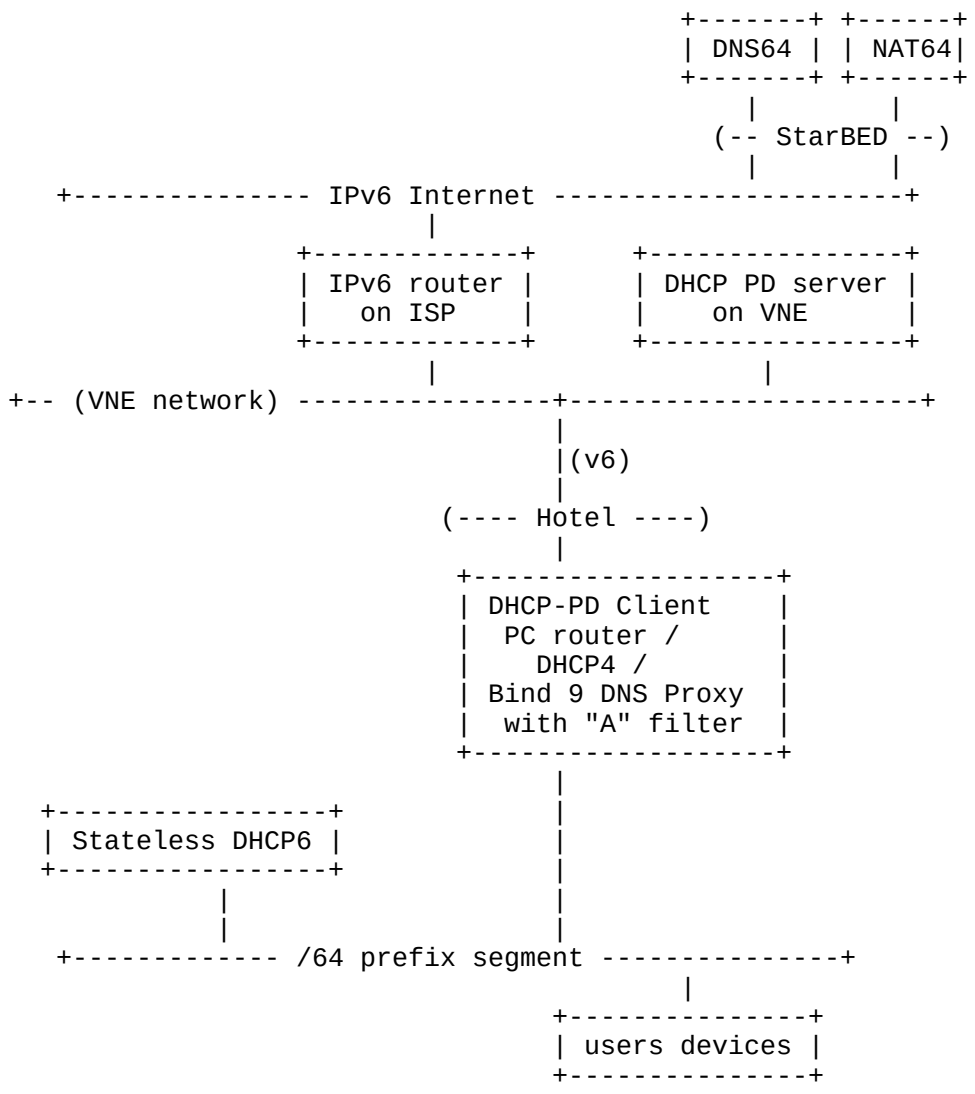
4.2.4.2. Result

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.



Test Topology on Experiment 3 (v6only-fallback)

Figure 6

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.

  o 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.

  o 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.

  o 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

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

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

Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.

[RFC3736] Droms, R., "Stateless Dynamic Host Configuration Protocol (DHCP) Service for IPv6", RFC 3736, April 2004.

[RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

8.2. Informative References

[I-D.draft-ietf-softwire-map-02]
Troan, O., Bao, C., Matsushima, S., and T. Murakami, "Mapping of Address and Port with Encapsulation (MAP)", September 5, 2012, <draft-ietf-softwire-map-02 (work in progress)>.

[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]
Matsuhira, N., Horiba, K., Ueno, Y., and O. Nakamura, "SA46T Address Translator", July 2011, <draft-matsuhira-sa46t-at-00 (work in progress)>.

[I-D.draft-matsuhira-sa46t-spec]
Matsuhira, N., "Stateless Automatic IPv4 over IPv6 Tunneling: Specification", July 2012, <draft-matsuhira-sa46t-spec-05 (work in progress)>.

[I-D.draft-murakami-softwire-4rd]
Murakami, T., Troan, O., and S. Matsushima, "Stateless Automatic IPv4 over IPv6 Tunneling: Specification", September 2011, <draft-murakami-softwire-4rd-01 (work in progress)>.

[RFC0894] Hornig, C., "Standard for the transmission of IP datagrams over Ethernet networks", STD 41, RFC 894, April 1984.

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[RFC2516] Mamakos, L., Lidl, K., Evarts, J., Carrel, D., Simone, D., and R. Wheeler, "A Method for Transmitting PPP Over Ethernet (PPPoE)", RFC 2516, February 1999.

[RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.

[RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.

[RFC3927] Cheshire, S., Aboba, B., and E. Guttman, "Dynamic Configuration of IPv4 Link-Local Addresses", RFC 3927, May 2005.

[RFC4074] Morishita, Y. and T. Jinmei, "Common Misbehavior Against DNS Queries for IPv6 Addresses", RFC 4074, May 2005.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC5942] Singh, H., Beebee, W., and E. Nordmark, "IPv6 Subnet Model: The Relationship between Links and Subnet Prefixes", RFC 5942, July 2010.

[RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.

[RFC6106] Jeong, J., Park, S., Beloeil, L., and S. Madanapalli, "IPv6 Router Advertisement Options for DNS Configuration", RFC 6106, November 2010.

[RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, April 2011.

[RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.

[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.

[RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 2011.

[RFC6384] van Beijnum, I., "An FTP Application Layer Gateway (ALG) for IPv6-to-IPv4 Translation", RFC 6384, October 2011.

[RFC6586] Arkko, J. and A. Keranen, "Experiences from an IPv6-Only Network", RFC 6586, April 2012.

[RFC6603] Korhonen, J., Savolainen, T., Krishnan, S., and O. Troan, "Prefix Exclude Option for DHCPv6-based Prefix Delegation", RFC 6603, May 2012.

[YasudaAPRICOT2011] Yasuda, A., "Building for IPv6 by IPv6 Promotion Council Japan.", February, 2011, <http://meetings.apnic.net/__data/assets/pdf_file/0003/30981/Ayumu-Yasuda-apricot.pdf>.

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.

T. Jimei of Internet Systems Consortium for his quick hack on A filter of Bind 9.

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

Network Working Group
Internet-Draft
Obsoletes: 6204 (if approved)
Intended status: Informational
Expires: May 3, 2013

H. Singh
W. Beebee
Cisco Systems, Inc.
C. Donley
CableLabs
B. Stark
AT&T

October 30, 2012

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.

Table of Contents

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

End-User Network	one or more links attached to the IPv6 CE router that connect IPv6 hosts.
IPv6 Customer Edge Router	a node intended for home or small-office use that forwards IPv6 packets not explicitly addressed to itself. The IPv6 CE router connects the end-user network to a service provider network.
IPv6 Host	any device implementing an IPv6 stack receiving IPv6 connectivity through the IPv6 CE router.
LAN Interface	an IPv6 CE router's attachment to a link in the end-user network. Examples are Ethernet (simple or bridged), 802.11 wireless, or other LAN technologies. An IPv6 CE router may have one or more network-layer LAN interfaces.
Service Provider	an entity that provides access to the Internet. In this document, a service provider specifically offers Internet access using IPv6, and may also offer IPv4 Internet access. The service provider can provide such access over a variety of different transport methods such as DSL, cable, wireless, and others.
WAN Interface	an IPv6 CE router's attachment to a link used to provide connectivity to the service provider network; example link technologies include Ethernet (simple or bridged), PPP links, Frame Relay, or ATM networks, as well as Internet-layer (or higher-layer) "tunnels", such as tunnels over IPv4 or IPv6 itself.

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.

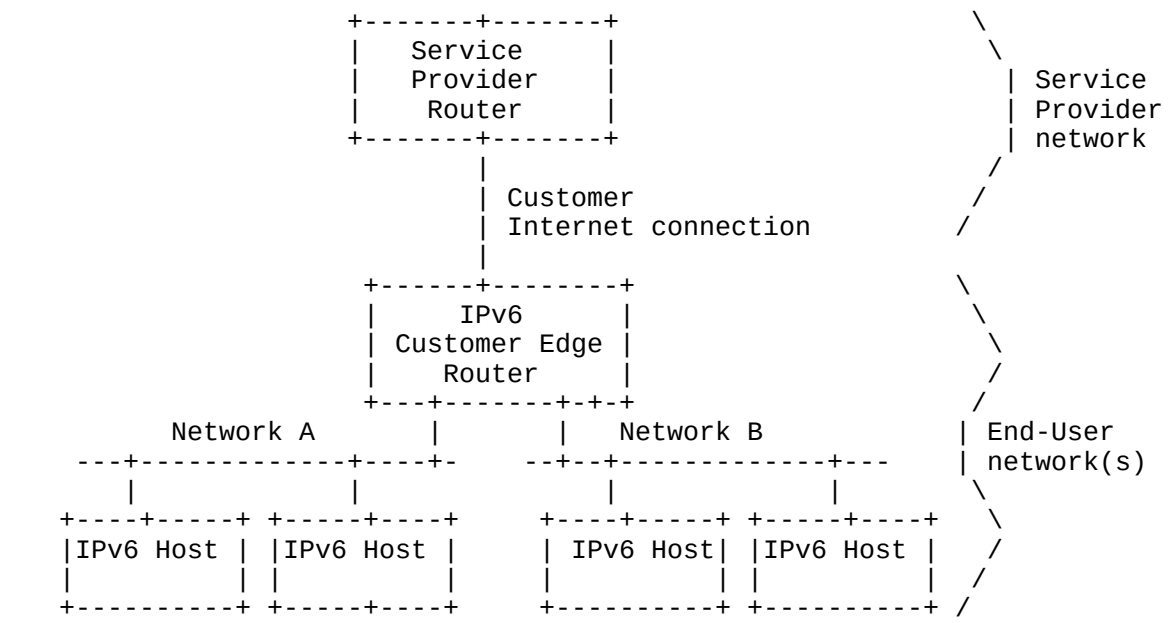


Figure 1: An Example of a Typical End-User Network

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 de-couples 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:

Mikael Abrahamsson, Tore Anderson, Merete Asak, Rajiv Asati, Scott Beuker, Mohamed Boucadair, Rex Bullinger, Brian Carpenter, Tassos Chatzithomaoglou, Lorenzo Colitti, Remi Denis-Courmont, Gert Doering, Alain Durand, Katsunori Fukuoka, Brian Haberman, Tony Hain, Thomas Herbst, Ray Hunter, Kevin Johns, Joel Jaeggli, Erik Kline, Stephen Kramer, Victor Kuarsingh, Francois-Xavier Le Bail, Arifumi Matsumoto, David Miles, Shin Miyakawa, Jean-Francois Mule, Michael Newbery, Carlos Pignataro, John Pomeroy, Antonio Querubin, Daniel Roesen, Hiroki Sato, Teemu Savolainen, Matt Schmitt, David Thaler, Mark Townsley, Sean Turner, Bernie Volz, Dan Wing, Timothy Winters, James Woodyatt, Carl Wuyts, and Cor Zwart.

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-dhcpv6-solmaxrt-update]
Droms, R., "Modification to Default Value of SOL_MAX_RT", draft-droms-dhc-dhcpv6-solmaxrt-update-03 (work in progress), August 2012.

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

"Prefix Exclude Option for DHCPv6-based Prefix Delegation", draft-ietf-dhc-pd-exclude-04 (work in progress), December 2011.

[I-D.ietf-pcp-base] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-28 (work in progress), October 2012.

[RFC1122] Braden, R., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.

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

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[RFC2464] Crawford, M., "Transmission of IPv6 Packets over Ethernet Networks", RFC 2464, December 1998.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.

[RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.

[RFC3646] Droms, R., "DNS Configuration options for Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3646, December 2003.

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.

[RFC3736] Droms, R., "Stateless Dynamic Host Configuration Protocol (DHCP) Service for IPv6", RFC 3736, April 2004.

[RFC4191] Draves, R. and D. Thaler, "Default Router Preferences and More-Specific Routes", RFC 4191, November 2005.

[RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

[RFC4242] Venaas, S., Chown, T., and B. Volz, "Information Refresh Time Option for Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 4242, November 2005.

[RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.

[RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.

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

[RFC4779] Asadullah, S., Ahmed, A., Popoviciu, C., Savola, P., and J. Palet, "ISP IPv6 Deployment Scenarios in Broadband Access Networks", RFC 4779, January 2007.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.

[RFC4864] Van de Velde, G., Hain, T., Droms, R., Carpenter, B., and E. Klein, "Local Network Protection for IPv6", RFC 4864, May 2007.

[RFC5072] S.Varada, Haskins, D., and E. Allen, "IP Version 6 over PPP", RFC 5072, September 2007.

[RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

[RFC5908] Gayraud, R. and B. Lourdelet, "Network Time Protocol (NTP) Server Option for DHCPv6", RFC 5908, June 2010.

[RFC5942] Singh, H., Beebee, W., and E. Nordmark, "IPv6 Subnet Model: The Relationship between Links and Subnet Prefixes", RFC 5942, July 2010.

[RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification",

RFC 5969, August 2010.

[RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service", RFC 6092, January 2011.

[RFC6106] Jeong, J., Park, S., Beloeil, L., and S. Madanapalli, "IPv6 Router Advertisement Options for DNS Configuration", RFC 6106, November 2010.

[RFC6177] Narten, T., Huston, G., and L. Roberts, "IPv6 Address Assignment to End Sites", BCP 157, RFC 6177, March 2011.

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

[RFC6334] Hankins, D. and T. Mrugalski, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Option for Dual-Stack Lite", RFC 6334, August 2011.

[RFC6434] Jankiewicz, E., Loughney, J., and T. Narten, "IPv6 Node Requirements", RFC 6434, December 2011.

8.2. Informative References

[I-D.ietf-dhc-dhcpv6-stateful-issues]
Troan, O. and B. Volz, "Issues with multiple stateful DHCPv6 options", draft-ietf-dhc-dhcpv6-stateful-issues-00 (work in progress), May 2012.

[MULTIHOMING-WITHOUT-NAT]
Troan, O., Ed., Miles, D., Matsushima, S., Okimoto, T., and D. Wing, "IPv6 Multihoming without Network Address Translation", Work in Progress, December 2010.

[RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, March 2011.

[UPnP-IGD]
UPnP Forum, "Universal Plug and Play (UPnP) Internet Gateway Device (IGD)", November 2001, <http://www.upnp.org/>.

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

Network Working Group
Internet-Draft
Updates: 6145 (if approved)
Intended status: Standards Track
Expires: April 15, 2013

X. Li
C. Bao
CERNET Center/Tsinghua University
D. Wing
R. Vaithianathan
Cisco
G. Huston
APNIC

October 12, 2012

Stateless Source Address Mapping for ICMPv6 Packets
draft-ietf-v6ops-ivi-icmp-address-07

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.

Status of this Memo

This Internet-Draft is submitted to IETF 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 15, 2013.

Table of Contents

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 Levkowetz, Henrik Levkowetz, Randy Bush and Warren Kumari for their comments and suggestions.

9. References

9.1. Normative References

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

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

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.

[RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.

[RFC5837] Atlas, A., Bonica, R., Pignataro, C., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, April 2010.

[RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X.

Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.

[RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.

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

v6ops
Internet-Draft
Intended status: Informational
Expires: September 4, 2012

I. Gashinsky
Yahoo!
J. Jaeggli
Zynga
W. Kumari
Google Inc

March 03, 2012

Operational Neighbor Discovery Problems
draft-ietf-v6ops-v6nd-problems-05

Abstract

In IPv4, subnets are generally small, made just large enough to cover the actual number of machines on the subnet. In contrast, the default IPv6 subnet size is a /64, a number so large it covers trillions of addresses, the overwhelming number of which will be unassigned. Consequently, simplistic implementations of Neighbor Discovery (ND) can be vulnerable to deliberate or accidental denial of service, whereby they attempt to perform address resolution for large numbers of unassigned addresses. Such denial of attacks can be launched intentionally (by an attacker), or result from legitimate operational tools or accident conditions. As a result of these vulnerabilities, new devices may not be able to "join" a network, it may be impossible to establish new IPv6 flows, and existing IPv6 transported flows may be interrupted.

This document describes the potential for DOS in detail and suggests possible implementation improvements as well as operational mitigation techniques that can in some cases be used to protect against or at least alleviate the impact of such attacks.

This Internet-Draft will expire on September 4, 2012.

Table of Contents

1. Introduction

This document describes implementation issues with IPv6's Neighbor Discovery protocol that can result in vulnerabilities when a network is scanned, either by an intruder or through the use of scanning tools that perform network inventory, security audits, etc. (e.g. "nmap").

This document describes the problem in detail, suggests possible implementation improvements, as well as operational mitigation techniques, that can in some cases protect against such attacks.

The RFC series documents generally describe the behavior of protocols, that is, "what" is to be done by a protocol, but not exactly "how" it is to be implemented. The exact details of how best to implement a protocol will depend on the overall hardware and software architecture of a particular device. The actual "how" decisions are (correctly) left in the hands of implementers, so long as implementations differences will generally produce proper on-the-wire behavior.

While reading this document, it is important to keep in mind that discussions of how things have been implemented beyond basic compliance with the specification is not within the scope of the neighbor discovery RFCs.

1.1. Applicability

This document is primarily intended for operators of IPV6 networks and implementors of [RFC4861]. The Document provides some operational considerations as well as recommendations to increase the resilience of the Neighbor Discovery protocol.

2. The Problem

In IPv4, subnets are generally small, made just large enough to cover the actual number of machines on the subnet. For example, an IPv4 /20 contains only 4096 address. In contrast, the default IPv6 subnet size is a /64, a number so large it covers literally billions of billions of addresses, the overwhelming majority of which will be unassigned. Consequently, simplistic implementations of Neighbor Discovery may fail to perform as desired when they perform address resolution of large numbers of unassigned addresses. Such failures can be triggered either intentionally by an attacker launching a Denial of Service attack (DoS)[RFC4732] to exploit this vulnerability, or unintentionally due to the use of legitimate operational tools that scan networks for inventory and other

purposes. As a result of these failures, new devices may not be able to "join" a network, it may be impossible to establish new IPv6 flows, and existing IPv6 transport flows may be interrupted.

Network scans attempt to find and probe devices on a network. Typically, scans are performed on a range of target addresses, or all the addresses on a particular subnet. When such probes are directed via a router, and the target addresses are on a directly attached network, the router will attempt to perform address resolution on a large number of destinations (i.e., some fraction of the 2^64 addresses on the subnet). The router's process of testing for the (non)existence of neighbors can induce a denial of service condition, where the number of necessary Neighbor Discovery requests overwhelms the implementation's capacity to process them, exhausts available memoryand replaces existing in-use mappings with incomplete entries that will never be completed. A directed DoS attack may seek to intentionally create similar conditions to that created unintentionally by a network scan. The resulting network disruption may impact existing traffic, and devices that join the network may find that address resolution attempts fail. The DOS as a consequence of network scanning was previously described in [RFC5157]

In order to mitigate risk associated with this DoS threat, some router implementations have taken steps to rate-limit the processing rate of Neighbor Solicitations (NS). While these mitigations do help, they do not fully address the issue and may introduce their own set of issues to the neighbor discovery process.

3. Terminology

Address Resolution Address resolution is the process through which a node determines the link-layer address of a neighbor given only its IP address. In IPv6, address resolution is performed as part of Neighbor Discovery [RFC4861], p60

Forwarding Plane That part of a router responsible for forwarding packets. In higher-end routers, the forwarding plane is typically implemented in specialized hardware optimized for performance. Steps in the forwarding process include determining the correct outgoing interface for a packet, decrementing its Time To Live (TTL), verifying and updating the checksum, placing the correct link-layer header on the packet, and forwarding it.

Control Plane That part of the router implementation that maintains the data structures that determine where packets should be forwarded. The control plane is typically implemented as a "slower" software process running on a general purpose processor

and is responsible for such functions as communicating network status changes via routing protocols, maintaining the forwarding table, performing management, and resolving the correct link-layer address for adjacent neighbors. The control plane "controls" the forwarding plane by programming it with the information needed for packet forwarding.

Neighbor Cache As described in [RFC4861], the data structure that holds the cache of (amongst other things) IP address to link-layer address mappings for connected nodes. As the information in the Neighbor Cache is needed by the forwarding plane every time it forwards a packet, it is usually implemented in an ASIC.

Neighbor Discovery Process The Neighbor Discovery Process (NDP) is that part of the control plane that implements the Neighbor Discovery protocol. NDP is responsible for performing address resolution and maintaining the Neighbor Cache. When forwarding packets, the forwarding plane accesses entries within the Neighbor Cache. When the forwarding plane processes a packet for which the corresponding Neighbor Cache Entry is missing or incomplete, it notifies NDP to take appropriate action (typically via a shared queue). NDP picks up requests from the shared queue and performs any necessary discovery action. In many implementations the NDP is also responsible for responding to router solicitation messages, Neighbor Unreachability Detection (NUD), etc.

4. Background

Modern router architectures separate the forwarding of packets (forwarding plane) from the decisions needed to decide where the packets should go (control plane). In order to deal with the high number of packets per second, the forwarding plane is generally implemented in hardware and is highly optimized for the task of forwarding packets. In contrast, the NDP control plane is mostly implemented in software processes running on a general purpose processor.

When a router needs to forward an IP packet, the forwarding plane logic performs the longest match lookup to determine where to send the packet and what outgoing interface to use. To deliver the packet to an adjacent node, the forwarding plane encapsulates the packet in a link-layer frame (which contains a header with the link-layer destination address). The forwarding plane logic checks the Neighbor Cache to see if it already has a suitable link-layer destination, and if not, places the request for the required information into a queue, and signals the control plane (i.e., NDP) that it needs the link-layer address resolved.

In order to protect NDP specifically and the control plane generally from being overwhelmed with these requests, appropriate steps must be taken. For example, the size and fill rate of the queue might be limited. NDP running in the control plane of the router dequeues requests and performs the address resolution function (by performing a neighbor solicitation and listening for a neighbor advertisement). This process is usually also responsible for other activities needed to maintain link-layer information, such as Neighbor Unreachability Detection (NUD).

By sending appropriate packets to addresses on a given subnet, an attacker can cause the router to queue attempts to resolve so many addresses that it crowds out attempts to resolve "legitimate" addresses (and in many cases becomes unable to perform maintenance of existing entries in the neighbor cache, and unable to answer Neighbor Solicitation). This condition can result in the inability to resolve new neighbors and loss of reachability to neighbors with existing ND-Cache entries. During testing it was concluded that 4 simultaneous nmap sessions from a low-end computer was sufficient to make a router's neighbor discovery process unusable and therefore forwarding became unavailable to the destination subnets.

The failure to maintain proper NDP behavior whilst under attack has been observed across multiple platforms and implementations, including the largest modern router platforms available (at the inception of work on this document).

5. Neighbor Discovery Overview

When a packet arrives at (or is generated by) a router for a destination on an attached link, the router needs to determine the correct link-layer address to use in the destination field of the layer 2 encapsulation. The router checks the Neighbor Cache for an existing Neighbor Cache Entry for the neighbor, and if none exists, invokes the address resolution portions of the IPv6 Neighbor Discovery [RFC4861] protocol to determine the link-layer address of the neighbor.

[RFC4861] Section 5.2 (Conceptual Sending Algorithm) outlines how this process works. A very high level summary is that the device creates a new Neighbor Cache Entry for the neighbor, sets the state to INCOMPLETE, queues the packet and initiates the actual address resolution process. The device then sends out one or more Neighbor Solicitations, and when it receives a corresponding Neighbor Advertisement, completes the Neighbor Cache Entry and sends the queued packet.

6. Operational Mitigation Options

This section provides some feasible mitigation options that can be employed today by network operators in order to protect network availability while vendors implement more effective protection measures. It can be stated that some of these options are "kludges", and can be operationally difficult to manage. They are presented, as they represent options we currently have. It is each operator's responsibility to evaluate and understand the impact of changes to their network due to these measures.

6.1. Filtering of unused address space.

The DoS condition is induced by making a router try to resolve addresses on the subnet at a high rate. By carefully addressing machines into a small portion of a subnet (such as the lowest numbered addresses), it is possible to filter access to addresses not in that assigned portion of address space using Access Control Lists (ACLs), or by null routing, features which are available on most existing platforms. This will prevent the attacker from making the router attempt to resolve unused addresses. For example if there are only 50 hosts connected to an interface, you may be able to filter any address above the first 64 addresses of that subnet by null-routing the subnet carrying a more specific /122 route or by applying ACLs on the WAN link to prevent the attack traffic reaching the vulnerable device.

As mentioned at the beginning of this section, it is fully understood that this is ugly (and difficult to manage); but failing other options, it may be a useful technique especially when responding to an attack.

This solution requires that the hosts be statically or statefully addressed (as is often done in a datacenter) and may not interact well with networks using [RFC4862]

6.2. Minimal Subnet Sizing.

By sizing subnets to reflect the number of addresses actually in use, the problem can be avoided. For example, [RFC6164] recommends sizing the subnets for inter-router links to only have 2 addresses (a /127). It is worth noting that this practice is common in IPv4 networks, in part to protect against the harmful effects of ARP request flooding.

Subnet prefixes longer than a /64 are not able to use stateless auto-configuration [RFC4862] so this approach is not suitable for use with hosts that are not statically configured.

6.3. Routing Mitigation.

One very effective technique is to route the subnet to a discard interface (most modern router platforms can discard traffic in hardware / the forwarding plane) and then have individual hosts announce routes for their IP addresses into the network (or use some method to inject much more specific addresses into the local routing domain). For example the network 2001:db8:1:2:3::/64 could be routed to a discard interface on "border" routers, and then individual hosts could announce 2001:db8:1:2:3::10/128, 2001:db8:1:2:3::66/128 into the IGP. This is typically done by having the IP address bound to a virtual interface on the host (for example the loopback interface), enabling IP forwarding on the host and having it run a routing daemon. For obvious reasons, host participation in the IGP makes many operators uncomfortable, but can be a very powerful technique if used in a disciplined and controlled manner. One method to help address these concerns is to have the hosts participate in a different IGP (or difference instance of the same IGP) and carefully redistribute into the main IGP.

6.4. Tuning of the NDP Queue Rate Limit.

Many implementations provide a means to control the rate of resolution of unknown addresses. By tuning this rate, it may be possible to ameliorate the issue, as with most tuning knobs (especially those that deal with rate limiting), the attack may be completed more quickly due to the lower threshold. By excessively lowering this rate you may negatively impact how long the device takes to learn new addresses under normal conditions (for example, after clearing the neighbor cache or when the router first boots). Under attack conditions you may be unable to resolve "legitimate" addresses sooner than if you had just left the parameter untouched.

It is worth noting that this technique is worth investigating only if the device has separate queues for resolution of unknown addresses and the maintenance of existing entries.

7. Recommendations for Implementors.

The section provides some recommendations to implementors of IPv6 Neighbor Discovery.

At a high-level, implementors should program defensively. That is, they should assume that attackers will attempt to exploit implementation weaknesses, and should ensure that implementations are robust to various attacks. In the case of Neighbor Discovery, the following general considerations apply:

Manage Resources Explicitly Resources such as processor cycles, memory, etc. are never infinite, yet with IPv6's large subnets it is easy to cause NDP to generate large numbers of address resolution requests for non-existent destinations. Implementations need to limit resources devoted to processing Neighbor Discovery requests in a thoughtful manner.

Prioritize Some NDP requests are more important than others. For example, when resources are limited, responding to Neighbor Solicitations for one's own address is more important than initiating address resolution requests that create new entries. Likewise, performing Neighbor Unreachability Detection, which by definition is only invoked on destinations that are actively being used, is more important than creating new entries for possibly non-existent neighbors.

7.1. Prioritize NDP Activities

Not all Neighbor Discovery activities are equally important. Specifically, requests to perform large numbers of address resolutions on non-existent Neighbor Cache Entries should not come at the expense of servicing requests related to keeping existing, in-use entries properly up-to-date. Thus, implementations should divide work activities into categories having different priorities. The following gives examples of different activities and their importance in rough priority order. If implmented, the operation and priority of these should be configurable by the operator.

1. It is critical to respond to Neighbor Solicitations for one's own address, especially for a router. Whether for address resolution or Neighbor Unreachability Detection, failure to respond to Neighbor Solicitations results in immediate problems. Failure to respond to NS requests that are part of NUD can cause neighbors to delete the NCE for that address, and will result in followup NS messages using multicast. Once an entry has been flushed, existing traffic for destinations using that entry can no longer be forwarded until address resolution completes successfully. In other words, not responding to NS messages further increases the NDP load, and causes on-going communication to fail.

2. It is critical to revalidate one's own existing NCEs in need of refresh. As part of NUD, ND is required to frequently revalidate existing, in-use entries. Failure to do so can result in the entry being discarded. For in-use entries, discarding the entry will almost certainly result in a subsequent request to perform address resolution on the entry, but this time using multicast. As above, once the entry has been flushed, existing traffic for destinations using that entry can no longer be forwarded until address resolution

completes successfully.

3. To maintain the stability of the control plane, Neighbor Discovery activity related to traffic sourced by the router (as opposed to traffic being forwarded by the router) should be given high priority. Whenever network problems occur, debugging and making other operational changes requires being able to query and access the router. In addition, routing protocols dependent on Neighbor Discovery for connectivity may begin to react (negatively) to perceived connectivity problems, causing additional undesirable ripple effects.

4. Traffic to unknown addresses should be given lowest priority. Indeed, it may be useful to distinguish between "never seen" addresses and those that have been seen before, but that do not have a corresponding NCE. Specifically, the conceptual processing algorithm in IPv6 Neighbor Discovery [RFC4861] calls for deleting NCEs under certain conditions. Rather than delete them completely, however, it might be useful to at least keep track of the fact that an entry at one time existed, in order to prioritize address resolution requests for such neighbors compared with neighbors that have never been seen before.

7.2. Queue Tuning.

On implementations in which requests to NDP are submitted via a single queue, router vendors should provide operators with means to control both the rate of link-layer address resolution requests placed into the queue and the size of the queue. This will allow operators to tune Neighbour Discovery for their specific environment. The ability to set, or have per interface or per prefix queue limits at a rate below that of the global queue limit might limit the damage to the neighbor discovery processing to the network targeted by the attack.

Setting those values must be a very careful balancing act - the lower the rate of entry into the queue, the less load there will be on the ND process, however, it will take the router longer to learn legitimate destinations as a result. In a datacenter with 6,000 hosts attached to a single router, setting that value to be under 1000 would mean that resolving all of the addresses from an initial state (or something that invalidates the address cache, such as a STP TCN) may take over 6 seconds. Similarly, the lower the size of the queue, the higher the likelihood of an attack being able to knock out legitimate traffic (but less memory utilization on the router).

8. IANA Considerations

No IANA resources or consideration are requested in this draft.

9. Security Considerations

This document outlines mitigation options that operators can use to protect themselves from Denial of Service attacks. Implementation advice to router vendors aimed at ameliorating known problems carries the risk of previously unforeseen consequences. It is not believed that these mitigation techniques or the implementation of finer-grained queuing of NDP activity create additional security risks or DOS exposure.

10. Acknowledgements

The authors would like to thank Ron Bonica, Troy Bonin, John Jason Brzozowski, Randy Bush, Vint Cerf, Tassos Chatzithomaoglou, Jason Fesler, Wes George, Erik Kline, Jared Mauch, Chris Morrow and Suran De Silva. Special thanks to Thomas Narten and Ray Hunter for detailed review and (even more so) for providing text!

Apologies for anyone we may have missed; it was not intentional.

11. References

11.1. Normative References

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

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.

[RFC6164] Kohno, M., Nitzan, B., Bush, R., Matsuzaki, Y., Colitti, L., and T. Narten, "Using 127-Bit IPv6 Prefixes on Inter-Router Links", RFC 6164, April 2011.

11.2. Informative References

[RFC4732] Handley, M., Rescorla, E., and IAB, "Internet Denial-of-Service Considerations", RFC 4732, December 2006.

[RFC5157] Chown, T., "IPv6 Implications for Network Scanning", RFC 5157, March 2008.

Authors' Addresses

Igor Gashinsky
Yahoo!
45 W 18th St
New York, NY
USA

Email: igor@yahoo-inc.com

Joel Jaeggli
Zynga
111 Evelyn
Sunnyvale, CA
USA

Email: jjaeggli@zynga.com

Warren Kumari
Google Inc
1600 Amphitheatre Parkway
Mountain View, CA
USA

Email: warren@kumari.net

v6ops
Internet-Draft
Intended status: Informational
Expires: April 11, 2012

V. Kuarsingh, Ed.
Rogers Communications

October 9, 2011

Wireline Incremental IPv6
draft-kuarsingh-wireline-incremental-ipv6-02

Abstract

Operators worldwide are in various stages of preparing for, or deploying IPv6 into their networks. The operators often face challenges related to both IPv6 introduction along with a growing risk of IPv4 run out within their organizations. The overall problem for many of there operators will be to meet the simultaneous needs of IPv6 connectivity and continue support for IPv4 connectivity for legacy devices and systems with a depleting supply of IPv4 addresses. The overall transition will take most networks form an IPv4-Only environment to a dual stack network environment and potentially an IPv6-Only operating mode. This document helps provide a framework for Wireline providers who may be faced with many of these challenges as they consider what IPv6 transition technologies to use, how to use the selected technologies and when to use them.

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 11, 2012.

Copyright Notice

Table of Contents

1. Introduction

IPv6 represents the strategic IP protocol version which will meet the addressing needs of the Internet into the future. Many operators are already working on implementing IPv6 within their networks, and other operators may just be starting this process. A solid IPv6 plan will need to include both the baseline requirements to enable IPv6 within the network, but must also include facilities to provide continued support for IPv4 connectivity. Given the vast number of technological options now available to operators for transition to IPv6, the task may seem daunting when attempting to identify which technologies are appropriate for a given network, and how these technologies can be introduced.

This draft sets out to help operators who may be just starting the evaluation process or well underway, by identifying which technologies can be used in an incremental fashion to transition from an IPv4-only environment to an efficient IPv6/IPv4 dual stack environment. Some plans may also include IPv6-Only end state targets, but there is not clear consensus on how long IPv4 support is required. Although no single plan will work for for all operators, generically, options listed herein provide a baseline which can be included in many plans.

This draft is specifically catered towards wireline environments which may use technologies such as Cable, DSL and/or Fibre as the access method to the end consumer. This draft also attempts to follow the methodologies set out in [I-D.ietf-v6ops-v4v6tran-framework] to identify how the technologies can be used individual and in combination. This document also attempts to follow the principles laid out in [RFC6180] which provides guidance on using IPv6 transition mechanisms. This document does not show the IPv6-Only end state architecture since it is years away from existing mainstream Internet service connections. This document will show how tunnelling using 6RD [RFC5969] and DS-Lite [RFC6333] as well as translation via CGN can be used with Native Dual Stack to deliver effective IPv4 and IPv6 services in an evolving wireline network.

2. Motivation

Wireline Operators are increasingly becoming aware of the need to support IPv6. The depletion of unassigned IPv4 addresses within IANA and the RIRs has highlighted the need to move beyond IPv4-Only operation. In many operator environments, the main task will be the addition of IPv6 into the network. As straightforward as this task may seem, it will require forethought and planning. However, of greater concern is that the introduction of IPv6 may need to take

place in a volatile environment where IPv4 resources are depleted complicating what technologies can be used, and how Dual Stack services may be offered to customers.

Operators will want to understand which of the prevailing technologies can be used in a changing network environment while adapting to the needs and conditions of the network. IPv6 will be a focal point in the Operators plans, but the realities of IPv4, and it's demand by legacy equipment and system needs to be acknowledged and managed. The Operator's main goal will be to maintain quality IP services to Internet customers while the world moves from a predominately IPv4 centric system to a Dual Dtack IPv6/IPv4 system and eventually to an IPv6 centric world. The IPv6 centric world may not preclude the use of IPv4 altogether, but focuses on a time where most functions and and will be delivered over IPv6.

3. Operator Assumptions

For the purposes of this document, it's assumed the operator is considering deploying IPv6. It is also assumed that the operator has a legacy IPv4 customer base which will continue to exist and for a long period of time (years). Other assumptions include that that operator will want to minimize the level of disruption to the existing and new customers by minimizing number of technologies and functions that are needed to mediate any given set of customer flows (overall preference for Native IP flows).

These assumptions translate into analyzing technologies and subsequently selecting technologies which minimize how many flows must be tunnelled, translated or intercepted at any given time. Technology selections would be made to manage the non dominant flows and allow Native IP routing (IPv4 and/or IPv6) to manage the bulk of the traffic. This allows the operator to minimize the cost of IPv6 transition technologies by containing the scale required by the relevant systems.

Not all operators may see these assumptions as valid, but most operators who have built and optimized their networks for efficient delivery of IP traffic from their customer base to the Internet (and vice versa) would typically agree with the approch suggested herein.

4. Reasons and Considerations for a Phased Approach

When faced with the challenges described in the Introductory portion of this document, operators may need to consider a phased approach to IPv6 service introduction and IPv4 service continuance. Both IPv4

and IPv6 play critical role in connectivity throughout the IPv6 transition yet each protocol will be based with challenges as time progresses. Some of these challenges include the depletion of IPv4 which will occur in many networks long before most traffic is able to delivered over IPv6. IPv6 will also be added into many networks and pose many operational challenges to organizations and customers since much of the hardware, software and processes will be relatively new. Connectivity modes will move from single stack to dual stack in the home further challenging the transition as operators contend with many functional behaviours in the home network.

These challenges, as noted, will occur over time which means the operator's plans need to address the every changing requirements of the network and customer demand. The following few sections highlight some of the key reasons why a phase approach to IPv6 transition may be warranted and desired.

4.1. Relevance of IPv6 and IPv4

The reality for operators over the next few years will be that both IPv4 and IPv6 will play a role in the Internet experience. Although many IPv6 advocates seek to move the Internet to IPv6 quickly, the fact that many older operating systems and hardware support IPv4-Only operating modes will need to be accepted and managed. Internet customers don't buy IPv4 or IPv6 connections, they buy Internet connections, which demands the need to support both IPv4 and IPv6 for as long at the customer's home network demands such support.

The Internet is made of of many interconnecting systems, networks, hardware, software and content sources - all of which will move to IPv6 at different rates. The Operator's mandate during this time of transition will be to support connectivity to both IPv6 and IPv4 through various technological means. The operator may be able to leverage one or the other protocol to help bridge connectivity, but the home network will demand both IPv4 and IPv6 for the foreseeable future.

4.2. IPv4 Resource Challenges

Since connectivity to IPv4-Only endpoints and/or content will remain prevalent for a long period of time, IPv4 resource challenges are of key concern to operators. The lack of new IPv4 addressees for additional endpoints means that growth in demand of IPv4 connections in some networks will be based on address sharing.

Networks are growing at different rates based on a number of factors which may be related to emerging markets and/or proliferation of Internet based services and endpoints. Given that reality, growth on

the Internet will continue. IPv4 address constraints will likely impact many if not most operators at some point. This will play an important role when considering what technologies are viable as the transition period moves on. Of note will be any use of technologies which rely on IPv4 as the mechanism to supply IPv6 services such as 6RD. Also, if Native Dual Stack is considered by the operator, challenges on the IPv4 path is also of concern.

Some operators may be able to achieve some level of IPv4 address reclamation through various levels of efficiency in the network and replacement of GUA assignments with private addresses such as those in [RFC1918], but these measures are tactical in nature and do not support a longer term strategic option. The lack of new IPv4 addresses 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.

4.3. IPv6 Introduction and Maturity

Operators will want to or be forced to support IPv6 at some point. The introduction of IPv6 will require the operationalization of IPv6. The IPv4 environment we have today was built over many years and was matured by experience. Although many of these experiences are transferable from IPv4 to IPv6, new experience specific to IPv6 will be needed.

Engineering and Operational staff will need to become acclimatized to IPv6 which and gain this needed experience. During this ramp up period, Operators will need to be aware that instability may occur in the IPv6 deployment and should be taking this into account when selecting what technologies are viable during early 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. This plays a role during initial transition when considering technologies which require IPv6 to support IPv4 services such as DS-Lite.

Of consideration as well will be the reality that some of these technologies are new and require refinement within running code and operations. Deployment experience may be needed to vet these technologies out and stabilize them 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, along with other systems.

Although the base technological capabilities exist to enable and run IPv6 in most environments; until such time as each key technical

member of an operator's organization can identify IPv6, understand it's relevance to the IP Service offering, how it operates and how to troubleshoot it - it's still maturing.

4.4. Service Management

Services are managed within most networks and is often based on the gleaning and monitoring of IPv4 addresses. 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 128-bits, but often both IPv4 and IPv6 at the same time.

With any Dual Stack service - whether Native, 6RD based, DS-Lite based or otherwise - two address families need to be managed simultaneously to help provide for the full Internet experience. 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 which require IPv6 as the base protocol to delivery IPv4. Instability on the IPv6 service in such case would impact IPv4 services.

4.5. Sub-Optimal Operation of Transition Technologies

Yet another important concept for an operator to understand is the difference between a native path and a path which requires a transition technology to bridge certain connectivity. Native paths are often well understood and most networks are optimized to send traffic to and from the customer (to/from Internet) in an efficient manner.

The addition of transition technologies may alter the normal path of traffic and delay or hinder the IP flows due to tunnelling and translation operation. New logical nodes in the network will be needed to supply the full IP path, all of which will be slower and less agile then the native alternative.

The consideration for this issue may be that an operator minimize the amount of traffic that needs to be delivered over a transition technology platform by optimizing the technologies deployed over time. During earlier phases of transition, IPv6 traffic volumes may be lower, so tunnelling of IPv6 traffic may be reasonable. Over time, these traffic volumes will increase, raising the benefits of native delivery of this traffic. Also, as IPv4 content diminishes, translation and tunnelling of this protocol may become more tolerable

when considering performance.

Operators may wish to align their own internal service delivery with the deployment of transition technologies including Native IPv6 and potential CGN deployments. An operator may not want to enable many of their services, especially high traffic flow services, for IPv6 delivery if IPv6 tunnelling is used. The operator may which to constrain such customers to IPv4 delivery until Native IPv6 is available. Also, the operation may likewise which to constrain customers to IPv6 content versus IPv4 if CGN is deployed in the future to deal with IPv4 address depletion.

5. IPv6 Transition Technology Analysis

Understanding the main IPv6 transition technologies and those related to dealing with IPv4 run out should be a primary goal of any operator. Although this draft is not designed to list all options or to provide a full technical analysis of each of the identified technologies, it provides a brief description and explains some of the mainstream technological options can be used in an operator network.

In this analysis, common automatic tunnelling, provider controlled tunnelling, translation and native modes of operations are considered. The analysis also includes technologies such as NAT64 which may not be appropriate for near term wireline transition due to the nature of the home network. This analysis is also focused primarily on the applicability of technologies to deliver residential services and less focused on commercial or support for the provider's infrastructure. It is assume the operator is able to Dual Stack their own core network and transition their own services to support IPv6.

5.1. Automatic Tunnelling using 6to4 and Teredo

Operators may not be actively deploying IPv6, but automatic mechanisms do exist on deployed operating systems and hardware that should be of note. Such technologies include 6to4 described within [RFC3056] which is mostly commonly used in a deployment mode using anycast relays as described in [RFC3068]. Additionally, Teredo [RFC4380] is also used widely by many Internet hosts as a means to reach the IPv6 world when no native or operator provided path is made present.

The operator may not want or have intended for these technologies to be active in their networks, but should be aware that the traffic exists. The operator may be inclined to provide the best possible

experience for endpoints using automatic tunnelling technologies. Documents such as [RFC6343] have been written to help operators understand observed problems and provide guidelines on how to manage such protocols. An Operator may want to incrementally 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 address are used for CGN 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 can is likely to occur if squat space (not assigned to local operator) is used in place of RFC1918 space or if Shared CGN Space is used [I-D.weil-shared-transition-space-request]. The operator can used options such as 6to4-PMT to help mitigate this issue as described in [I-D.kuarsingh-v6ops-6to4-provider-managed-tunnel] or attempt to block 6to4 operation entirely.

5.2. Carrier Grade NAT (NAT444)

Carrier Grade NAT (GGN), specifically as deployed in a NAT444 scenario [I-D.ietf-behave-lsn-requirements], is also a relevant technology. Although CGN is not a IPv6 specific function, it may prove beneficial for those operators who offer Dual Stack services to customer endpoints once they exhaust their pools of IPv4 addresses. CGNs, and address sharing overall, are known to cause certain challenges for the IPv4 service path as described in documents like [RFC6269], but will often be necessary for a time.

In a network where IPv4 address availability is low or no new addressees can be assigned to Internet hosts, a CGN deployment may be a viable way to provide continued access to the IPv4 path. Other technologies may also be used, but a provider may choose to use this method earlier on since it's a well understood method of delivering IPv4 connectivity - notwithstanding the challenges of CGN and address sharing. Some of the advantages of using CGN include the similarities in provisioning and activation IPv4 hosts within a network and operational procedures in managing such hosts or CPEs (i.e. DHCPv6, DNSv4, TFTP, TR-069 etc).

When considered in the overall IPv6 transition, CGN may play a vital role in the delivery of Internet services.

5.3. 6RD

6RD as described in [RFC5969] does provide a quick and effective way to deliver IPv6 services to access network endpoints which do not yet support Native IPv6 on the operator's access network (WAN Side connection). 6RD provides tunnelled connectivity to IPv6 over the existing IPv4 path. The lack of Native IPv6 support at customer premise may be related to technological challenges of delivering IPv6 on a given access type or related to other operational or technical impediments that may existing in the operator's network.

6RD defiantly offers a solid early transition option to operators by eliminating the bottle neck of needing to deploy Native IPv6 to the access edge and customer CPE. Over time, as the access edge is upgraded and customer premise equipment is replaced, 6RD can be superseded by Native IPv6 access. 6RD can be delivered along with CGN, but this mode of operation would be a sub-optimal way of delivering service since the operator would then need to relay all IPv6 traffic as well as provide NAT functionally for all Internet bound IPv4 flows.

6RD may also be seen as advantageous during 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. Scaling of 6RD may be required by adding relays to the operator's network, but since 6RD is stateless, this task is quite manageable. In the case where CGN is used, there are stateful considerations to be made on the NATed IPv4 path.

Operators may want to use 6RD, as noted, while traffic volumes are low and while internal services are mainly on IPv4. As higher capacities are reached on the IPv6 path, the operator may want to move away from delivering heavy loads on a tunnelled connection. 6RD can continue to run indefinitely if the operator wishes to continue this service, but over time, Native IPv6 would be a much more efficient way of delivering robust IPv6 services.

Of specific consideration for 6RD is the client support required needed at the CPE. Most currently deployed CPEs do not have 6RD client functionality built into them and may or may not be upgradable. 6RD deployments would most likely require the replacement of the home CPE. An advantage of this technology over DS-Lite is that the WAN side interface does not need to implement IPv6 to function correctly which may make it easier to deploy to field hardware which is restricted in memory footprint, processing power and storage space. 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.

5.4. 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 which is already used in many existing IPv6 deployments. Native Dual Stack does however require that Native IPv6 be delivered to the customer premise. This technology option is desirable in many cases and can be used immediately if the access network and customer premise equipment supports Native IPv6 to the operators access network.

As time progresses, continued delivery new Native Dual Stack service connections may be challenging should the operator run out of free IPv4 addresses to assign to CPEs. For a sub-set of the IPv6 Native Dual Stack Customers, operators may include NATed IPv4 path as an assist, leveraging CGN. Delivering Native Dual Stack would require the operator’s core and access network support IPv6. Additionally, other systems like DHCPv6, DNS, and diagnostic/management facilities need to be upgraded to support IPv6. The upgrade of such systems may often not be trivial.

5.5. DS-Lite

DS-Lite, as described in [RFC6333], is an architecturally desirable way of delivery both IPv4 and IPv6 services in an IPv4 constrained environment. DS-Lite is able to provide IPv4 services to customer networks which are only addressed with IPv6. DS-Lite uses tunnelling mechanisms to pass IPv4 traffic between the customer’s network device (often a CPE) and the IPv4 internet using a provider managed AFTR.

DS-Lite however can only be used where there are native IPv6 facilities to the customer premise endpoint. This may mean that the technology’s use may not be viable during early transition. The operator may also not want to use DS-Lite immediately after IPv6 introduction as the organization may be development and maturing their IPv6 environment and may not want to subject the customers IPv4 connection to the IPv6 path. This is likely an early transition consideration and would diminish over time as IPv6 service delivery is matured. The provider may also want to make sure that most of their internal services, and external provider content is available over IPv6 before deploying DS-Lite. This would lower the overall load on the AFTR devices helping reduce cost and load on that layer of the network. Nothing precludes an operator from using DS-Lite earlier in the transition, but the operator needs to be aware of the challenges that can arise. If DS-Lite is used during early transition the operator will face scenario where they have support

personnel learning to troubleshoot IPv6 while this new protocol is supporting the legacy IPv4 service.

One of the strongest benefits of DS-Lite is the technology's ability to facilitate continued growth of IPv4 services if required without the need to deploy more IPv4 addressees to customer endpoints. This is quite advantageous as the transition period progresses and IPv4 resources become more and more challenging to secure.

Similar to 6RD, DS-Lite requires client support on the CPE to function. Client functionality is likely to be more prevalent in the future as IPv6 capable (WAN side) CPEs begin to penetrate the market. This includes both retail and operator provided gateways.

5.6. NAT64

NAT64 as described in [RFC6146] provides the ability to connection IPv6-Only connected clients and hosts to IPv4 Servers (or other like hosts). This technology, although useful in many circumstances, is not considered viable by many operators during early transition. NAT64 requires that the client, host or by extension the home network, supports IPv6-Only modes of operation. This type of environment is not considered typical in most traditional Wireline connections.

It is possible that in the future, NAT64 may become more viable for Wireline provides as home networking environments support IPv6-Only attachment modes, but until then, this technology is less useful for mass deployments in Wireline networks. As noted earlier, alternate technologies such as DS-Lite which still provide in-home IPv4 services though an IPv6-Only network (WAN) attachment are still of strong consideration.

6. IPv6 Transition Phases

The Phases described in this document are not provided as a ridged set of steps, but are considered a guideline which should be analyzed by an operator planning their IPv6 transition. The phases presented reflect the need to support IPv4 and IPv6 during the early to mid-term transition. The phased approach as presented in this document, attempts to match the most appropriate technologies for the various phases of the transition. The other key point of note with respect to this position on transition is the relationship between selected IPv6 transition technologies and overall traffic flow volumes.

During early transition, it is possible IPv6 traffic volumes will be present in most operator networks serving the Internet. As time

moves on more content is becoming available over IPv6 so this variable must be monitored by the operator. The early low volume conditions will most likely be attributable to IPv4-Only equipment in the home network and the Operator's access network. During these earlier time periods, technologies which "tunnel" IPv6 may be quite appropriate as operators attempt to provide IPv6 before the access network supports it . As time progresses and IPv6 traffic volumes rise, it may be desirable to provide a Native path for IPv6 service to better deal with the increased traffic volumes. Over time, IPv4 traffic volumes may be reduced as IPv6 traffic becomes the primary load in the Network. As the IPv4 traffic volumes lower, the operator may consider tunnelling this traffic if IPv4 resources are depleted or in short supply. Since the traffic levels are low, the scale needs to support this type of configuration would also be lower.

The overall objective with the phases provided is to also make sure the operator has prepared a solid foundation for IPv6 Services and is able to supply this in a timely manor to the customer base. Not all technologies which are technical available to the operator are included in this document and additional guidelines and information on utilizing IPv6 transition mechanisms can also be found in [RFC6180].

6.1. Phase 0 - Foundation

An operator considering an IPv6 service offering must initially be prepared to support it. These preparation steps are likely be to somewhat unique to each operator, but some basic items are well known, or at least common to most environments. These foundational steps include those listed below.

6.1.1. Phase 0 - Foundation: Training

Training is one of the most important steps in preparing an organization to support IPv6. Most resources in an organization have little to no experience with IPv6. Resources in organizations may only have a trivial understanding of IPv4 and given it's long history on the Internet, most may not be familiar with the intricacies of IP. Since there is likely to be many challenges with implementing IPv6 due to immature code on hardware and the evolution of many applications and systems to support IPv6 - it is of utmost important that organizations train their staff on IPv6 (and IP in general to that point).

Training should also be provided within reasonable timelines from actual IPv6 deployment. This means the operator needs to plan in advance as they train the various parts of their organization. New Technology and Engineering staff will require upfront training as

they plan and draw the designs for the network. Operation staff which support the network and other systems need to be trained closer to the deployment timeframes allowing them to more immediately use their new found knowledge and limiting memory loss issues. Customer support staff would require much more basic, but large scale training as may organizations have massive call centres to support the customer base.

6.1.2. 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 and related network functions. Since IPv6 is quite different from IPv4 in number of ways including the number of addresses which are made available, careful attention to a scalable and manageable architecture needs to be made. Also, given that home networks environments will no longer receive a token single address as is common in IPv4, operators will need to understand the impacts of delegating larges sums of addresses (Prefixes) to consumer endpoints. Delegating prefixes can be of specific importance in access network environments where downstream customers 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).

6.1.3. Phase 0 - Foundation: Network Policy and Security

Like many principles, network policy and security needs to be considered for IPv6. Although it is possible that many of the IPv4 policies may transfer transparently over to the IPv6 world, others may not be straight forward. There is also a potential that new policies need to be made to deal with issues specifically related to IPv6. This document does not highlight these specific issues, but raises the awareness they are of consideration and should be addressed when delivering IPv6 services.

6.1.4. Phase 0 - Foundation: Transition Architecture

The operator may want to plan out their transition architecture in advance (with obvious room for flexibility) to help optimize how they will build out and scale their networks. If the operator should want to use multiple technologies like CGN, DS-Lite and 6RD, they may want to plan out where such equipment may be located and potentially choose locations which can be used for all three functional roles (i.e. placement of NAT44 translator, AFTR and 6RD relays). This would allow for the least disruption as the operator evolves the transition environment to meet the needs of the network. This approach may also prove beneficial if traffic patterns change rapidly

in the future and the operator may need to evolve their network quick then originally anticipated.

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 of importance for both network resident gear and more importantly CPEs. Once deployed it's difficult and expensive to replace equipment. Vendors need to be brief and ready to pre-load or upgrade their systems to support the technology suites planned for deployment.

6.1.5. Phase 0- Foundation: Tools and Management

Although many of the tools and and service management systems may change over the course of the IPv6 transition, this area is of specific note. The operator may want to do a thorough analysis in advance as to what systems will need to be modified to deal with the interowrking models related to IPv6 service delivery. This will include address concepts related to the 128-bit addressing field, the notation of an assigned IPv6 prefix (PD) and the ability to detect either or both address families when determining if a customer has full Internet service.

If an operator stores usage information, this would need to be aggregated to include both the IPv4 and IPv6 traffic flows. Also, tools. that verify connectivity may need to query or interrogate the IPv4 and IPv6 addresses.

6.2. Phase 1 - Tunnelled IPv6

During the initial phase of transition the operator may want to support IPv6 Services before Native IPv6 can be supported by the access network. During this period of time, tunnelled access to IPv6 is a viable alternative to Native IPv6. Providers can deploy relays for automatic tunnelling technologies like 6to4 and Teredo, and can more importantly deploy technologies like 6RD. It should be noted that technologies like 6to4 and Teredo do not share the same address selection behaviours as those like 6RD as per address [RFC3484]. Additional guidelines on deploying and supporting 6to4 can be found in [RFC6343].

The operator can deploy 6RD relays quite easily and scale them as needed to meet the early customer needs of IPv6. Since 6RD requires the upgrade or replacement of most CPEs, the operator may want ensure that the CPEs support not just 6RD but Native Dual Stack and other tunnelling technologies if possible. 6RD client side deployments are now available in some retail channel products and within the OEM market making it a viable option for a wide range of operators.

Retail availability of 6RD is important since not all operators control or have influence over what equipment is deployed in the consumer home network which connects to the operator's network.

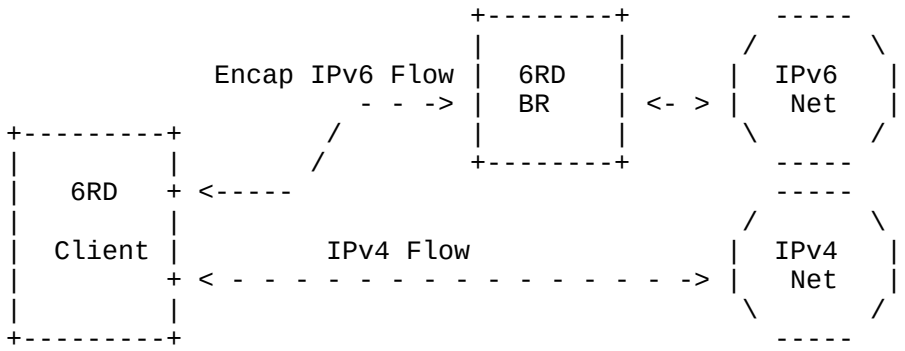


Figure 1: 6RD Basic Model

If the operator is able to support Native IPv6 right away, they may want to skip this phase. However, the operator may still want to deploy 6to4 and/or Teredo relays to assist connectivity for IPv4-Only connected customers which may have hosts using those protocols. 6RD used as an initial phase technology also provides the added benefit of a deterministic IPv6 prefix which is 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 customer host/CPE. So a simple tool and/or method can be built to help the operational folks in an organization know what the IPv6 prefix is for 6RD based on to knowledge of the IPv4 address.

An operator may choose to not offer internal services over IPv6 if such services generate a large amount of traffic. This mode of operation should avoid the need to greatly increase the scale of the 6RD Relay environment.

6.2.1. 6RD Deployment Considerations

Deploying 6RD can greatly speed up an operators ability to support IPv6 to the customer network. If considering deploying 6RD, an operator may want to consider who the system would be deployed, provisioned, scaled and managed. The operator may have additional considerations particular to their environment but these represent the core items which should be addressed.

The first core consideration is deployment models. 6RD requires the

CPE (6RD client) to send traffic to a 6RD relay. These relays can often share a common anycast address or use unique addresses. Both of these options are viable but each share benefits and challenges. Anycast options exist since 6RD is stateless by nature. 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 here is that it may be difficult to control large segments (or small segments) of the 6RD customer base as placement of the relays (in proximity to client) is the only way to steer traffic to new or alternate nodes. Proximity in this case actually refers to netowrk cost (i.e. in IGP) and not necessarily actual physical distance (although these can often be related). Use of specific addresses can help provide more control but has the disadvantage of being more complex to provision as CPEs will contain different information. An alternative approach is to use a hybrid model using multiple anycast service IPs for clusters of 6RD relays should the operator anticipate massive scaling of the environment. This way, the operator has multiple vectors by which to scale the service.

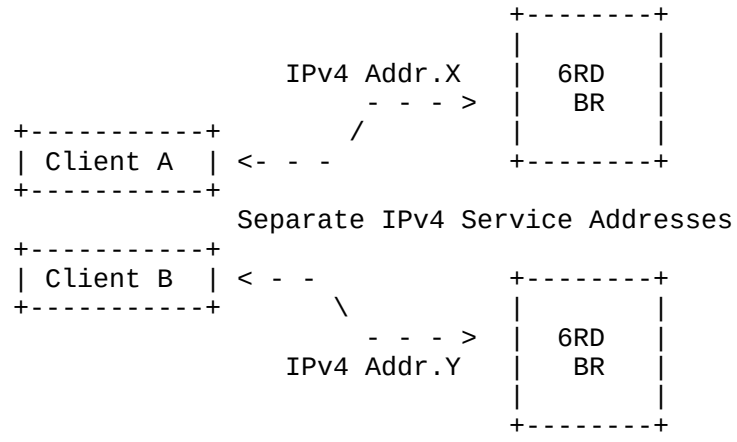


Figure 2: 6RD Multiple IPv4 Service Address Model

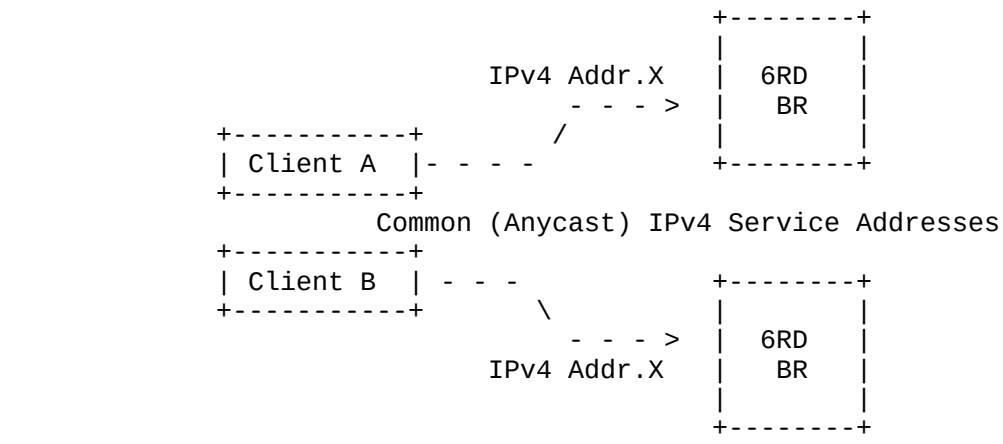


Figure 3: 6RD Anycast IPv4 Service Address Model

Provisioning of the endpoints is of consideration to the operator. This provisioning is also impacted by the deployment model chose (i.e. Anycast vs. specific service IPs). Using multiple IPs may require more planning and management as CPEs will have different sets of data to be provisioned into the devices. The operator will also need to decide if they will use DHCPv4, manual provisioning or other mechanisms to set the parameters into the CPEs.

If the operator wishes to managed the CPEs they will need to have access to new management tools or functions which are able to report the status of the 6RD tunnel to the inquiring support personnel. Also, if an operator needs to collect usage information, they would need to understand where this operation can take place. If the usage information includes understanding actual source/destination flow details, this information would likley be best collected after the 6RD relay (IPv6 side of connection). The operator will also need to be mindful of what tools they will need to manage such connections.

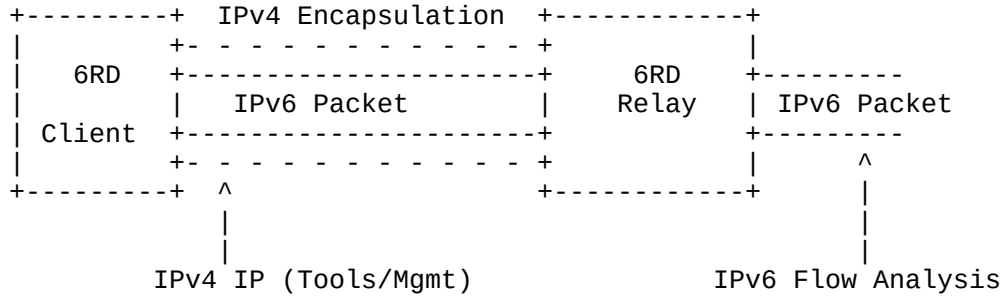


Figure 4: 6RD Tools and Flow Management

6.3. Phase 2: Native Dual Stack

Either as a follow-up phase to "Tunnelled IPv6" or as an initial step, the operator may deploy Native IPv6 to the customer premise. This phase would then allow for both IPv6 and IPv4 to be natively accessed by the customer home gateway/CPE. The Native Dual Stack phase be rolled out across the network while the tunnelled IPv6 service remains running. As areas begin to support Native IPv6, customer home equipment can be set to use it in place of technologies like 6RD. If 6to4 and/or Teredo was the sole method of connectivity prior to IPv6 service deliver then the internal home network hosts will naturally prefer the IPv6 address delivered via Native IPv6 (assumed to be a Delegated Prefix as per [RFC3769]).

As one of the most desirable options, Native Dual Stack should be sought as soon as possible if the operator’s network allows. 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 allows both protocols (IPv6 and IPv4) to work efficiently within the network. Efficiency in this context refers to the need (or lack there of) to translate, tunnel, incrementally route or relay customer traffic within the operator’s network.

6.3.1. Native Dual Stack Deployment Considerations

Native Dual Stack is a very desirable option for deployment. That said, it also requires a number of things to be in place before IPv6 it should be turned on. The operator is assumed to have a fully operational IPv6 network core and peering before they attempt to turn on Native IPv6 services. Additionally, supporting systems such as DHCPv6, DNS6 and other functions which support the customers IPv6 Internet connection need to be in place.

The operator will need make sure the IPv6 environment is stable and secure to ensure fluid operation. Poor IPv6 service may be worse then not offering an IPv6 service at all. Given that many platforms have very recent code which has enabled IPv6 or other functions which support IPv6 operation, instability may be experienced at first. The operator will need to be fully aware of the IPv6 service and it’s attributes to make sure they catch erroneous behaviour and address it promptly.

Of particular importance is the management of delegated prefixes. Prefix assignment and routing is a new concept for common residential services. The ability to assign the IPv6 prefix may be somewhat

strait forward (DHCPv6 using IA_PDs) but installation and propagation of this information is not. Operators who may see access layer instability impacting service if the route is not re-installed. Incrementally the operator may often re-assign customers to new IP Access nodes (such as in a Cable network) may need to consider this as PD information may not be transferable to the new location.

Operators will also needs to build new tools that help managed the IPv6 connection and will need to update systems to keep track of both the dynamically assigned IPv4 and IPv6 addresses. Any additional dynamic elements, such as auto-generated DNS names, need to be considered and planed for.

6.4. Intermediate Phase for CGN

As some point during the first two phases, acquiring more IPv4 addresses may become challenging or impossible, therefore CGN may be required on the IPv4 path. The CGN infrastructure can be enabled if needed during either phase. CGN is less optimal in a 6RD deployment (if used with 6RD to a given endpoint) since all traffic must transverse some type of operator service node (relay and translator).

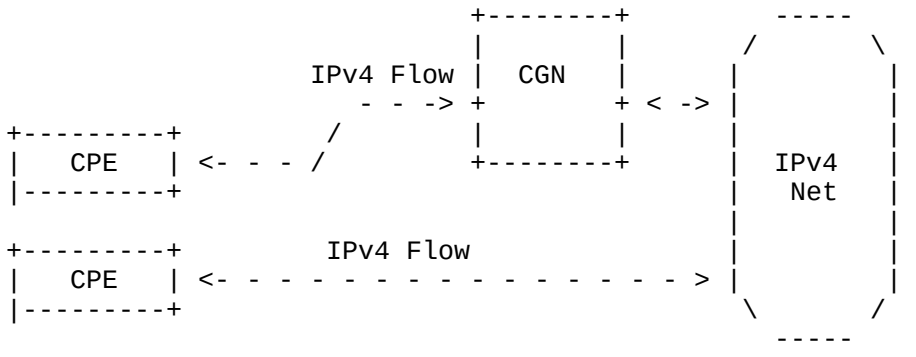


Figure 5: Overlay CGN Deployment

In the case of Native Dual Stack, CGN 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 model the Native IPv6 path is available for higher quality connectivity helping host operation over the network while the CGN path may offer a less then optimal performance.

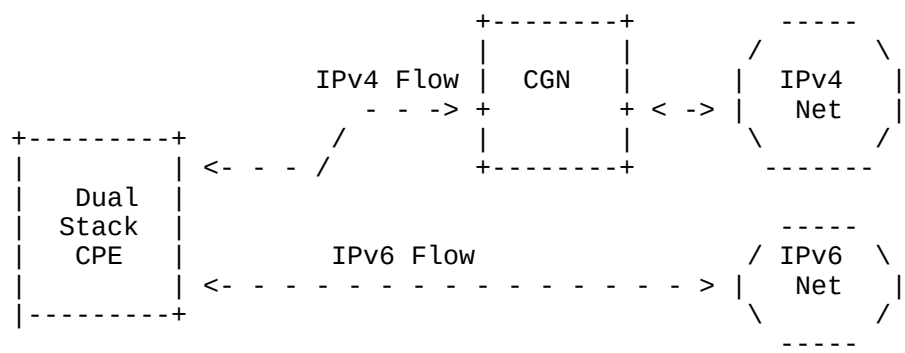


Figure 6: Dual Stack with CGN

CGN deployments may make use of a number of address options which include RFC1918 or Shared CGN Address Space [I-D.weil-shared-transition-space-request]. It is also possible that operators may use part of their own RIR assigned address space for CGN zone addressing if RFC918 address 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.

6.4.1. CGN Deployment Considerations

CGN is often considered undesirable by operators but required in many cases. An operator who needs to deploy CGN services should consider it's impacts 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 also be needed at low scale at first and grown to meet future demands based on traffic and connection dynamics of the customer, 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 and only spend money on equipment with the actual growth of traffic (demand on CGN system). The operator will need a deployment model and architecture which allows the system to scale as needed.

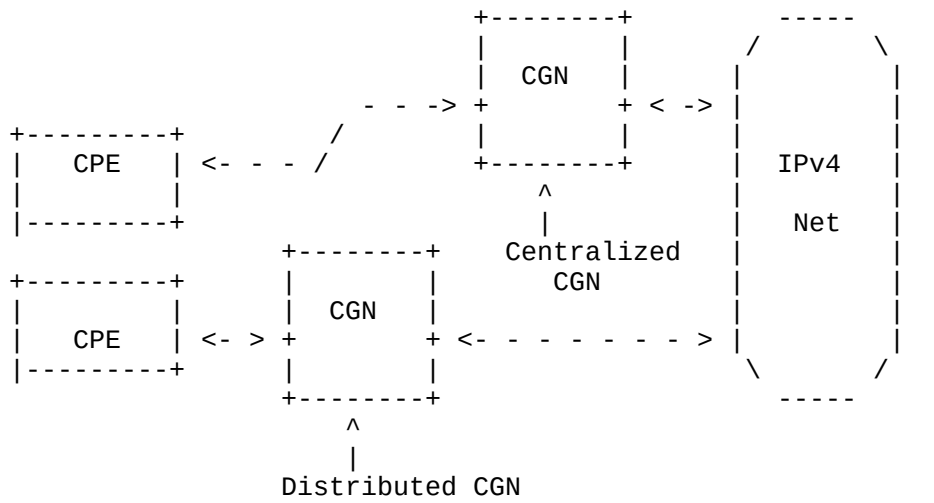


Figure 7: CGN Deployment: Centralized vs. Distributed

CGNs also increase the demands (potentially) for operators due to new phenomenon related to shared addressing. This includes logging of translation information for lawful response. This logging may require significant investment in external systems which ingest, aggregate and report on such information.

6.5. Phase 3 - Tunnelled IPv4

Over time, the operator will mature the IPv6 service and have more ubiquitous coverage within the network. Once the operator is familiar with IPv6, tools have been developed and operational procedures refined, more efficient modes of connectivity can be enabled. Once such technology is DS-Lite. DS-Lite allows the operator to grow the IPv4 customer base if needed without the need to deploy more IPv4 addresses to customer home networks. DS-Lite still requires IPv4 address sharing for IPv4 Internet connectivity, but this is seen as no worse and often more advantageous then CGN (NAT44) because only a single layer of NAT is required.

The operator can also move endpoints (Dual Stack) to DS-Lite retroactively in an attempt to reclaim IPv4 addresses for redeployment. Redeployment of addressees may be desirable if IPv4 resources are needed for legacy equipment and service connections which cannot be upgraded to IPv4 and no new IPv4 addressees can be acquired otherwise. The operator may want to have already moved most external content and internal content to IPv6 before this phase implemented. By having a significant amount of traffic on IPv6, the

operator would limit the amount of translation resources which are needed at the AFTR layer to support IPv4 flows. This would also be a benefit to the customer as their traffic need not be translated by a operator device improving performance.

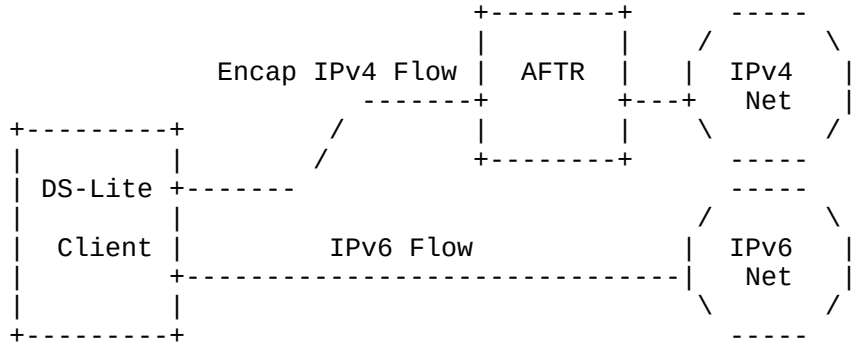


Figure 8: DS-Lite Basic Model

If the operator was forced to enable CGN for a NAT444 deployment, they may be able to co-locate the AFTR and CGN functions within the network to simplify capacity management and the engineering of flows. This phase can also co-exist with Native Dual Stack if desired since the same basic foundation is needed for both technologies on the IPv6 side. DS-Lite however requires incremental functions in the network such as the programming of the CPE and the implementation of the AFTRs'.

6.5.1. DS-Lite Deployment Considerations

DS-Lite although quite useful has a number of considerations for the operator. First all the same deployment considerations associated with Native IPv6 deployments are applicable to DS-LIte. The IPv6 network and service must be running well to ensure a quality experience for the end customer. IPv4 will now be subject to IPv6 service quality - this is a very important point. Tools will need be written or used to help manage the encapsulated IPv4 service which to not likely exist in most operators arsenal today. If flow analysis is required for IPv4 traffic, this may need to be enabled at a point beyond the AFTR or the operator will need equipment that can decapsulate DS-Lite to see inside the packets.

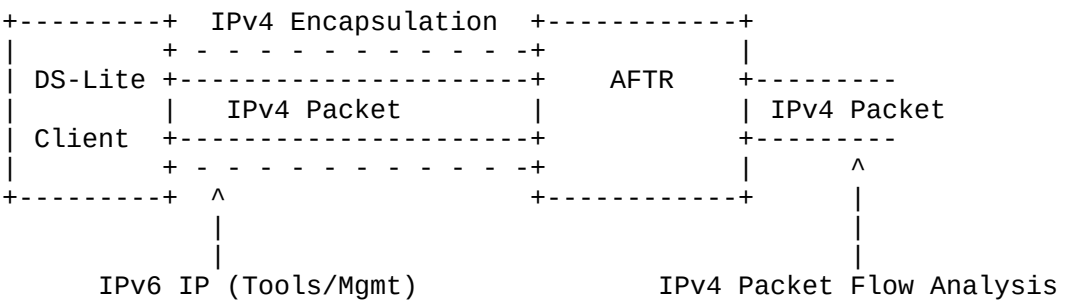


Figure 9: DS-Lite Tools and Flow Analysis

DS-Lite also requires client support. If the operator has chose to have a vendor support multiple transition technologies, the activation logic will need to be clearly articulated such that the correct behaviour is manifest in the network. As an example, an operator may use 6RD in the outset of the transition, then move to Native Dual Stack followed by DS-LIte.

7. IANA Considerations

No IANA considerations are defined at this time.

8. Security Considerations

No Additional Security Considerations are made in this document.

9. Acknowledgements

Thanks to the following people for their textual contributions and/or guidance on IPv6 deployment considerations: John Brzozowski, Lee Howard, Jason Weil, Nik Lavorato, John Cianfarani, Chris Donley, Wesley George and Tina TSOU.

10. References

10.1. Normative References

[I-D.ietf-v6ops-v4v6tran-framework]
Carpenter, B., Jiang, S., and V. Kuarsingh, "Framework for IP Version Transition Scenarios", draft-ietf-v6ops-v4v6tran-framework-02 (work in progress),

July 2011.

[RFC6180] Arkko, J. and F. Baker, "Guidelines for Using IPv6 Transition Mechanisms during IPv6 Deployment", RFC 6180, May 2011.

10.2. Informative References

[I-D.donley-nat444-impacts]
Donley, C., Howard, L., Kuarsingh, V., Chandrasekaran, A., and V. Ganti, "Assessing the Impact of NAT444 on Network Applications", draft-donley-nat444-impacts-01 (work in progress), October 2010.

[I-D.ietf-behave-lsn-requirements]
Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common requirements for Carrier Grade NAT (CGN)", draft-ietf-behave-lsn-requirements-03 (work in progress), August 2011.

[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-04 (work in progress), September 2011.

[I-D.weil-shared-transition-space-request]
Weil, J., Kuarsingh, V., Donley, C., Liljenstolpe, C., and M. Azinger, "IANA Reserved IPv4 Prefix for Shared CGN Space", draft-weil-shared-transition-space-request-07 (work in progress), October 2011.

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

[RFC3056] Carpenter, B. and K. Moore, "Connection of IPv6 Domains via IPv4 Clouds", RFC 3056, February 2001.

[RFC3068] Huitema, C., "An Anycast Prefix for 6to4 Relay Routers", RFC 3068, June 2001.

[RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.

[RFC3769] Miyakawa, S. and R. Droms, "Requirements for IPv6 Prefix Delegation", RFC 3769, June 2004.

[RFC4380] Huitema, C., "Teredo: Tunneling IPv6 over UDP through Network Address Translations (NATs)", RFC 4380, February 2006.

[RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.

[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.

[RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.

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

[RFC6343] Carpenter, B., "Advisory Guidelines for 6to4 Deployment", RFC 6343, August 2011.

Author's Address

Victor Kuarsingh (editor)
Rogers Communications
8200 Dixie Road
Brampton, Ontario L6T 0C1
Canada

Email: victor.kuarsingh@gmail.com
URI: http://www.rogers.com

Network Working Group
Internet Draft
Intended status: Best Current Practice
Expires: August 28, 2013

B. Liu
S. Jiang
Huawei Technologies
C. Byrne
T-Mobile USA
February 25, 2013

Guidance of Using Unique Local Addresses
draft-liu-v6ops-ula-usage-analysis-05.txt

Status of this Memo

Copyright Notice

Abstract

This document provides guidance of how to use ULA. It analyzes ULA usage scenarios and recommends use cases where ULA address may be beneficially used.

Table of Contents

1. Introduction

Unique Local Addresses (ULAs) are defined in [RFC4193] as provider-independent prefixes that can be used on isolated networks, internal networks, and VPNs. Although ULAs may be treated like global scope by applications, normally they are not used on the publicly routable internet.

However, the ULAs haven’t been widely used since IPv6 hasn’t been widely deployed yet.

The use of ULA addresses in various types of networks has been confused for network operators. Some network operators believe ULAs are not useful at all while other network operators run their entire networks on ULA address space. This document attempts to clarify the advantages and disadvantages of ULAs and how they can be most appropriately used.

2. ULA usage analysis

2.1. The features of ULA

2.1.1. Self-assigned

ULA is self-assigned, this feature allows automatic address allocation, which is beneficial for some lightweight systems and can leverage minimal human management.

2.1.2. Globally unique

ULA is intended to be globally unique to avoid collision. Since the hosts assigned with ULA may occasionally be merged into one network, this uniqueness is necessary. The prefix uniqueness is based on randomization of 40 bits and is considered random enough to ensure a high degree of uniqueness (refer to [RFC4193] section 3.2.3 for details)and make merging of networks simple and without the need to renumbering overlapping IP address space. Overlapping is cited as a deficiency with how [RFC1918] addresses were deployed, and ULA was designed to overcome this deficiency.

Notice that, as described in [RFC4864], in practice, applications may treat ULAs like global-scope addresses, but address selection algorithms may need to distinguish between ULAs and ordinary GUA (Global-scope Unicast Address) to ensure bidirectional communications.

2.1.3. Independent address space

ULA provides an internal address independence capability in IPv6 that is similar to how [RFC1918] is commonly used. ULA allows administrators to configure the internal network of each platform the same way it is configured in IPv4. But the ability to merge two ULA networks without renumbering (because of the uniqueness) is a big advantage over [RFC1918].

On the other hand, many organizations have security policies and architectures based around the local-only routing of [RFC1918] addresses and those policies may directly map to ULA. ULA can be used for internal communications without having any permanent or only intermittent Internet connectivity. And it needs no registration so that it can support on-demand usage and does not carry any RIR documentation burden or disclosures.

2.1.4. Well known prefix

The prefixes of ULAs are well known and they are easy to be identified and easy to be filtered.

This feature may be convenient to management of security policies and troubleshooting. For example, the administrators can decide what parameters have to be assembled or transmitted globally, by a separate function, through an appropriate gateway/firewall, to the Internet or to the telecom network.

2.1.5. Stable or Temporary Prefix

A ULA prefix can be generated once, at installation time or "factory reset", and then never change unless the network manager wants to change. Alternatively, it could be regenerated regularly, if desired for some reason.

2.2. Enumeration of ULA use scenarios

In this section, we try to cover plausible possible ULA use case. Some of them might have been discussed in other documents and are briefly reviewed in this document as well as other potential valid usage is discussed.

2.2.1. Isolated network

IP is used ubiquitously. Some networks like RS-485, or other type of industrial control bus, or even non-networked digital interface like MIL-STD-1397 began to use IP protocol. In such kind of networks, the system might lack the ability/requirement of connecting to the Internet.

Besides, some networks are explicitly designed to not connect to the internet. These networks may include machine-to-machine, sensor networks, or other types of SCADA networks which may include very large numbers of addresses and explicitly prohibited from connect to the global internet (electricity meters...).

ULA is a straightforward way to assign the IP addresses in these kinds of networks with minimal administrative cost or burden.

2.2.2. Connected network

2.2.2.1. ULA-only Deployment

In some situations, hosts/interfaces are assigned with ULA-only, but the networks need to communicate with the outside. For example, just like many implementations of private IPv4 address space [RFC1918]. One important reason of using private address space is the lack of IPv4 addresses, but this it is not an issue any more in IPv6. Another reason is regarding with security, private address space is designed by some administrators as one layer of a multilayer security. Such design is also applicable in IPv6 with using ULAs.

But we should eliminate the misunderstanding that ULA is designed to be the IPv6 version of [RFC1918] deployment model. If you chose non-globally routable address space for some reasons, ULA is a nature selection, but we need to know ULA itself is not designed for this intention.

ULA-only in connected network may include the following two models.

o Using Network Prefix Translation

Network Prefix Translation (NPTv6) [RFC6296] is an experimental specification that provides a stateless one to one mapping between internal addresses and external addresses.

In some very constrained situations(for example, in the sensors), the network needs ULA as the on-demand and stable addressing which doesn’t need much code to support address assignment mechanisms like DHCP or ND. And the network also needs to connect to the outside, then there can be a gateway to be the NAT which may not be so sensitive to the constrained resource. This behavior could refer NPTv6 [RFC6296].

o Using application-layer proxies

The proxies terminate the network-layer connectivity of the hosts and delegate the outgoing/incoming connections.

There may be some scenarios that need this kind of deployment for some special purpose (strict application access control, content monitoring, e.g.).

2.2.2.2. ULA along with GUA

There are two classes of network probably to use ULA with GUA addresses:

- Home network. Home networks are normally assigned with one or more globally routed PA prefixes to connect to the uplink of some an ISP. And besides, they may need internal routed networking even when the ISP link is down. Then ULA is a proper tool to fit the requirement. And in [RFC6204], it requires the CPE to support ULA.

- Enterprise network. An enterprise network is usually a managed network with one or more PA prefixes or with a PI prefix, all of which are globally routed. The ULA could be used for internal connectivity redundancy and better internal connectivity or isolation of certain functions like OAM of servers.

3. Recommended ULA Use Cases

3.1. Used in Isolated Networks

As analyzed in section 2.2.1, ULA is very suitable for isolated networks. Especially when you have subnets in the isolated networks, ULA is almost the only choice.

3.2. ULA along with GUA

For either home networks or enterprise networks, the main purpose of using ULA along with GUA is to provide a logically local routing plane separated from the globally routing plane. The benefit is to ensure stable and specific local communication regardless of the ISP uplink failure. This benefit is especially meaningful for the home network or private OAM function in an enterprise.

In some special cases such as renumbering, enterprise administrators may want to avoid the need to renumber their internal-only, private nodes when they have to renumber the PA addresses of the whole network because of changing ISPs, ISPs restructure their address allocations, or whatever reasons. In these situations, ULA is an effective tool for the internal-only nodes.

Besides the internal-only nodes, the public nodes can also benefit from ULA for renumbering. When renumbering, as RFC4192 suggested, it has a period to keep using the old prefix(es) before the new prefix(es) is(are) stable. In the process of adding new prefix(es) and deprecating old prefix(es), it is not easy to keep the local communication immune of global routing plane change. If we use ULA

for the local communication, the separated local routing plane can isolate the affecting by global routing change.

But for the separated local routing plane, there always be some argument that in practice the ULA+PA makes terrible operational complexity. But it is not a ULA-specific problem; the multiple-addresses-per-interface is an important feature of IPv6 protocol. Running multiple prefixes in IPv6 might be very common, and we need to adapt this new operational model than that in IPv4.

Another issue is mentioned in [RFC5220], there is a possibility that the longest matching rule will not be able to choose the correct address between ULAs and global unicast addresses for correct intra-site and extra-site communication. In [RFC6724] , it claimed that a site-specific policy entry can be used to cause ULAs within a site to be preferred over global addresses.

3.3. Special Use Cases

3.3.1. Special routing

If you have a special routing scenario, of which [draft-baker-v6ops-b2b-private-routing] is an example, for various reasons you might want to have routing that you control and is separate from other routing. In the b2b case, even though two companies each have at least one ISP, they might choose to also use direct connectivity that only connects stated machines, such as a silicon foundry with client engineers that use it. A ULA provides a simple way to obtain such a prefix that would be used in accordance with an agreement between the parties.

3.3.2. Used as NAT64 prefix

Since the NAT64 pref64 is just a group of local fake addresses for the DNS64 to point traffic to a NAT64, the pref64 is a very good use of ULA. It ensures that only local systems can use the translation resources of the NAT64 system since the ULA is not globally routable and helps clearly identify traffic that is locally contained and destine to a NAT64. Using ULA for Pref64 is deployed and it is an operational model.

But there’s an issue should be noticed. The NAT64 standard [RFC6146) mentioned the pref64 should align with [RFC6052], in which the IPv4-Embedded IPv6 Address format was specified. If we pick a /48 for NAT64, it happened to be a standard 48/ part of ULA (7bit ULA famous prefix+ 1 "L" bit + 40bit Global ID). Then the 40bit of ULA is not violated to be filled with part of the 32bit IPv4 address. This is

important, because the 40bit assures the uniqueness of ULA, if the prefix is shorter than /48, the 40bit would be violated, and this may cause conformance issue. But it is considered that the most common use case will be a /96 PREF64, or even /64 will be used. So it seems this issue is not common in current practice.

It is most common that ULA Pref64 will be deployed on a single internal network, where the clients and the NAT64 share a common internal network. ULA will not be effective as Pref64 when the access network must use an Internet transit to receive the translation service of a NAT64 since the ULA will not route across the internet.

3.3.3. Used as identifier

Since ULA could be self-generated and easily grabbed from the standard IPv6 stack, it is very suitable to be used as identifiers by the up layer applications. And since ULA is not intended to be globally routed, it is not harmful to the routing system.

Such kind of benefit has been utilized in real implementations. For example, in [RFC6281], the protocol BTMM (Back To My Mac) needs to assign a topology-independent identifier to each client host according to the following considerations:

- TCP connections between two end hosts wish to survive in network changes.
- Sometimes one needs a constant identifier to be associated with a key so that the Security Association can survive the location changes.

It should be noticed again that in theory ULA has the possibility of collision. However, the probability is desirable small enough and could be ignored by most of the cases when used as identifiers.

4. Security Considerations

Security considerations regarding ULAs, in general, please refer to the ULA specification [RFC4193].

5. IANA Considerations

None.

6. Conclusions

ULA is a useful tool, it have been successfully deployed in a diverse set of circumstances including large private machine-to-machine type networks, enterprise networks with private systems, and within service providers to limit Internet communication with non-public services such as caching DNS servers and NAT64 translation resources.

We should eliminate the misunderstanding that ULA is just an IPv6 version of [RFC1918]. The features of ULA could be beneficial for various use cases.

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, BCP14, March 1997.

[RFC4193] Hinden, R., B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

7.2. Informative References

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

[RFC4864] Van de Velde, G., Hain, T., Droms, R., Carpenter, B., and E. Klein, "Local Network Protection for IPv6", RFC 4864, May 2007.

[RFC5220] Matsumoto, A., Fujisaki, T., Hiromi, R., and K. Kanayama, "Problem Statement for Default Address Selection in Multi-Prefix Environments: Operational Issues of RFC 3484 Default Rules", RFC 5220, July 2008.

[RFC6281] Cheshire, S., Zhu, Z., Wakikawa, R., and L. Zhang, "Understanding Apple’s Back to My Mac (BTMM) Service", RFC 6281, June 2011.

[RFC6296] Wasserman, M., and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, June 2011.

[RFC6724] Thaler, D., Draves, R., Matsumoto, A., and Tim Chown, "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 6724, September, 2012.

[draft-baker-v6ops-b2b-private-routing]
          F. Baker, "Business to Business Private Routing", Expired

8. Acknowledgments

Many valuable comments were received in the mail list, especially from Fred Baker, Brian Carpenter, Anders Brandt and Wesley George.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Bing Liu
Huawei Technologies Co., Ltd
Huawei Q14 Building, No.156 Beiqing Rd.,
Zhong-Guan-Cun Environmental Protection Park, Beijing
P.R. China

EMail: leo.liubing@huawei.com

Sheng Jiang
Huawei Technologies Co., Ltd
Huawei Q14 Building, No.156 Beiqing Rd.,
Zhong-Guan-Cun Environmental Protection Park, Beijing
P.R. China

EMail: jiangsheng@huawei.com

Cameron Byrne
T-Mobile USA
Bellevue, Washington 98006
USA

Email: cameron.byrne@t-mobile.com

v6ops
Internet-Draft
Intended status: Informational
Expires: September 13, 2012

C. Xie
China Telecom
X. Li
Tsinghua University
J. Qin
Consultant
M. Chen
FreeBit

A. Durand

Juniper Networks

March 12, 2012

Practice of IPv4/IPv6 transition system for data center
draft-sunq-v6ops-contents-transition-03

Abstract

This document describes deployment practice of IPv4/IPv6 translation technologies for data center transition, aiming at rapidly increasing the amount of IPv6 accessible contents for users from IPv6 Internet while preserving the continuity of IPv4 service delivery. System based on this design has been deployed in production network to provide transition service for several ICP websites.

Table of Contents

1. Introduction

Facing the pressure of IPv4 address shortage, the operators may like to provide services through IPv6 by upgrade their IP infrastructure to support IPv6. As part of the Infrastructure, Data center (in short, IDC) is the main faculty to house service system that provides services and contents. It is obvious that data center also plays an important role in IPv6 transition in accordance with the transition of IP network. Dual-stack is the basic transition strategy for most data centers, as well as IP transport network. However, in our practices, we found that dual-stack alone is not enough to meet the transition demand of ICPs(in short, ICP) in data centers. The reason behind this is that providing IPv6 services requires the service software of ICP, i.e., website system, database system, supporting system, etc., should be IPv6-aware and can deal with IPv6-related information. Upgrading the service system to support IPv6 is technological-complicated and financially costly, especially for some small and medium-sized ICPs, which is the main reason that the IPv6 transition on the ICP sides moves even more slowly than the readiness of operators' IP network. The lack of IPv6-reachable contents becomes one of the main obstacles. On the other hand, some progressive ICPs who are willing to setup an IPv6-only system also would like to offer IPv4 continuity for end-users.

Under such circumstances, we propose to deploy IDC transition system in data center, aiming at aiding CP/SP to provide IPv6 services rapidly and smoothly. Another purpose of our approach is to increase the amount of IPv6 accessible contents for users from IPv6 Internet. It can also keep the IPv4 continuity for IPv6-only contents.

This document describes our current experiences on two deployment models for the transition of data center based on the approaches specified by IETF (e.g., NAT64 [RFC6146], Dual-Stack [RFC4213], IVI[RFC6219], etc.), targeting different use cases or conditions. Based on these models, an IDC transition system was designed and developed by China Telecom to provide transition services to ICPs in data centers. Some issues and considerations were also identified from the actual deployment.

2. 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].

3. Motivations

As mentioned above, IDC's transition is closely related to the IPv6 service provisioning of ICPs. There have been statements from several popular ICPs that they have turned on IPv6 (no matter by which means), which do have a beneficial effect on encouraging end users' transition to IPv6. However, given the operational cost, it is still difficult for most ICPs (especially the great many ones of small-to- medium size) to immediately make their publically-facing services accessible through both IPv4 and IPv6 natively. It will involve a lot of workload for upgrading numerous application systems and the supporting systems in ICPs. On the other hand, from the users' perspective, the IPv6 reachability of resources required for their daily lives is one of the foremost concerns when making the decision on whether or not to access Internet using IPv6. It is a chicken or egg dilemma, but the two perspectives are interdependent. If the transition of one side passes the point of inflexion, the other side will be speeded up after. So, more efforts are needed to encourage the IPv6 adoption and reach the point.

Moreover, some progressive ICPs are willing to maintain a separated IPv6-only system, which will lower the risk of the potential impact on their existing wildly used IPv4 system in the early phase. Besides, single-stack system is also easy for operation, management and troubleshooting. There are no duplicated policies need to be applied, including e.g. ACL control, accounting, authentication, etc. In this case, it is also the requirement to offer IPv4 continuity to IPv6-only contents.

Therefore, the transition system provided by operators in data centers will not only help promote ICP transition in a step-by-step way, but also break out the chicken or egg dilemma for the whole IPv6 industry.

3.1. Transition As A Service

In China Telecom, we have deployed a transition platform in our IDC network. It can be regarded as transition services offered by the operators, to small-to-medium size ICPs (e.g., those who rent servers from the operators).

The ICPs can choose to take different approaches according to their scenarios and business strategies. For the conservative ones, the IPv4 services can be still offered natively, and the IPv6 services can be offered by the stateful IPv4/IPv6 translation [RFC6146]. While for progressive ones and newly incomers, the stateless IVI [RFC6219], [RFC6052] can be employed to offer native IPv6 services reachable via IPv4.

3.2. Guiding the traffic to IPv6 network

IPv4 address shortage has driven some network providers began to run IPv6 in part or the whole network. However, even if IPv6 is ready in the IP network, most ICPs in IDC have not been ready to provide IPv6 services. As a result, almost all the traffic is still IPv4-based, which makes the IPv6 network nearly empty. With this in mind, IPv4/ IPv6 translation system deployed in IDC can translate the IPv4 packets sourced from the existing servers into IPv6 packets, and forward them into IPv6 network, which is equal to move the traffic from IPv4 network to IPv6 network. and encourage the customers to use IPv6 from the beginning. Furthermore, only translation will be performed on the edge of the network and it is independent of user-side transition mechanisms.

4. Deployment practice one: Communication from IPv6 users to IPv4 server

4.1. Deployment scenario

We have deployed transition service gateway in the exit of our IDCs. It is a shared platform which can serve multiple servers simultaneously. It can be integrated with existing network element of our IDC, e.g. egress router, load balancer, etc., or can be deployed as a new standalone device. The integrated deployment scenario would have little impact on existing network topology; however, it is highly coupled with existing devices. The standalone deployment scenario would be easier to implement on existing network incrementally. However, it will result in extra cost for new devices.

The egress router of our IDC is IPv6-reachable, however, either the content servers or the whole IDC infrastructure have been upgraded to IPv6 directly. With the help of transition gateway, we can provide IPv6 reachable content to customers in a quick manner. Our deployment model is depicted in the following picture.

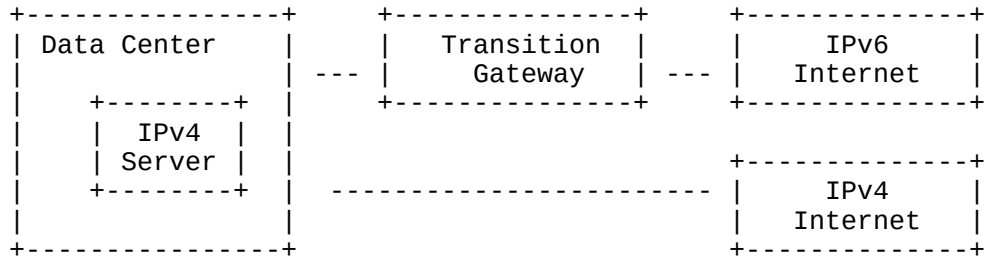


Figure 1: Deployment Model 1

In this deployment model, the Stateful NAT64 is performed to translate IPv6 packets to IPv4 and vice versa. The guidance in [RFC6146] should be followed. The communications are initiated from the IPv6 side. When an IPv6 packet arrives, a lookup of the mapping table will be carried out to get the IPv4 address used for the translation. If there is no one matched, a new entry will be created.

The server-side deployment model is independent of user-side transition When a dual-stack user gets both A and AAAA records for a remote server, it will be encouraged to reach IPv4 content via IPv6 connectivity through the only NAT64 gateway along the path. So even if there are some other CGNs deployed in the customer-side, IPv6 traffic will be forwarded in a traditional way. Therefore, there will be no double-translation problems around here.

Up to now, there are 8 sites including the official website of China Telecom have been upgrading to IPv6 with this mechanism. More than 15 thousands different IPv6 users ever accessing the above eight ICPs through the transition box totally, with 4000 to 6000 active users every day. www.voc.com.cn is the most popular one accessed by more than 4000 IPv6 users daily, and www.chinatelecom.com.cn (the official website of china telecom) has amounts of access from 1200 IPv6 users on average every day.

4.2. Mapping and Addressing

The Stateful NAT64 can support the following two mapping modes:

- 1:1, one IPv6 address is mapped to one IPv4 address (exclusively for given lifetime);
- N:1, each of the IPv4 addresses (i.e. IPv4 address pool) will be shared by multiple IPv6 users from Internet.

To save global IPv4 addresses which has become scarce resource, private blocks, for instance 10.0.0.0/8 may be used for the Stateful NAT64. This private address block can only be seen within the IDC network.

Considering the scale of traffic in the foreseeable future, the 1:1 Mapping Mode with private blocks (one IPv6 address mapped to one private IPv4 address within 10.0.0.0/8) is selected as the default mode for the Stateful NAT64. In this mode, there is only address-layer mapping and no TCP/UDP session maintenance anymore. By this mean, the efficiency of stateful operations could be improved and the

problems introduced by the address sharing could be alleviated (for example, the burden of logging will be reduced in this mode).

However, there may be conflicts if the same private space is used internally for the interconnection of servers (e.g. multiple servers for load balancing). In this case, N:1 mode with public blocks can be used. In order to reduce state management burden in N:1 stateful NAT64 gateway as well as logging system, a bulk of ports can be allocated for each subscriber. In this port-set based mapping mode, one IPv6 address will be mapped to the same IPv4 address and a given port-set.

In addition, an IPv6 prefix is used to serve the IPv4 servers in the IDC, and the route of the prefix has been advertised to the IPv6 Internet. The IPv4 address of the server can be embedded in the IPv6 prefix following the algorithm specified in [RFC6052].

4.3. DNS

To make sure the addresses of servers can be retrieved by IPv6 users before initiating sessions, the AAAA records which formed through IPv4-translated addresses have been added directly on the domain's authoritative DNS, or upgrade authoritative DNS to support DNS64. In this way, the AAAA records under one domain name could be retrieved by IPv6 users around the world.

Please note that if the authoritative DNS of given ICPs' domain names are maintained by some third-party DNS Providers but not by themselves or the operator from whom this transition service (i.e. the deployment model of Stateful NAT64 discussed herein) is purchased, the ICPs must make sure the authoritative AAAA records can be added.

4.4. Fragmentation

Basically, the processing of packets carrying fragments follows the guidance specified in [RFC6145] and [RFC6146] with exceptions that fragmented IPv4/IPv6 packets will be firstly reassembled to an integrated packet before doing packet translation and so on.

4.5. Logging

The logging is essential for tracing back specific users in stateful NAT64. In 1:1 mode, only per-user logging events need to be recorded as {IPv6 address, IPv4 address, timestamp}. For N:1 mode, in order to reduce the number of sessions need to be logged, we adopt port-set based mechanism to assign a bulk of ports to each subscriber. Therefore, one subscriber will only create one corresponding log

report, e.g. {IPv4 address, IPv6 address, port-set, timestamp}.

4.6. Geographically aware services

Since converted IPv4 address would not represent any geographical feature anymore, applications that assume such geographic information may not work as intended.

Two solutions were designed and implemented, one is to maintain the above logging information in geographic server as well, and offer an open API to ICPs to retrieve its original IPv6 address when necessary. It will have little impact on NAT64 gateway since there is no application-layer procedure. However, due to the transmission and computational latency in geographic servers, it is more suitable for ICPs to retrieve IPv6 users' source address offline. Another way is to embed user's source IPv6 address in x-forward field of user's request when it traverses NAT64 gateway. This involves application-layer process which will bring extra burden on NAT64 gateway. So only for ICPs who really need online users' source address will be offered with this additional service.

4.7. ALG issues

Since the types of applications are relatively limited due to the deployment policy, it would be easier to solve the ALG issue compared to client-side deployment. For example, Web-based ICPs might be introduced in the first stage, and so specific ALGs can be applied accordingly.

Since video traffic constitutes a great portion of the whole Internet traffic, we have implemented HTTP AGLs for video traffic in particular.

In our test for TOP100 Websites in China, there are basically three types of HTTP ALGs for video traffic.

HTTP/1.1 302 Found: This is a common way of performing a redirection. Usually, IPv4 address literals for redirected server will be embedded in Location header.

HTTP/1.1 301 Moved Permanently: This is also a redirect way indicating the requested resource has been assigned a new permanent place, and the IPv4 address literals for redirected server will also be embedded in Location header.

HTTP/1.1 200 ok: This code means the request has succeeded. However, some ICPs will still embed the IPv4 address literals to indicate the redirected server in the following communication, and

they will use a great variety of keywords. For example, www.sina.com.cn uses the keyword "CDATA[http://" followed by a list of IPv4 addresses, and v.6.cn use "watchip" as its keyword.

Since the first two types occupy the great majority of existing ALGs for HTTP-based videos traffic, we have implemented the ALG for the first two cases to synchronize an IPv4-translated address if the server of the embedded IPv4 address is located within the NAT64 region.

4.8. High Availability

In general, there are two mechanisms to achieve high reliability, i.e. cold-standby and hot-standby. In cold-standby mode, the NAT64 states are not replicated from the Primary NAT64 gateway to the Backup NAT64 gateway. When the Primary NAT64 gateway fails, all the existing established sessions will be flushed out. The hosts are required to re-establish sessions with the external hosts. Another high availability option is the hot standby mode. In this mode the NAT64 gateway keeps established sessions while failover happens. The 1:1 mapping mode will greatly reduce the amount of sessions needed to be replicated on-the-fly from the Primary NAT64 gateway to the Backup gateway. Another option is to deploy an Anycast NAT64 prefix. This is similar to cold-standby that NAT64 states are not replicated between Primary gateway and Backup gateway, except that the heartbeat line is not needed anymore.

4.9. Security

The security issues and considerations discussed in [RFC6146] apply to the deployment model described in this document. However, when deploying stateful NAT64 in server side, it is hard to apply source-based filtering policy. As a result, we have introduced alarming mechanism to report the current status of state-consuming speed in NAT64 gateway.

Besides, both 1:1 mapping mode and port-set based N:1 mapping mode can guarantee that one IPv6 source address will be mapped to a single IPv4 address. Therefore, the ICP can identify a single subscriber either by IPv4 source address in 1:1 mapping, or IPv4 source address plus port-set in N:1 mapping.

4.10. Deployment practices

Up to now, there are 8 sites including the official website of China Telecom have been upgrading to IPv6 with this mechanism. More than15 thousands different IPv6 users ever accessing the above eight Content Providers through the transition box totally, with 4000 to 6000

active users every day. www.voc.com.cn is the most popular one accessed by more than 4000 IPv6 users daily, and www.chinatelecom.com.cn (the official website of china telecom) has amounts of access from 1200 IPv6 users on average every day.

5. Deployment practice two: communications from IPv4 users to IPv6 server

5.1. Deployment scenario

Considering in the foreseeable future, IPv6 will be a widely accepted protocol in the Internet, some ICPs, especially newcomers, will setup IPv6-only servers, to reduce the operation and maintenance complexity. When the server in question itself is IPv6-capable,communications initiated from IPv6 users will not encounter any transition problem. What we are concerned is the communications initiated from IPv4 users. To mitigate this problem, IPv4/IPv6 translation is utilized in the IDC that the server resides. In this scenario, the IPv4 node will firstly get A/AAAA records of the server from DNS, and then the communication will follow the path to NAT64 Gateway. When an IPv4 packet arrives at NAT64 Gateway, it would be translated to an IPv6 packet based on stateless 1:1 mapping algorithm [RFC6219].

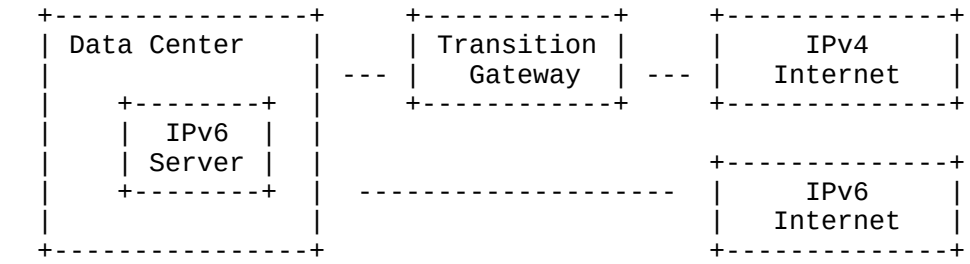


Figure 2: Deployment Model2

5.2. Mapping and Addressing

To eliminate the state management burden, we adopted stateless transition gateway to do the Interworking between IPv4 Internet and IPv6-only server within IDC, IPv6-only server should be configured with an IPv4-translatable address. Then both source address and destination address are applied with 1:1 mapping to keep the simplicity and transparency.

In addition, an IPv4 address within the range of a given IPv4 prefix is used to represent the IPv6 server, and the route of the IPv4

prefix has been advertised to the IPv4 Internet. An IPv6 prefix will be assigned to the IDC to represent the whole IPv4 Internet, when IPv4 packet traverse the transition gateway, IPv6 addresses, e.g., source address and destination address, will be formed by combine the IPv4 address with a IPv6 prefix following the algorithm specified in [RFC6052]. In this way, the server can be reachable from IPv4 Internet without mapping states in transition gateway.

5.3. DNS

To make sure that addresses of servers can be retrieved by IPv4 users before initiating sessions, the A records which are extracted from IPv4-translated addresses should be added directly on the domain's authoritative DNS, or upgrade authoritative DNS to support DNS64. Other considerations are actually the same with Section 4.

5.4. Logging

There is no logging issue in stateless transition solution.

5.5. Geographically aware services

When a ICP gets an IPv4-converted IPv6 addresses with a pre-defined Prefix, it should extract the embedded IPv4 address which would reflects its original geographical information.

5.6. ALG issues

ALG issues would be the same with section 4.6.

5.7. High Availability

Since there is no state maintained in the transition gateway, state replication or re-establishment encountered in the HA of the first deployment model will not exist in the second one.

5.8. Security

IPv4/IPv6 translators which can be modeled as special routers, are subject to the same risks, and can implement the same mitigations. (The discussion of generic threats to routers and their mitigations is beyond the scope of this document.) There is, however, a particular risk that often happens in IPv4 Internet: address spoofing.

An attacker could use a faked IPv4 address as the source address of malicious packets. After translation, the packets will appear as IPv6 packets from the specified source, and the attacker may be hard

to track. If left without mitigation, the attack would allow malicious IPv4 nodes to spoof arbitrary IPv4 addresses.

The mitigation is to implement reverse path checks and to verify throughout the network that packets are coming from an authorized location.

5.9. Deployment practices

The following IPv6-only websites has been setup to provide native IPV6 service to IPv6 users, all of them are hosted in a dual-stack IDC.

http://iptv.bupt.edu.cn

http://www.mayan.cn

http://www.ivi.buptnet.edu.cn

In order to accommodate the access of great volume of existing IPv4-only users, stateless transition gateway was deployed to provide translation in the exit of the IDC. Currently, the peak of the traffic is around 900Mbps.

6. Additional Author List

Qiong Sun

China Telecom

Room 708 No.118, Xizhimenneidajie

Beijing, 100035

P.R.China

Phone: +86 10 5855 2923

Email: sunqiong@ctbri.com.cn

Qian Liu

China Telecom

No.359 Wuyi Rd.,

Changsha, Hunan 410011

P.R.China

Phone: +86 731 8226 0127

Email: 18973133999@189.cn

Qin Zhao

BUPT

Beijing 100876

P.R.China

Phone: +86 138 1127 1524

Email: zhaoqin@bupt.edu.cn

7. IANA Considerations

This document includes no request to IANA.

8. Acknowledgements

The authors would like to thank Fred Baker, Joel Jaeggli, Erik Kline, Randy Bush for their comments and feedback.

9. References

9.1. Normative References

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

[RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

[RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.

[RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, April 2011.

[RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation

Algorithm", RFC 6145, April 2011.

[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.

[RFC6154] Leiba, B. and J. Nicolson, "IMAP LIST Extension for Special-Use Mailboxes", RFC 6154, March 2011.

[RFC6219] Li, X., Bao, C., Chen, M., Zhang, H., and J. Wu, "The China Education and Research Network (CERNET) IVI Translation Design and Deployment for the IPv4/IPv6 Coexistence and Transition", RFC 6219, May 2011.

9.2. Informative References

[I-D.wing-behave-http-ip-address-literals]
Wing, D., "Coping with IP Address Literals in HTTP URIs with IPv6/IPv4 Translators", draft-wing-behave-http-ip-address-literals-02 (work in progress), March 2010.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

Authors' Addresses

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

Phone: +86 10 5855 2116
Email: xiechf@ctbri.com.cn

Xing Li
Tsinghua University
Room 225, Main Building
Beijing 100084
P.R.China

Phone: +86 10 6278 5983
Email: xing@cernet.edu.cn

Jacni Qin
Consultant
Shanghai,
China

Phone: +86 1391 861 9913
Email: jacniq@gmail.com

Maoke Chen
FreeBit Co., Ltd.
13F E-space Tower, Maruyama-cho 3-6
Shibuya-ku, Tokyo 150-0044
Japan

Email: fibrib@gmail.com

Alain Durand
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089-1206
USA

EMail: adurand@juniper.net

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: May 17, 2012

W. Townsley
Cisco
A. Cassen
Free Telecom

November 14, 2011

6rd Sunsetting
draft-townsley-v6ops-6rd-sunsetting-00

Abstract

This document provides guidelines for transitioning an IPv6 deployment using IPv6 Rapid Deployment (6rd) to an IPv6 deployment using Native IPv6. It is targeted at both 6rd operators and 6rd implementors."

Table of Contents

1. Introduction

6rd [RFC5969] specifies a protocol mechanism to deploy IPv6 to sites via a service provider's (SP's) IPv4 network. The 6rd mechanism uses an algorithmic mapping between IPv4 and IPv6 within the SP network. This mapping allows for automatic IPv6 prefix delegation as well as determination of IPv6 over IPv4 tunnel endpoints within the SP, providing for stateless operation.

Unlike 6to4 [RFC3056], 6rd is designed to be configured and operated by an SP. It is expected than an SP providing 6rd configuration will do so with the same considerations as with native IPv6.

This document describes two incremental 6rd "sunsetting" models, each with different tradeoffs. The best model for an SP will depend on the specific operational considerations for that SP. One model requires a 6rd user to be renumbered to a new IPv6 prefix when native configuration is applied. The other model allows an SP to move to native IPv6 in a "seamless" mode which does not require renumbering or multihoming with separate prefixes.

The CE requirements identified in this document provide a basis for both modes.

2. 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].

3. Terminology

This document uses terms as defined in section 3 of [RFC5969], in particular Customer Edge (CE) to refer to a router at the edge of a customer site, 6rd Border Relay (BR), 6rd domain, 6rd delegated prefix, CE LAN side, CE WAN side, etc. Please refer to the definitions in [RFC5969] for more information.

4. Incremental Sunsetting With Renumbering

Perhaps the most obvious method for sunsetting 6rd is to bring up native IPv6 users in a separate prefix from that provided by 6rd, then disabling 6rd at the same time or later. The CE LAN side is then either being served by 6rd from one IPv6 prefix, native from another IPv6 Prefix, or both at any given time.

Using this mode, end-user sites are renumbered when moving from 6rd to native IPv6. Recommendations for "Renumbering Without a Flag Day" described in [RFC4192] should be followed. When 6rd and native IPv6 are both active with different prefixes, the CE is effectively multihomed via two separate interfaces (one physical, one virtual).

Paradoxically, traffic may decrease less than expected or even increase at the 6rd BRs as users are moved from 6rd to native IPv6. This is because the native IPv6 users are considered by the rest of the 6rd users to be outside of their 6rd domain, as such the BR will be required to be traversed for traffic that before was handled directly between 6rd CEs. Ideal BR placement may also change as new native IPv6 users are added and the footprint of the 6rd domain is altered.

5. Incremental Sunsetting Without Renumbering

A perhaps less obvious sunsetting approach allows for incremental native deployment without requiring site renumbering and no increase of traffic at the Border Relays as native IPv6 is deployed. In this "seamless mode" the native link is configured by the ISP with the same IPv6 prefix as 6rd calculates from IPv4. The aim is to allow the network to use the native interface when it can and the 6rd interface when it cannot via simple forwarding metrics. As there is no new delegated prefix introduced, this mode avoids complications with multihoming and renumbering.

Following is a set of basic steps an operator might employ:

1. 6rd Deployment.

2. CEs reachable by Native IPv6 are configured via DHCPv6-PD [RFC3633] with the same delegated prefix calculated and in use by 6rd.

3. CEs keep native and 6rd interfaces active, with a single (unchanged) prefix for the CE LAN side. There is no affect on the home site.

4. When native IPv6 becomes active on a given CE, an upstream default route is installed for IPv6 as it normally would, while assigning a metric that causes the native link to be preferred over 6rd. 6rd routes remain as long as 6rd is configured by the SP, allowing the more-specific route for inter-domain 6rd traffic to be selected over the native route (again, following normal IP forwarding rules). This allows CE to CE traffic to continue over 6rd, while "off-net" IPv6 traffic destined for outside the 6rd

domain will be sent over the native link.

5. If the operator wishes to direct incoming traffic from outside the 6rd domain towards native CEs directly, this may be done by injecting an IPv6 prefix within the ISPs IGP, or by configuring static routes directly on the BR. The level of granularity here is up to the operator, anywhere from a single site for testing, up to a set of sites corresponding to a specific aggregation level, region, or block of addresses as long as all of the sites for which the prefix is being injected have native IPv6 enabled. This allows a progressive removal of traffic which must traverse the 6rd BR function commensurate with the deployment of native IPv6.

6. Once Native IPv6 is fully deployed, 6rd is disabled at the BRs. This should be done first at the BRs allowing all native traffic to and from outside the 6rd domain to switch to native, and then at the CEs to move all intradomain 6rd traffic to native. Again, this may be done incrementally, by first testing a handful of CEs without 6rd enabled, and then moving forward at a pace determined by the individual operator.

7. The final result will be a native IPv6 deployment with the same IPv4-based numbering plan that 6rd required. If the SP would like to obtain better aggregation or move to a different IPv6 prefix entirely (as may be required by some RIR policies), renumbering may then be safely performed now that 6rd is fully decommisioned.

6. CE Requirements

The following CE requirements are sufficient for "Incremental Sunsetting Without Renumbering", and provide a basis for "Incremental Sunsetting With Renumbering."

1. A 6rd CE MUST continue to allow 6rd packets to be sent and received as long as 6rd configuration is provided by the ISP, even while links on the router are configured with native IPv6.

2. A 6rd CE MUST assign a forwarding metric such that native IPv6 egress is preferred for traffic outside the 6rd domain when 6rd and native IPv6 interfaces are active.

3. 6rd and native IPv6 MUST allow for an identical IPv6 delegated prefix.

Specific CE requirements for renumbering of a residential site itself

are out of scope of this document.

7. Security Considerations

There are no specific additional security issues identified at this time.

8. IANA Considerations

None.

9. Acknowledgements

10. References

10.1. Normative References

[RFC1332] McGregor, G., "The PPP Internet Protocol Control Protocol (IPCP)", RFC 1332, May 1992.

[RFC1661] Simpson, W., "The Point-to-Point Protocol (PPP)", STD 51, RFC 1661, July 1994.

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

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

[RFC2491] Armitage, G., Schulter, P., Jork, M., and G. Harter, "IPv6 over Non-Broadcast Multiple Access (NBMA) networks", RFC 2491, January 1999.

[RFC2516] Mamakos, L., Lidl, K., Evarts, J., Carrel, D., Simone, D., and R. Wheeler, "A Method for Transmitting PPP Over Ethernet (PPPoE)", RFC 2516, February 1999.

[RFC3056] Carpenter, B. and K. Moore, "Connection of IPv6 Domains via IPv4 Clouds", RFC 3056, February 2001.

[RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.

[RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.

[RFC3964] Savola, P. and C. Patel, "Security Considerations for 6to4", RFC 3964, December 2004.

[RFC4192] Baker, F., Lear, E., and R. Droms, "Procedures for Renumbering an IPv6 Network without a Flag Day", RFC 4192, September 2005.

[RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.

10.2. Informative References

[RFC3068] Huitema, C., "An Anycast Prefix for 6to4 Relay Routers", RFC 3068, June 2001.

[RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.

[RFC5569] Despres, R., "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd)", RFC 5569, January 2010.

Authors' Addresses

Mark Townsley
Cisco
Paris,
France

Phone:
Email: mark@townsley.net

Alexandre Cassen
Free Telecom
Paris,
France

Phone:
Email: acassen@freebox.fr