

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 5, 2012

G. Chen
T. Yang
L. Li
H. Deng
China Mobile
July 4, 2011

IPv6 Practices on China Mobile IP Bearer Network
draft-chen-v6ops-ipv6-bearer-network-trials-00

Abstract

This memo has introduced IPv6 practices on China Mobile IP bearer network, in which IP backbone network and broadband access routers have been covered. In the practice, IPv6 protocol conformance and data packages forwarding capabilities have been tested in multi-vendors environments. Several IPv6 transition schemes have been deployed to validate interoperabilities. Based on concrete testing data, IPv6-enable deployment experiences have been learned to share with community.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
2. IPv6 trial on IP backbone network	4
2.1. IP Backbone Topology for Trials	4
2.2. IPv6-only Routing Protocol Testing	6
2.3. Dual-stack Routing Protocol Testing	7
2.4. 6PE/6vPE Protocol Testing	7
2.5. Tunnel Protocol Interoperabilities	7
2.6. IPv6 ACL and Policy Routing Configuration	8
2.7. Summary for IPv6 Trial on IP Backbone Network	8
3. IPv6 Testing on BRAS	9
3.1. Test Topology	9
3.2. Test Cases- Basic IPv6 protocols	12
3.3. Test Cases- DUT in Network Test	12
3.4. Test Cases- Performance in IPv6 environment	13
3.5. Summary for BRAS Testing	13
4. IANA Considerations	13
5. Security Considerations	13
6. Normative References	13
Authors' Addresses	14

1. Introduction

With fast development 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 only way to satisfy demands of Internet development. Operators have to accelerate the process of deploying IPv6 networks in order to address IP address strains.

With significant demands of service development, China Mobile has officially kicked off first IPv6 pre-commercial trials on IP bearer network on June 2011 after several standalone tests of IP equipment in labs. The trials have taken place on major IP backbone network and broadband access equipments. In order to verify IPv6 feasibility and applicability, IPv6 protocol conformance and data packages forwarding capabilities have been tested in multi-vendors environments. Several IPv6 transition schemes, i.e. dual-stack, 6PE[RFC4789], 6vPE[RFC4659], have been deployed to validate interoperabilities. Based on the IPv6 trials, concrete testing data have been generated and analysed to provide informative assessment to facilitate IPv6 deployment in next steps.

This memo has described detailed testing topology, cases and process both on IP backbone network and BRAS. The testing results have been summarized and analyzed to provide explicit conclusions for further deployment.

2. IPv6 trial on IP backbone network

This section will describe IPv6 trial on IP backbone network in details. It includes testing topology, testing cases. Based on collected testing data, we have summarized testing results.

2.1. IP Backbone Topology for Trials

Figure 1 depicts the overall topology for IP backbone trials, which is constituted by hierarchical IPv6 enable routers. The same level would deploy double-routers due to redundancy considerations. The top-level is national IP backbone network to connect provincial level networks. The middle level is provincial IP backbone to connect metropolitan area networks. The bottom level stands for core IP routers to connect local area networks. The trials have been taken place on these three levels. In order to simulate user-generated data, two router testers have been positioned to access metropolitan core routers. They have responsibilities to propagate routing information into the network under testing.

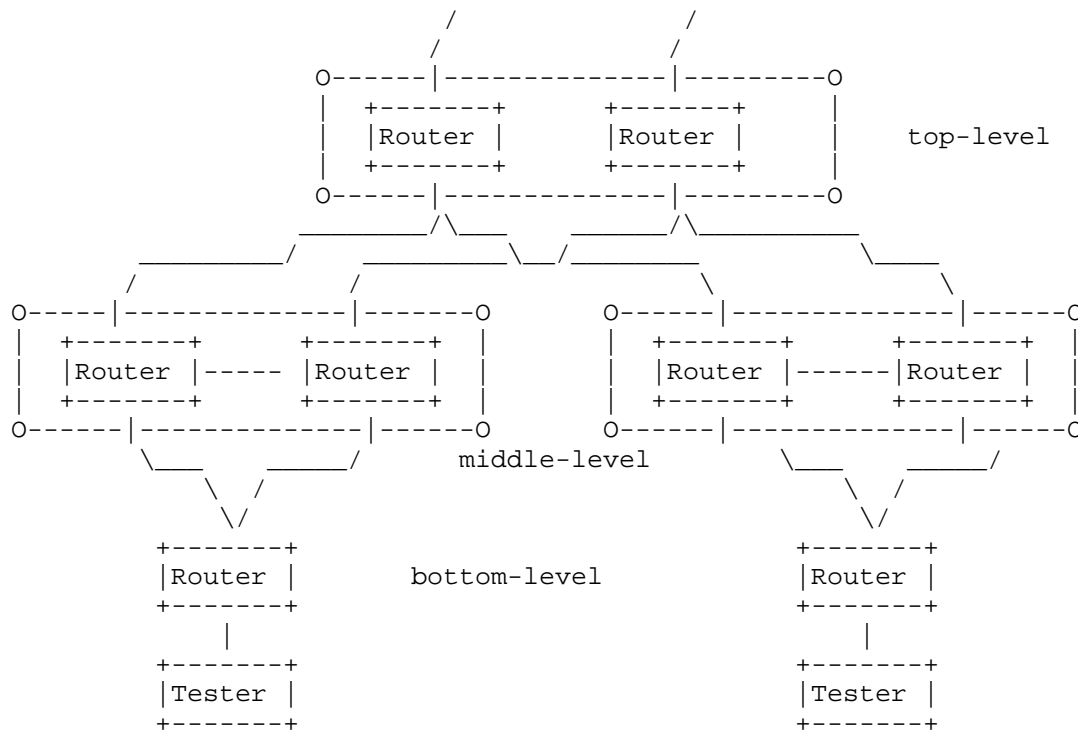


Figure 1: IP Backbone Topology for Trials

The test cases mainly are composed by several parts:

- o IPv6-only routing protocol interoperabilities: the test case would shutdown IPv4 stack on tested routers. Only IPv6 stack is running to process IPv6 EGP(i.e. BGP4+[RFC2545]) and IGP routing protocol(i.e. OSPFv3[RFC2740] and ISIS[RFC5308]). It will validate IPv6 routing protocol conformance in multi-vendor environments.
- o Dual-stack routing protocol interoperabilities: the tested router would run both IPv4 and IPv6 IGP and EGP protocol simutanously. It will verify if remote peers could totally learn spreaded IPv4/IPv6 routing information.
- o 6PE/6vPE protocol interoperabilities: the test case would require PE routers to be upgraded to support 6PE/6vPE functionalities and remain MPLS core network staying IPv4-enable. It will testify MP-BGP routing learning and package forwarding capabilities.

- o Tunnel protocol interoperabilities: the test case would configure PE routers as 6over4 tunnel end-point. After configured 6over4 tunnel is established, the encapsulated package would be forwarded through MPLS network.
- o IPv6 ACL and policy routing capabilities: the target of this test case aiming at IPv6 traffic restrainability. A pair PE would be configured with a particular policy to constrain data forwarding for specific IPv6 traffic. The verification will help to enhance network security.

The following sub-sections would state testing results and related observations.

2.2. IPv6-only Routing Protocol Testing

The testing is going to validate IPv6 routing protocol interconnection between multi-vendor routers. OSPFv3 and ISIS have been deployed as Interior Gateway Protocol. Based on IGP, BGP4+ has been configured as EGP to communicate routing information.

In the case of OSPFv3 deployment, routers on the top-level and middle-level have been schemed as area 0, which takes responsibility of backbone area. And, routers on the bottom-level has been configured as area 100 and area 200 accessing to backbone area. The testers inject IPv6 routing information to see whether propagated IPv6 routing information can be learned by remote routers located on the other side of bottom-level. The testing is finalized that routers on bottom-level still remain Exstar/Exchange states. OSPFv3 can't create adjacencies between neighboring routers for the purpose of exchanging routing information. After troubleshooting, IPv6 MTU between neighboring routers is inconsistent, which causes it failed to establish adjacencies. It is fixed by adjusting IPv6 MTU as identical. It's recommended that IPv6 MTU should be configured as unified benchmark.

In the case of ISIS deployment, routers on the top-level and middle-level have been schemed as level 2, which takes responsibility of backbone area. And, routers on the bottom-level has been configured as level 1. The testers inject IPv6 routing information to see whether propagated IPv6 routing information can be learned by routers located on level 1. The testing has found out that Multi Topology (MT) Routing[RFC5120] in IS-IS would easily cause disorders in ISIS routing area. It's recommended that MT Routing in IS-IS should be enable or disable simultaneously.

In the case of BGP4+ deployment, one of routers on top-level has been selected as reflectors. The router on others level will establish

neighboring relationship with the router based on running ISIS route protocol. The testing has shown BGP4+ is running well on IPv6-enable network.

2.3. Dual-stack Routing Protocol Testing

Dual-stack routing testing runs IPv4 and IPv6 protocols simultaneously. The routing area planning is similar with IPv6-only routing protocol testing. For IPv4 routing protocol, OSPF and ISIS have been taken as IGP and BGP has been taken as EGP. Testing has shown that routing path have been computed by IPv4 and IPv6 routing algorithm independently. The IPv4/IPv6 routing information learning and data forwarding are conformed with testing expectation.

2.4. 6PE/6vPE Protocol Testing

6PE and 6vPE could shift network to provide IPv6 access depending on existing Multiprotocol Label Switching (MPLS) [RFC3032] core network. Operators could deploy IPv6 network without modifying IPv4 enable MPLS cloud. In the testing, the routers on bottom-level take responsibility of PE and routers tester simulate CE to inject IPv6 routing information and package flows. The routers on the top-level and middle-level would still remain IPv4 enable situation. MPLS protocol has been configured on these routers. During the testing, tester would serve for CE to inject routing information. In the case of 6PE, the tester will propagate normal IPv6 IGP routing information to the PE. In the case of 6vPE, the tester would generate IPv6 VPN routing information to communicate with remote peer through MP-BGP. The testing shown remote peers could learn total IPv6 routing information through MPLS enable cloud. The basic functionalities of 6PE/6vPE are going well in the testing network. The caution has been raised by forwarding big data packages. The intermedia routers would like to drop big packages surpassing network MTU since they can't fragment big package in the middle of the forwarding path. The data fragmentation functionalities are taken by initiated router. Therefore, it's recommended that the package size do not exceed network MTU by configuring maximum MTU on initiated router or enabling Path MTU Discovery on initiated router.

2.5. Tunnel Protocol Interoperabilities

Tunnel protocol requires dedicated board to support. The testing is focusing on configured 6over4 tunnel. In order to coordinate with existing deployed MPLS cloud. The tunnel end-point has been located on PE routers. In this case, IPv6 package is expected encapsulated in IPv4 package going through MPLS network, in which additional MPLS header with label would route 6over4 packages into remote peer. Therefore, the transmitted IPv6 data packages are not only

encapsulated in native IPv4 package, but headed into MPLS tunnel. Double tunneling is requested in this situation. According to the testing results, router are failed to support such encapsulation manner. The tunnel board can't forward data packages carrying both tunnel id and MPLS lable. However, MPLS encapsulation supporting is essential for IP bearer network since MPLS is widely deployed.

2.6. IPv6 ACL and Policy Routing Configuration

IPv6 ACL(Access Control List) will help to reduce potential risks of harmful invasion by preventing IPv6 traffic from a specific source address or network to a specified destination network. In the test scenario, the routers have been configured as IPv6-enable. Routers on bottle-level have configured with ACL deny policies which have applied for both outbound and inbound IPv6 traffics. Testers would inject IPv6 traffic to the routers to validate if the ACL takes effect. The results shown that only outbound ACL deny policies is valid. And inbound policies are failed to constrain IPv6 traffic due to the lacking of supports from the router board.

Another testing is taken place at policy routing redistribution through BGP4+. An restricted routing policy would be added into BGP4+ UPDATE message to propagate into the remote peer. After the testing, the remote peer could suceessfully learn the redistributed routing policies and take effect to limite paticular IPv6 traffic .

2.7. Summary for IPv6 Trial on IP Backbone Network

In general, the tests on several aspects indicate that IPv6-enable backbone network has already qualified for carrying IPv6 data packages in IPv6-only or dual-stack conditions. The IPv6 routing protocol has mature supporting in current routers. It also shows high-interoperabilities in multi-vendor environments. According to testing results, two points needed to be highlighted for next step in IPv6 deployment:

- o MTU configuration needs to be carefully configured and checked up. The inappropriate MTU configuration will cause failures of IGP neighboring establishment and packages dropping. The unified maxmum MTU configuration is recommended.
- o Multi Topology (MT) Routing in IS-IS should be configured in unified manner to avoid routing disorders in ISIS area.
- o Compared to other tunneling technologies, 6PE and 6vPE are recommended to be deployed on IP bearer network, in which MPLS is widely enable.

3. IPv6 Testing on BRAS

BRAS is the key equipment in MAN, it distributes IP addressess to the subscribers through DHCP protocols, authorize the subscribers to log on, and calculate their online time for billing. It is the "central control node" in MAN.

In the past several years, BRAS helped ISPs to control and record the subscribers' bahaviors in IPv4 networks. Along with IPv6 approaching day by day, several BRAS manufacturers announce their equipments could support IPv6 functions. In order to checkout that, we carried on testing five types of BRAS produced by four manufacturers.

Test results show that BRAS's IPv6 functions and capabilities are not optimistic. The following subsections describe the specific results of the tests.

3.1. Test Topology

We tested BRAS in four network scenarios in our labotory. The first scenario checks out the basic IPv6 protocols in stand-alone environment, the following two mostly test the DUTs' communicating ability in networks, and the last simplest configuration verifies the BRAS performance in IPv6 environment.

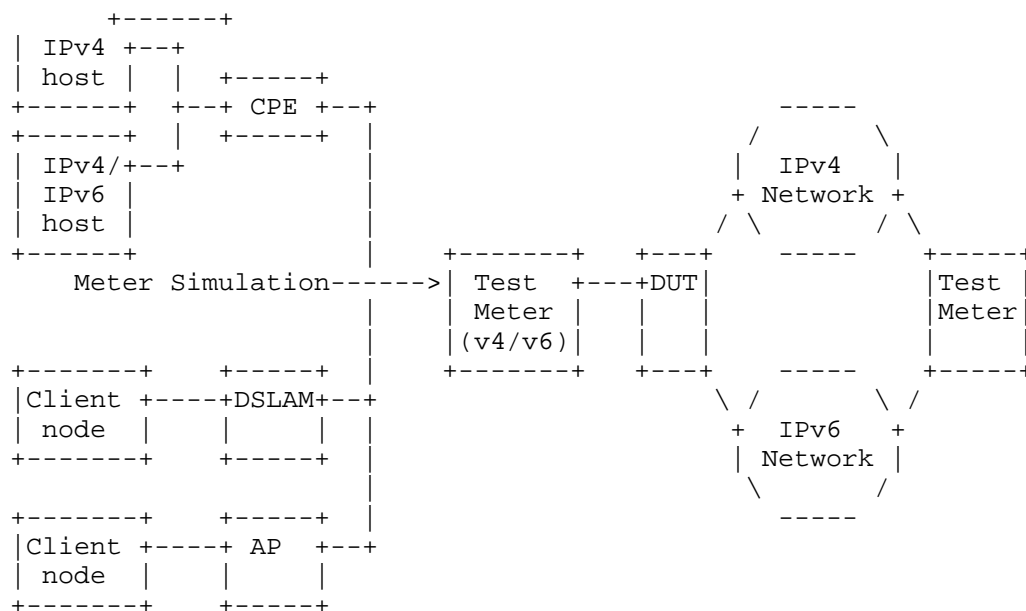


Figure 2: Test Topology 1 (stand-lone test)

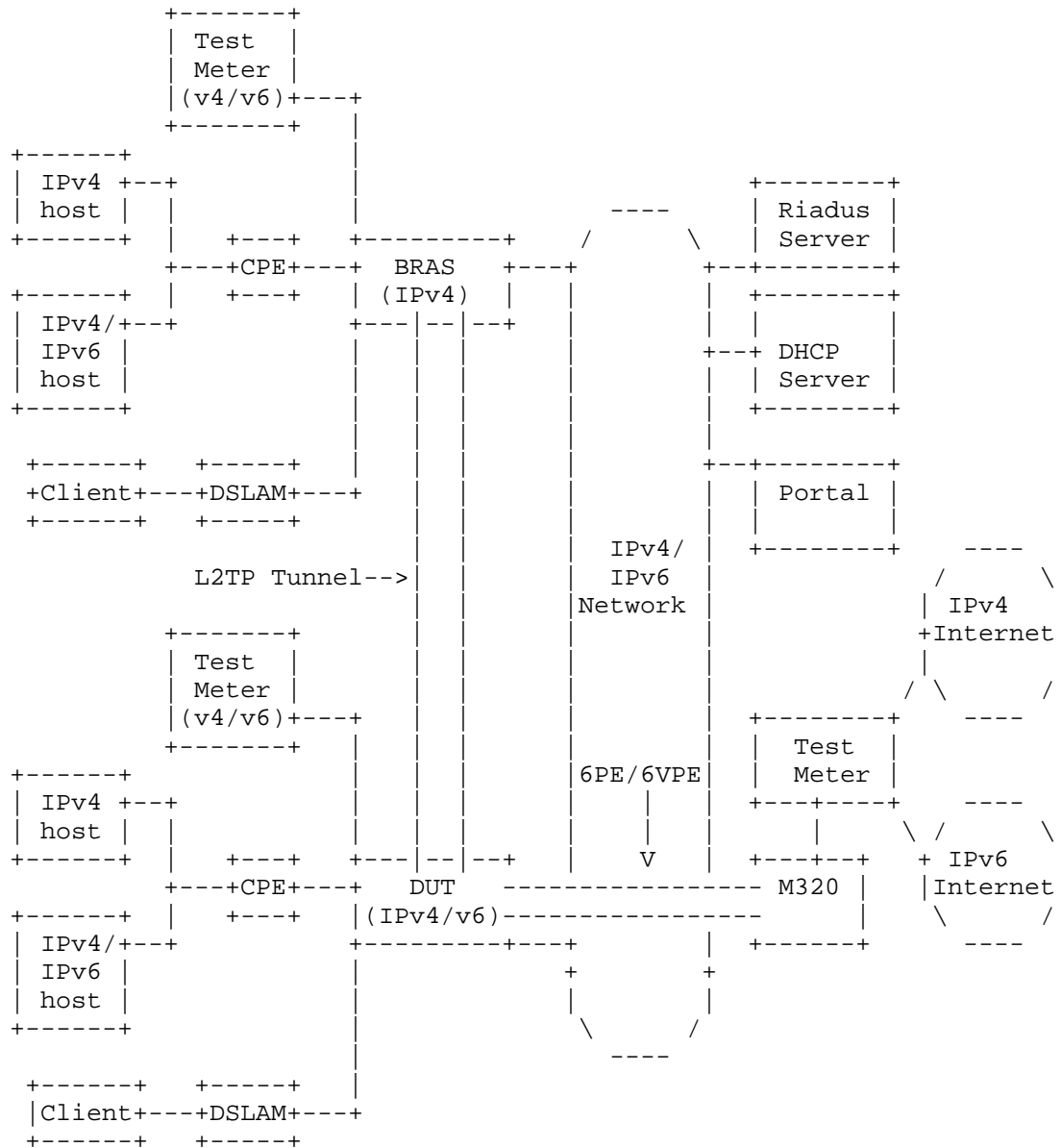


Figure 3: Test Topology 2 (DUT in networks)

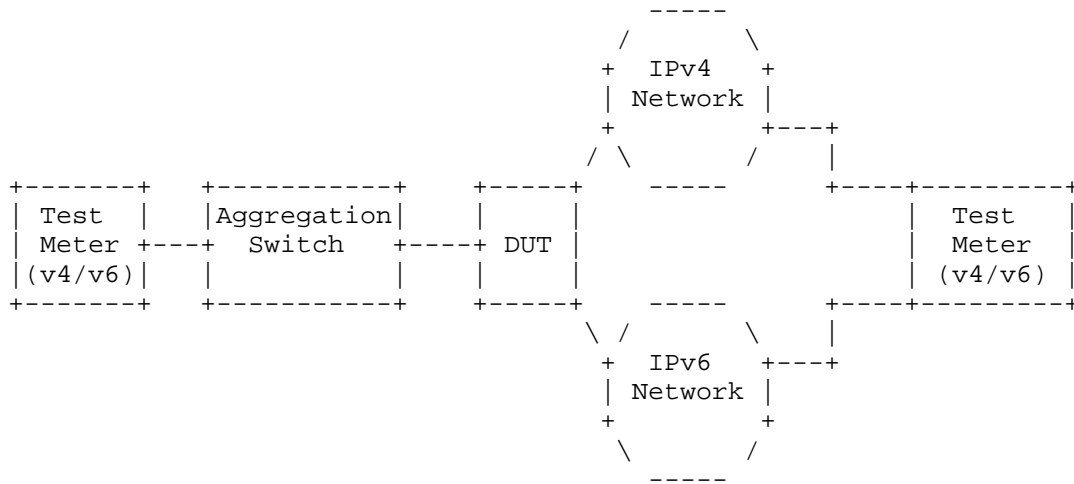


Figure 4: Test Topology 3 (DUT in networks)

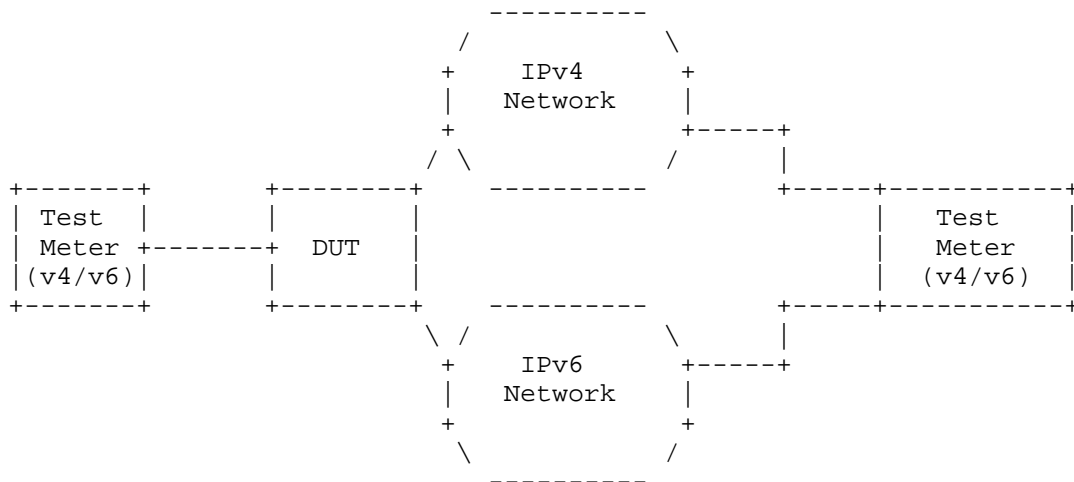


Figure 5: Test Topology 4 (performance Test)

3.2. Test Cases- Basic IPv6 protocols

Along with devices transiting from IPv4 to IPv6, BRAS's IP-based protocol stack will undergo a major change. Some new protocols will come into being, and some existed protocols' features will be changed, such as NDP, MLD, ICMPv6, DHCPv6, etc. We tested the BRAS equipments' new features particularly. At the same time, we also did some experiments on L2 or L4 protocols, such as TCP, UDP, PPP, to verify whether they can work well in IPv6 environment.

Testing conclusions show that all the equipments made by four manufacturers can support the basic IPv6, ICMPv6, TCP, UDP, PPP, and IPv6 routing protocols, but just two devices can pass most of the access functions test cases.

The most serious problem is that two DUTs do not support DHCP server functions in IPv6, which causes a negative impact when the BRAS distributes IPv6 addresses to the subscribers or even does DHCP-PD, because there are not local IPv6 address pool in the devices. ISPs must purchase IPv6 DHCP servers additionally.

The second trouble is about multicast. One DUT does not only support MLD protocols but PIM-SM which has nothing to do with IPv6 or IPv4. Another two cannot duplicate multicast stream for each subscriber neither in PPPoE nor IPE. considering this problem, ISPs can hardly carry out triple-play, such as IPTV, by deploying these devices.

At last, only one DUT does not support local PPPoEv6 authentication.

3.3. Test Cases- DUT in Network Test

In order to achieve the gradual transition from IPv4 to IPv6, and to protect the investment of the deployed devices in the current networks, IPv6 BRAS must have the ability to set L2TP tunnel with IPv4 BRAS, and should support IPv6 over IPv4 tunnel with IPv6 router. We did the experiment in topology 2 and 3 in our lab.

One of the problems is about L2TP tunnel. In scenario 2, IPv6 client initiates a PPPoE access request to the IPv4 BRAS, which sets up an L2TP tunnel with the DUT (IPv6 BRAS) after it get the client's IPv6 attribute from the Radius server. But unfortunately, the IPv6 client can not get an IPv6 address or prefix. This situation has taken place on three manufacturers' DUTs because of their IPv6CP function.

BRAS must record subscribers' online time and traffic information, which is a basic function in IPv4. We test that in topology 2, all the DUTs can make a record, but three of them cannot distinguish IPv6 and IPv4 for a dual-stack client. That means ISPs could not charge

separately according to the traffic is IPv4 or IPv6, which is very important for developing IPv6 if ISPs want to charge IPv6 traffic cheaper than IPv4 in the early of the IPv4-IPv6 transition years.

3.4. Test Cases- Performance in IPv6 environment

We also tested the BRAS capability in IPv6 environment in topology 4. The test cases included FIB capacity, ACL quantity, QoS queue quantity, and line card IPv6 throughput.

Performance test conclusions demonstrate that the DUTs' capability do not drop so much in IPv6 networks than in IPv4. The main problems are listed below.

One DUT's IPv6 FIB capacity is only 10% of its IPv4 FIB capacity because the related resources, such as Memory, are designed separately but not shared. But we do not consider that is a serious problem, because the design will certainly be changed if IPv6 traffic increasing over IPv4 traffic in current networks.

One DUT's line card throughput is much less than the nominal value when the packet length shorter than 128B no matter it is IPv4 or IPv6.

3.5. Summary for BRAS Testing

The specific testing results show that only one DUT's IPv6 functions can basically meet the requirement of current networks experiment or commercial trial. The IPv6 device industry need to be pay more attention by all the ISPs and manufacturers.

4. IANA Considerations

This document has no IANA actions.

5. Security Considerations

TBD

6. Normative References

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.

- [RFC2740] Coltun, R., Ferguson, D., and J. Moy, "OSPF for IPv6", RFC 2740, December 1999.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, September 2006.
- [RFC4789] Schoenwaelder, J. and T. Jeffree, "Simple Network Management Protocol (SNMP) over IEEE 802 Networks", RFC 4789, November 2006.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-IS)", RFC 5120, February 2008.
- [RFC5308] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, October 2008.

Authors' Addresses

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

Email: chengang@chinamobile.com

Tianle Yang
China Mobile
53A,Xibianmennei Ave.,
Xuanwu District,
Beijing 100053
China

Email: yangtianle@chinamobile.com

Lianyuan Li
China Mobile
53A,Xibianmennei Ave.
Beijing 100053
P.R.China

Phone: +86-13910750201
Email: lilianyuan@chinamobile.com

Hui Deng
China Mobile
53A,Xibianmennei Ave.
Beijing 100053
P.R.China

Phone: +86-13910750201
Email: denghui02@gmail.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 12, 2012

G. Chen
H. Deng
China Mobile
July 11, 2011

NAT64-CPE Mode Operation for Opening Residential Service
draft-chen-v6ops-nat64-cpe-02

Abstract

The document has summarized NAT64 usages on different modes, in which NAT64-CGN would serve for a large-scale network and NAT64-CPE would give 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 January 12, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	3
2. NAT64-CGN Mode	3
2.1. Usage Overviews	3
2.2. NAT64-CGN Mode Operations	4
2.3. NAT64-CGN Mode Requirements	5
3. NAT64-CPE Mode	6
3.1. Usage Overviews	6
3.2. NAT64-CPE Mode Operations	7
3.3. NAT64-CPE Mode Requirements	7
4. Security Considerations	8
5. IANA Considerations	8
6. Normative References	8
Authors' Addresses	9

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 upgradations 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 either CGN side, or CPE side. NAT64-CPE would give residential service opportunities to be accessed by remote subscribers going through IPv6 networks. In this usage, the NAT64-CPE may not need cooperate with DNS64[RFC6147] any more, whereby this mechanism allows an IPv6-only client (i.e. either a host with only IPv6 stack, or a host with both IPv4 and IPv6 stack, but only with IPv6 connectivity or a host running an IPv6 only application) to initiate communications to an IPv4-only residential service server.

The document has summarized NAT64 usages on different modes. Considering the existing deployment approaches, the memo has proposed different operational requirements for each particular NAT64-mode.

2. NAT64-CGN Mode

2.1. Usage Overviews

Figure 1 illustrates NAT64-CGN mode usage where an IPv6-only host would like to initiate communications with an IPv4-only server through NAT64 deployed at an edge of carried IPv4/IPv6 networks. The NAT64 mechanism allow IPv6-only host to access IPv4 resources and IPv6 resources simultaneously.

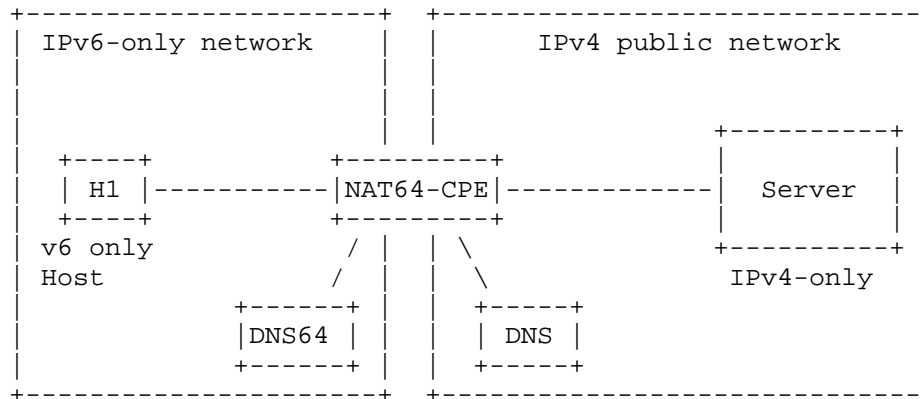


Figure 1: NAT64-CGN Mode Usage

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.

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. NAT64-CGN Mode Operations

The following has described the operational process of NAT64-CGN.

- o Step1: IPv6-only host performs an AAAA DNS query to DNS64 for the IPv6 address of the Pv4-only sever.
- o 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.
- o Step3: DNS4 return the A record to the DNS64.
- o 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.

- o 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-CPE Mode

3.1. Usage Overviews

Residential servers are usually going beyond the operator's management. They may not be 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.

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-CPE numbered with public IPv6 addresses and private IPv4 addresses. DNS server is located in dual stack Internet for naming-resolving.

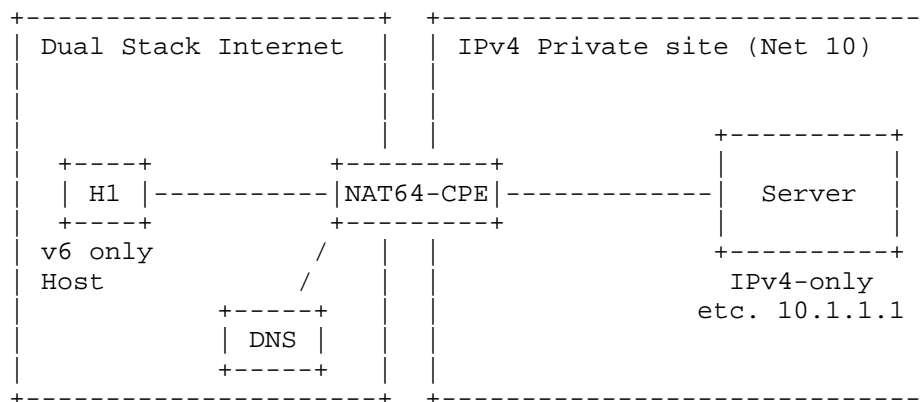


Figure 2: NAT64-CPE Mode Usage

This scenario appears in ISP network quite popular. 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.

3.2. NAT64-CPE Mode Operations

The operational process of NAT64-CPE operation involves CPE, DNS and addressing mechanism. This section illustrates different parts of functionalities.

For NAT64-CPE operations, two kinds of functions would take on. First, it will perform the functionalities that normal CPE does except NAT44 forwarding, like assigning private IPv4 address to their attached residential servers. Additionally, CPE will allocate private IPv4 address to the servers depending on the server MAC address. Therefore, the server could always get constant private IPv4 address. Second, CPE should carry NAT64 functionalities without integrating DNS64. According to normative handling, NAT64-CPE translates in-coming IPv6 destination address by stripping NAT64 IPv6-prefix and maintains a IPv4 pool for translating IPv6 sources address. Therein, the NAT64 IPv6 prefix will be NSP specified in IPv6 Addressing of IPv4/IPv6 Translators. And, ISP will reserve distinct NSP for each CPE. In order to maintain address mapping between inner IP address and outer IP address, PCP [PCP] could be adopted.

For DNS configuration, each residential services should be represented by FQDN format so as to users could easily remember and understand. The corresponding naming resource record should be stored as AAAA. The record's IPv6 address is synthesized by NAT64 prefix and private IPv4 address. The IPv6 format is compliant with assembling IPv6 address in DNS64. The deployed DNS just follow regular DNS handling. There is no demands for performing DNS64 process.

3.3. NAT64-CPE Mode Requirements

The following lists requirements for NAT64-CPE operations.

NAT64-CPE-R1: NAT64-CPE SHOULD be decoupled with DNS64.

NAT64-CPE-R2: NAT64-CPE SHOULD be capable of port forwarding. PCP, UPnP and NAT-PMP are RECOMMENDED to be deployed along with NAT64-CPE.

NAT64-CPE-R3: Network-Specific Prefix SHOULD be assigned to each NAT64-CPE for constructing ipv4-converted ipv6 addresses.

NAT64-CPE-R4: Each residential server SHOULD be represented by FQDN and stored as AAAA record along with ipv4-converted ipv6 address in DNS server. Dynamic updates in the domain name system MAY be applicable to reflect address renumbering of residential servers.

NAT64-CPE-R5: NAT64-CPE SHOULD follow WAN side configuration described in RFC6204[RFC6204].

NAT64-CPE-R6: When a residential server is attached to the LAN interface, NAT64-CPE SHOULD allocate private IPv4 addresses to the server depending on the server MAC address.

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

- [PCP] Wing, D., "Pinhole Control Protocol (PCP)", draft-ietf-pcp-base-13.txt (work in progress), July 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.
- [RFC6204] Singh, H., Beebee, W., Donley, C., Stark, B., and O. Troan, "Basic Requirements for IPv6 Customer Edge Routers", RFC 6204, April 2011.

Authors' Addresses

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

Email: chengang@chinamobile.com

Hui Deng
China Mobile
53A,Xibianmennei Ave.
Beijing 100053
P.R.China

Phone: +86-13910750201
Email: denghui02@gmail.com

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

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	3
2. NAT64-CGN Deployment	3
2.1. Deployment in IDC	3
2.2. Connecting with IPv4 Internet	4
2.3. NAT64-CGN Mode Requirements	5
3. NAT64-CE Mode	6
3.1. NAT64 at Enterprise Network Edge	6
3.2. NAT64 at Residential Network Edge	7
4. Security Considerations	7
5. IANA Considerations	7
6. Normative References	8
Author's Address	8

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 upgradations 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.

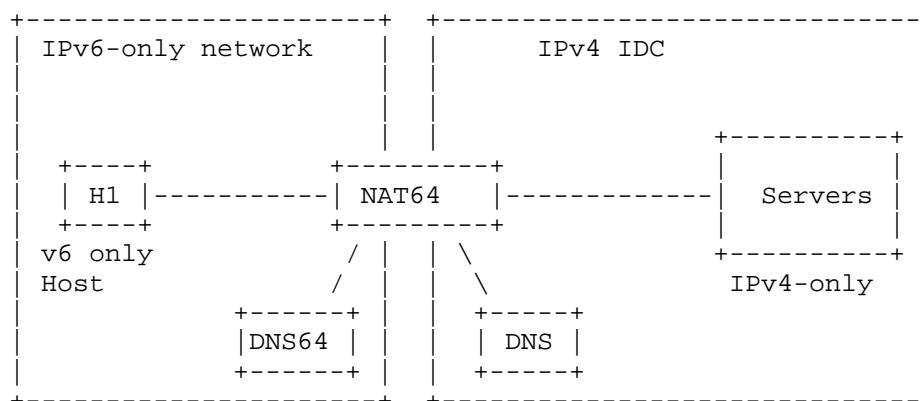


Figure 1: NAT64-CGN Mode Usage

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.

- o Step1: IPv6-only host performs an AAAA DNS query to DNS64 for the IPv6 address of the Pv4-only sever.
- o 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.
- o Step3: DNS4 return the A record to the DNS64.
- o 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.
- o 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.

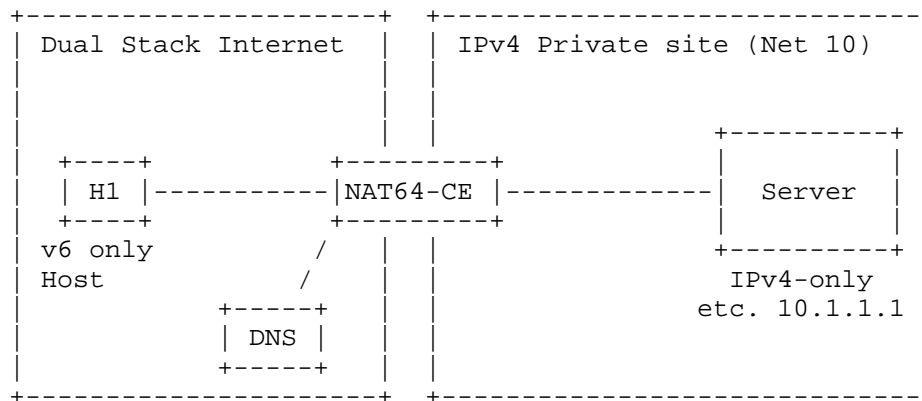


Figure 2: NAT64-CPE Mode Usage

3.2. NAT64 at Residential Network Edge

Residential servers are usually going beyond the operator's management. They may not be 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 appear 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 come back 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

IPv6 Operations
Internet-Draft
Intended status: Informational
Expires: January 13, 2012

T. Chown
University of Southampton
July 12, 2011

IPv6 Address Accountability Considerations
draft-chown-v6ops-address-accountability-01

Abstract

Hosts in IPv4 networks typically acquire addresses by use of DHCP, and retain that address and only that address while the DHCP lease remains valid. In IPv6 networks, hosts may use DHCPv6, but may instead autoconfigure their own global addresses, and potentially use many privacy addresses over time. This behaviour places an additional burden on network operators who require address accountability for their users and devices. There has been some discussion of this issue on various mail lists; this text attempts to capture the issues to encourage further discussion.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 13, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Accountability Approaches	3
2.1. Switch-router polling	3
2.2. Record all ND traffic	4
2.3. Force use of DHCPv6 only	4
2.4. Use SAVI mechanisms	4
3. Privacy Considerations	4
4. Conclusions	5
5. Security Considerations	5
6. IANA Considerations	5
7. Acknowledgments	5
8. Informative References	6
Author's Address	6

1. Introduction

Administrators of IPv4 networks are used to an address accountability model where devices acquire a single global address using DHCP and then use that address while the DHCP lease is valid. The model allows an administrator to track back an IP address to a user or device, in the event of some incident or fault requiring investigation. While by no means foolproof, this model, which may include use of DHCP option 82, is one that IPv4 network administrators are generally comfortable with.

There are many reasons why address stability is desirable, e.g. DNS mappings, ACLs using IP addresses, and logging. However, such stability may not typically exist in IPv6 client networks, particularly where clients are user managed.

In IPv6 networks, where hosts may use SLAAC [RFC4862] and Privacy Addresses [RFC4941], it is quite possible that a host may use multiple IPv6 addresses over time, possibly changing addresses used frequently, or using multiple addresses concurrently. Where privacy addresses are used, a host may choose to generate and start using a new privacy address at any time, and will also typically generate a new privacy address after rebooting. Clients may use different IPv6 addresses per application, while servers may have multiple addresses configured, one per service offered.

It is also worth noting that in an IPv4 network, it is more difficult for a user to pick and use an address manually without clashing with an existing device on the network, while in IPv6 networks, picking an unused address is simple to do without an address clash. Thus picking an unused IP address becomes as simple as picking an unused Layer 2 address. Continuing that comparison, some virtualised OSes may pick randomly generated link layer addresses, and may change these upon virtual host reboot.

2. Accountability Approaches

There are various approaches to address accountability, which have different costs, benefits and trade-offs.

2.1. Switch-router polling

By polling network switch and router devices for IPv4 ARP tables and IPv6 ND tables, and correlating the results with switch port MAC tables, it should be possible to determine which IP addresses are in use at any specific point in time and which addresses are being used on which switch ports (and thus users or devices).

This is the approach adopted by tools such as NAV and Netdot, but there is some concern expressed at the load that may be placed on devices by frequent SNMP or other polling. The polling frequency needs to be rapid enough to ensure that cached ND/ARP data on devices is not expired between polling intervals, i.e. the ND/ARP data should not be expired more frequently than the device is polled.

2.2. Record all ND traffic

If all ND traffic observed on a link can be captured, it should be possible for IPv6 address usage to be recorded. This would require appropriate capability on a device on any given subnet, e.g. as is currently achieved for RAmound or NDPmon, or a reporting mechanism for the subnet router. There may also be mechanisms such as a (filtered) RSPAN that may be suitable; at least one implementation of this has been published.

A benefit of this approach is that collecting all ND traffic would allow additional accounting and fault detection to be undertaken, e.g. rogue RA detection, or DAD DoS detection.

2.3. Force use of DHCPv6 only

One approach to accountability is to attempt to force devices to only use DHCPv6, which would in principle give the same address accountability model as exists for IPv4 today. [RFC4649] for DHCPv6 appears to give at least some of the functionality of DHCP option 82.

While it is possible to craft IPv6 Router Advertisements that give 'hints' to hosts that DHCPv6 should be used ('M' bit set), there is no obligation on the host to honour that hint. However, if the Autonomous (A) flag in the Prefix Information option is unset (as discussed in section 5.5.3 of RFC 4862), the Prefix Information option should be ignored. A user running the device will need to determine the on-link prefix if they wish to manually configure their own address.

2.4. Use SAVI mechanisms

Discussion of appropriateness of SAVI mechanisms to be added here. (In principle, SAVI mechanisms work by observing NDP and DHCP messages, allowing bindings to be set up and recorded.)

3. Privacy Considerations

This draft discusses mechanisms for a site or organisation to manage address accountability where IPv6 has been deployed. In most

networks there is a requirement to be able to identify which users have been using which addresses or devices at a given point in time. This draft was written in response to requests for improved accountability for IPv6 traffic in (mainly) UK academic sites, but the same rationale is likely to apply elsewhere.

While the sources of data that may be used for such purposes (e.g. state on routers or switches) is generally not available to general users of the network, it is available to administrators of the network. The use of privacy mechanisms, e.g. RFC 4941, gives the greatest benefit when the addresses are being observed by external third parties.

4. Conclusions

This text is an initial draft attempting to capture the issues related to IPv6 address accountability models. If an all-DHCPv6 model is not viable, IPv6 network administrators will need to deploy management and monitoring tools to allow them to account for hosts that will have multiple IPv6 addresses that may also change rapidly over time.

Some of the approaches described do not depend on a specific type of address management being used, and will thus work with other addressing methods if they emerge in the future.

Feedback on the issues discussed here is welcomed.

5. Security Considerations

There are no extra security consideration for this document.

6. IANA Considerations

There are no extra IANA consideration for this document.

7. Acknowledgments

The author would like to thank the following people for comments on this text: Mark Smith, and James Woodyatt.

8. Informative References

- [RFC4649] Volz, B., "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Relay Agent Remote-ID Option", RFC 4649, August 2006.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC4941] Narten, T., Draves, R., and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6", RFC 4941, September 2007.

Author's Address

Tim Chown
University of Southampton
Highfield
Southampton, Hampshire SO17 1BJ
United Kingdom

Email: tjc@ecs.soton.ac.uk

IPv6 Operations
Internet-Draft
Intended status: Informational
Expires: December 9, 2011

T. Chown
University of Southampton
M. Ford
Internet Society
S. Venaas
Cisco Systems
June 7, 2011

World IPv6 Day Call to Arms
draft-chown-v6ops-call-to-arms-03

Abstract

The Internet Society (ISOC) has declared that June 8th 2011 will be World IPv6 Day, on which some major organisations are going to make their content available over IPv6. With the likes of Google and Facebook providing IPv6 access to their production services and domains, it is very likely we will see more IPv6 traffic flowing across the Internet than has ever been seen before. With this in mind, it seems timely to issue a call to arms for systems and network administrators to review their organisation's IPv6 capabilities in order to mitigate common causes of IPv6 connectivity problems in advance of the day. The increased traffic on World IPv6 Day should also create an excellent opportunity to observe the behaviour and performance of IPv6; it is thus very desirable to have appropriate measurement tools in place in advance. We discuss some appropriate tools from the network and application perspective.

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 December 9, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
2. Connectivity Issues	4
2.1. Unmanaged Tunnels	4
2.2. Tunnel Broker first-hop delays	5
2.3. Connection Timeouts	5
2.4. PMTU Discovery	7
2.5. Rogue Router Advertisements	7
2.6. Tunnel performance	8
2.7. AAAA record advertised but service not enabled	8
2.8. IPv6 Reverse DNS	9
3. Instrumentation	9
3.1. IPv6 traffic levels	9
3.2. Network flow records	10
3.3. Client Web Access Success Rate	10
3.4. Tools to measure IPv6 brokenness	10
3.5. IPv4 Performance Comparison	11
3.6. User Tickets	11
3.7. Security monitoring	11
4. IPv6-only testing	11
5. Conclusions	11
6. Security Considerations	12
7. IANA Considerations	12
8. Acknowledgments	12
9. Informative References	12
Authors' Addresses	15

1. Introduction

Despite the recent exhaustion of the available IPv4 address pool, deployment of IPv6 remains limited. To help encourage organisations to trial production deployment, ISOC has declared June 8th 2011 as World IPv6 Day [ISOC]. Organisations are encouraged to use this day to test IPv6 in production by making their main, externally-facing websites available over IPv6. Sites planning to turn on IPv6 for access in their network in the interest of World IPv6 Day should ensure this is completed well before the day, and commit to leaving it active after the event, and thus using the method they would choose to do so indefinitely. At the current time, this would generally mean enabling dual-stack networking with IPv4 running alongside IPv6. However, IPv6-only networks are ultimately inevitable, and so some sites may choose to use June 8th to undertake some focused tests on that deployment model.

The purpose of this document is two-fold. One is to discuss common IPv6 connectivity issues that are likely to arise on June 8th, with a focus on dual-stack networking (which is likely to be how the vast majority of sites take part). Most of the issues discussed in this text are those that would affect an end site or enterprise network running IPv6, but may be applicable elsewhere. Highlighting the issues should help raise awareness of those problems and possible mitigations. The other purpose is to encourage organisations to think about how they might get useful instrumentation in place to observe what happens in and to/from their networks on the day, both from the network and application perspective. Such measurement tools are likely to be useful in the longer term, so once deployed they could be left in place beyond June 8th.

For sites providing content, June 8th will be a chance to make some public facing services available over IPv6, most likely web content using their production domain (e.g. www.example.com) rather than a contrived IPv6 test domain (e.g. www.ipv6.example.com). Enabling public-facing Internet services is a reasonable first step for any organisation deploying IPv6. For ISPs, supporting IPv6 for their Internet-facing services (web, mail, etc.) and recording the impact of World IPv6 Day on their IPv4-only customers is an appropriate action. For sites enabling clients, doing so initially in their IT department may be appropriate; for educational sites enabling IPv6 on eduroam wireless networks could be appropriate given the underlying 802.1x authentication technology is IP version independent.

It should be emphasised that while World IPv6 Day is in many senses an 'experiment' or 'test flight' for IPv6, organisations should strongly consider deploying IPv6 in exactly the same robust way that they would do if they were deploying IPv6 and leaving it enabled

indefinitely. Similarly, applying measures to improve IPv6 robustness, e.g. improved ICMPv6 filtering practice, should be considered long term benefits. That they 'affect' the experiment is not a problem; indeed all measures that improve the robustness of IPv6 deployment should be seen as worthwhile. There will still be problems found, but these can at least be recognised and work done to make them better.

The document also includes a brief section on tools that might be used to test IPv6-only operation.

The scope of this document is purely informational to provoke discussion.

2. Connectivity Issues

In this section we review some common causes of IPv6 connectivity issues, oriented towards those that end sites or enterprises may have some ability to influence or mitigate. Some issues, such as transit arrangements, are not included - currently the focus is on end sites (or users) who may take part in the World IPv6 Day. Some IPv6 connectivity test sites are emerging, for example [testipv6]. There is no significance to the order in which issues are listed.

2.1. Unmanaged Tunnels

One cause of connectivity problems is the use of unmanaged tunnels, in particular 'automated' methods that are not provisioned by the user's ISP. The most common example is 6to4 [RFC3056], or more specifically the 6to4 relay approach described in [RFC3068]. A native IPv6 host communicating with a 6to4 host will require both hosts to have access to an appropriately capable 6to4 relay (which may or may not be the same relay). If a host in a native IPv6 network has no route to 2002::/16 it cannot send traffic to a 6to4 host. Similarly, a 6to4 router that cannot reach the well-known IPv4 anycast relay address cannot send traffic to a native IPv6 network. There are also potential issues with Protocol 41 filtering at site borders close to the client.

A presentation by Geoff Huston at IETF80 [Huston2011] highlighted the connection failure rates with 6to4, measured in excess of 15%, as well as the additional latency in 6to4 communications, with 6to4 showing an average additional 1.2s latency per retrieval.

One approach to this problem is to encourage sites/ISPs to run local relays, as discussed in [I-D.carpenter-v6ops-6to4-teredo-advisory]. This draft discusses how to make 6to4 more robust in situations where

there is a conscious decision to use it. Sites using 6to4 should consider deploying local relays to increase the chance of a good IPv6 experience. The alternative to reduce such problems is simply to move 6to4 to Historic, as proposed in [I-D.troan-v6ops-6to4-to-historic]. This would mean 6to4 would not be enabled by default anywhere, and once its usage had reduced enough, relays could be turned off.

There may still be some CPE routers that do enable 6to4 by default; it is likely that devices behind such routers will experience problems on World IPv6 Day.

Connection failures and latency with the Teredo protocol [RFC4380] were also highlighted by Geoff Huston's IETF80 presentation. Teredo connection failure rates were as high as 35%, with 1-3s additional latency. One of the connection issues is reliance on the ICMPv6 probe packet being able to reach the destination host; in practice filters may block these. Thus Teredo should not be considered a reliable means of accessing the IPv6 Internet.

2.2. Tunnel Broker first-hop delays

IPv6 tunnel brokers, such as those provided by SixXS (<http://www.sixxs.net>) and Hurricane Electric (<http://tunnelbroker.net>) provide a more robust, managed approach to IPv6-in-IPv4 tunnelled access than 6to4. Individual users interested in IPv6 access for World IPv6 Day, in the absence of IPv6 support from their ISP, should consider registering to use a free tunnel broker. It would be sensible to register for and test your broker client well in advance of IPv6 Day, and ideally plan to keep it available beyond that date, until your ISP provides IPv6 natively for you. One set of test sites to use would be the list cited on the ISOC World IPv6 Day site [ISOCsites].

When choosing a broker service, it is prudent to pick one with a presence near to you that has a minimal round trip time. Providers such as SixXS and HE have tunnel broker servers in many countries. Beware picking a broker in another continent that may add 150ms+ to your round trip times.

2.3. Connection Timeouts

One of the main drivers for IPv6 Day is identifying and fixing the problems that can lead to connection timeouts. Because unreliable IPv6 connectivity leads to intensely frustrating problems for end-users, it is essential that people motivated to deploy IPv6 connectivity, whether for themselves, or for a larger network, only do so in a well-supported, production-quality fashion.

Where dual-stack systems - or rather the applications running on them - have a choice of IPv4 or IPv6 connectivity, timeouts can occur if there is no connectivity on the preferred protocol. For example, if both A and AAAA DNS records exist for a web server, and IPv6 connectivity is broken, there is likely to be some timeout for the browser before the connection drops back to IPv4.

A bigger problem exists if the application or OS tries IPv6 first and then does not fall back to IPv4. A bug in versions of Opera prior to 10.5 caused such behaviour, which was obviously a big issue for Opera users trying to access dual-stack web sites with broken IPv6 connectivity.

The author has undertaken some informal tests at his own site, which shows how different combinations of browsers and operating systems behave in the event of IPv6 connections failing or when ICMP unreachables are received. On Linux/Firefox, web connections timeout after 20 seconds for 'no response', but immediately for unreachables. In contrast, Windows Vista/IE was 20 seconds regardless of unreachables being received. Any non-trivial delay will cause significant user frustration.

A more complete set of tests was run by Teemu Savolainen and reported at IETF80 [Savolainen2011]. Although the tests were only samples, they confirmed the results, also showing experiences across a much broader range of platforms, and that the problems with Vista/IE are repeated with Win 7/IE. It's thus clear that if major content providers enable IPv6 on World IPv6 Day, and end users for some reason try to access the content with broken IPv6 connectivity, they are likely to experience significant timeout issues.

This problem is probably the main reason that Google implemented a AAAA whitelisting system for its test sites. The sites had to demonstrate they had good IPv6 connectivity before being allowed into the test programme. The topic is discussed in [I-D.ietf-v6ops-v6-aaaa-whitelisting-implications]. For the sake of World IPv6 Day, it is expected that no such whitelisting is in place - that is, after all, the point of having a day dedicated to testing IPv6 in production.

An interesting suggestion to handle the problem is the 'happy eyeballs' approach described in [I-D.ietf-v6ops-happy-eyeballs]. This approach is now also being suggested for multiple interface systems, as per [I-D.chen-mif-happy-eyeballs-extension]. The happy eyeballs philosophy is to try both IPv4 and IPv6 together, and keep the first working connection up, remembering the result for future connection attempts. It may prefer IPv6 slightly in initial connections rather than trying connections exactly simultaneously.

It is an interesting approach, though some people are concerned about the additional connection load, or that this 'workaround' is simply masking underlying problems that should be fixed.

2.4. PMTU Discovery

IPv6 mandates that fragmentation is only undertaken by the sending node, and thus IPv6 requires working PMTU Discovery [RFC1981]. An existing RFC gives Recommendations for Filtering ICMPv6 Messages in Firewalls [RFC4890]; if this guidance is not followed, connectivity problems are likely to arise. Blindly filtering all ICMPv6 messages is not good practise. Filtering ICMP is a common practice in some IPv4 networks today. Adopting the same approach to ICMPv6 when deploying IPv6 networks will cause connectivity issues for users of the network filtering ICMPv6 and hosts trying to reach the filtered network. RFC 4890 is therefore an important document for IPv6 deployment engineers to read and it is similarly important to verify that IPv6 firewall deployments support appropriate configurations for ICMPv6 filtering.

The minimum MTU for IPv6 is 1280 bytes. Checking the MTU is an important step when connectivity issues arise. Where PMTUD is not working or not implemented, the using the minimum MTU is likely to resolve the problem, though not give optimal performance (the cause should still be investigated and resolved for longer term benefit). Tunnel broker services such as SixXS and HE set their MTUs to default to 1280, probably due to the varying conditions their customers may be in. However, it is preferable for enterprise networks to configure appropriate ICMPv6 filtering to allow PMTUD to operate and establish the most efficient MTUs for a link.

2.5. Rogue Router Advertisements

Within a site, hosts may use IPv6 Stateless Address Autoconfiguration (SLAAC) [RFC4862]. However, it is possible for accidental (or malicious) rogue RAs to cause connectivity issues, as described in the Rogue Router Advertisement Problem Statement [RFC6104].

A typical cause of rogue RAs is Windows ICS, which can present a rogue 6to4 router on its wireless interface. This will cause hosts to potentially autoconfigure two global IPv6 addresses and pick the wrong default router, with unpredictable results. As a (bad) example the author experienced a scenario where he had a rogue 6to4 RA, but because the rogue 6to4 was working he was able to access IPv6 networks outside his own network, but could not access most internal hosts inside his own network because he was unwittingly using 6to4 from outside into his own network, and thus being firewalled from those internal hosts.

In many cases, default address selection [RFC3484] (and [I-D.ietf-6man-rfc3484-revise]) would avoid such cases, because the address selection rules should prefer, or can be configured to prefer, native IPv6 over 6to4. However not all operating systems implement RFC 3484 yet, in particular MacOS X (though support may be appearing in Lion). Where rogue RAs cause broken IPv6 behaviour, the timeout issues discussed above may apply.

Adding ACLs to your switches to block ICMPv6 Type 134 packets on ports that do not have routers connected would also minimise the impact of rogue RAs. A more elegant solution is RA Guard [RFC6105], and another is use of SEcure Neighbour Discovery (SEND) [RFC3971]. However neither is widely implemented yet. Indeed, any reported operational experience of SEND in an enterprise network would be very welcome.

Finally, there is a tool called RAMond, available freely from <http://ramond.sourceforge.net>, that can be configured to detect and issue deprecating RAs against observed rogue RAs. This software is based on rafixd.

2.6. Tunnel performance

In scenarios where sites currently have manually configured tunnels to gain IPv6 connectivity, it may be the case that such encapsulation is performed by a router's CPU, in which case unexpected high volumes of traffic may cause problems. Bear in mind that on World IPv6 Day, you may start using IPv6 by default for some high bandwidth applications that you had not used before, e.g. YouTube from Google. It may be prudent to estimate your load for such applications in advance, and test the capability of your tunnelling solution to handle that load.

2.7. AAAA record advertised but service not enabled

If enabling a service for World IPv6 Day, be aware of other existing services that may be running on the same system. If a server has multiple functions, all services should be IPv6 enabled before a AAAA record is entered into the DNS for services that may use that name.

A related consideration is to make sure that firewalls don't just drop IPv6 packets to ports that are not in use. It's better if the firewall or host sends an unreachable indication or a TCP RST to avoid a potential timeout. For example, if you add a AAAA record for your web server that also runs say FTP, where FTP is IPv4 only, either the firewall should have port 21 open or the firewall should be configured to send a TCP RST. There are of course tradeoffs in enabling ICMP unreachables.

2.8. IPv6 Reverse DNS

Presence of IPv6 reverse DNS records is used by many systems as a security method. For example, many mail exchangers will only accept SMTP connections from IP addresses with a reverse DNS entry. It is thus important for such records to exist where, for example, a site is sending mail out over IPv6 transport. It is not necessarily the case that such connections will fall back to IPv4 if reverse records are not present.

3. Instrumentation

In this section we discuss potential instrumentation approaches that may be configured in advance of World IPv6 Day, and then retained longer term after the event. These are particularly useful if your site is turning on AAAA records for its production web presence (for example) and wants to get the best insight into how the systems performed and the nature of the end user experience.

These measurements should complement informal, subjective reports from users at participating sites. It is probably prudent to make at least your organisation's IT staff aware of the 'at risk' day, and actions they should take should they experience problems. It may also be desirable to undertake some form of user survey soon afterwards; whether you inform general users in advance is an issue for each site. The ARIN IPv6 wiki is a good source of such advice [ARINwiki].

3.1. IPv6 traffic levels

It should be possible to measure raw IPv6 traffic levels independently on dual-stack switch/router platforms, given implementations of appropriate MIBs. Sites should take steps to ensure they have the tools in place to be able to view the relative levels of IPv4 and IPv6 traffic over time.

Application level measurement is also desirable, because handling of choice (preference) of protocol used lies with the application if both A and AAAA records are returned. Sites should be aware that due to IPv6 Privacy Extensions [RFC4941] application logs may show more apparent different clients connecting, due to clients cycling the source IPv6 address they use over time.

The types of information gathered might for example include:

- o IPv6 traffic volume, sources of IPv6 traffic by AS, types of IPv6 traffic (e.g. native, 6to4, Teredo, tunnelled);

- o IPv6 application mix, comparison with IPv4;
- o The number and type of IPv6 client connections.

3.2. Network flow records

Where available, sites should seek to generate and record network flow records for traffic, to maximise opportunities to analyse traffic patterns after the event, or in the case of reports of specific problems. Netflow v9 supports IPv6. Open source IPv6-capable Netflow collectors also exist, e.g. nfsen, from <http://nfsen.sourceforge.net>.

3.3. Client Web Access Success Rate

There have been some recent studies on the capabilities of web clients to access content on dual-stack servers by IPv4 or IPv6 in the presence of both A and AAAA records existing for a web domain.

One good example is that of [Anderson10], as reported at RIPE-61, where the author set up some application (web server) oriented tests for his newspaper content in Norway. The methodology was to add an invisible IFRAME to his site that would include IMG links randomly to 1x1 images that were served either via an IPv4-only target or a dual-stack target. Variation in the hit rates would imply IPv6 brokenness. By analysing the http metadata information could be gleaned on the cause of the brokenness. Results in Q4'2009 showed 0.2-0.3% brokenness, including the Opera bug mentioned above.

Recent figures published by Google suggest at most a 0.1% level of brokenness, indicating some improvement, but that level is still potentially 1 in 1000 users with a problem.

3.4. Tools to measure IPv6 brokenness

Sites may wish to make their own measurements of IPv6 brokenness rather than relying on third party reports. There are some openly available tools available that work along similar principles to the method proposed by Tore Anderson above.

The APNIC Labs test tool uses a combination of JavaScript and Google Analytics to measure various types of brokenness [APNIC]. Eric Vyncke's tool [Vyncke] measures a slightly smaller set of types of brokenness, but also looks very useful, with additional reports on the browser type for each failure. The author is currently using the latter tool, and plans to enable the APNIC measurement system shortly when other Analytics updates are applied locally.

3.5. IPv4 Performance Comparison

Where a dual-stack service is deployed, measuring the relative performance of both protocols is desirable. This may primarily be a measurement of throughput or delay, but may also include availability/uptime measurement. A site may choose to set up its own performance measuring framework, for example using open source bandwidth and throughput test tools. Participants in World IPv6 Day will be monitored from a broad range of locations and measurements will be available to show availability of AAAA records, reachability to http service, latency and availability over time.

3.6. User Tickets

It is possible a higher than usual user ticket rate for connectivity issues may be experienced. being able to categorise these cases for subsequent analysis is desirable.

3.7. Security monitoring

We mentioned RAMond above in the context of watching for rogue RAs. There is another useful package called NDPmon, also available freely from <http://ndpmon.sourceforge.net>, that can be configured to watch for certain types of IPv6 'abuse' on your local network. It may be interesting to run the tool to confirm whether any 'bad' traffic is observed within your network on World IPv6 Day.

4. IPv6-only testing

The long-term IPv6 deployment plan is IPv6-only networking, rather than dual-stack. It is not clear how quickly significant IPv6-only networks will emerge, but testing of approaches to IPv6-only operation is desirable as soon as possible. A draft by Jari Arkko and Ari Keranen describes some such experiences [I-D.arkko-ipv6-only-experience].

Some experience of NAT64 [RFC6146] has been described in [I-D.tan-v6ops-nat64-experiences], though this appears to have used only NAT-PT so far. An implementation of NAT64 is available at <http://ecdysis.viagenie.ca>. Operational experience of IVI is also desirable. An implementation of IVI is available at <http://www.ivi2.org/IVI>.

5. Conclusions

With the ISOC World IPv6 Day event due on June 8th 2011, this

document aims to help focus attention on both improving awareness and mitigations of common causes of IPv6 connectivity problems, and encouraging sites and organisations to introduce appropriate instrumentation into their networks so they can observe traffic behaviour appropriately.

This is still an early version of the text, and is thus a little drafty. All comments are very welcome towards a mature version in advance of June.

6. Security Considerations

There are no extra security consideration for this document.

7. IANA Considerations

There are no extra IANA consideration for this document.

8. Acknowledgments

To be added.

9. Informative References

- [RFC1981] McCann, J., Deering, S., and J. Mogul, "Path MTU Discovery for IP version 6", RFC 1981, August 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.
- [RFC3971] Arkko, J., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.
- [RFC4380] Huitema, C., "Teredo: Tunneling IPv6 over UDP through Network Address Translations (NATs)", RFC 4380, February 2006.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless

Address Autoconfiguration", RFC 4862, September 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.

[RFC6104] Chown, T. and S. Venaas, "Rogue IPv6 Router Advertisement Problem Statement", RFC 6104, February 2011.

[RFC6105] Levy-Abegnoli, E., Van de Velde, G., Popoviciu, C., and J. Mohacsi, "IPv6 Router Advertisement Guard", RFC 6105, February 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.

[I-D.carpenter-v6ops-6to4-teredo-advisory]
Carpenter, B., "Advisory Guidelines for 6to4 Deployment", draft-carpenter-v6ops-6to4-teredo-advisory-03 (work in progress), March 2011.

[I-D.ietf-v6ops-happy-eyeballs]
Wing, D. and A. Yourtchenko, "Happy Eyeballs: Trending Towards Success with Dual-Stack Hosts", draft-ietf-v6ops-happy-eyeballs-02 (work in progress), May 2011.

[I-D.tan-v6ops-nat64-experiences]
Tan, J., Lin, J., and W. Li, "Experience from NAT64 applications", draft-tan-v6ops-nat64-experiences-00 (work in progress), March 2011.

[I-D.troan-v6ops-6to4-to-historic]
Troan, O., "Request to move Connection of IPv6 Domains via IPv4 Clouds (6to4) to Historic status", draft-troan-v6ops-6to4-to-historic-01 (work in progress), March 2011.

[I-D.ietf-v6ops-v6-aaaa-whitelisting-implications]
Livingood, J., "IPv6 AAAA DNS Whitelisting Implications", draft-ietf-v6ops-v6-aaaa-whitelisting-implications-05 (work in progress), May 2011.

[I-D.chen-mif-happy-eyeballs-extension]

Chen, G. and C. Williams, "Happy Eyeballs Extension for Multiple Interfaces",
draft-chen-mif-happy-eyeballs-extension-01 (work in progress), March 2011.

[I-D.ietf-6man-rfc3484-revise]

Matsumoto, A., Kato, J., and T. Fujisaki, "Update to RFC 3484 Default Address Selection for IPv6",
draft-ietf-6man-rfc3484-revise-02 (work in progress),
March 2011.

[I-D.arkko-ipv6-only-experience]

Arkko, J. and A. Keranen, "Experiences from an IPv6-Only Network", draft-arkko-ipv6-only-experience-03 (work in progress), April 2011.

[APNIC] "IPv6 Capability Tracker", <<http://labs.apnic.net/>>.

[Vyncke] Vyncke, E., "Estimation of IPv6 brokenness",
<<http://test4.vyncke.org/testv6/>>.

[ARINwiki]

"ARIN IPv6 Wiki", <http://getipv6.info/index.php/Customer_problems_that_could_occur>.

[testipv6]

"Test IPv6", <<http://www.test-ipv6.com/>>.

[ISOC] "World IPv6 Day", <<http://isoc.org/wp/worldipv6day/>>.

[Huston2011]

Huston, G., "Stacking it Up: Experimental Observations on the operation of Dual Stack Services", 2011,
<<http://www.ietf.org/proceedings/80/slides/v6ops-1.pdf>>.

[Savolainen2011]

Savolainen, T., "Experiences of host behaviour in broken IPv6 networks", 2011,
<<http://www.ietf.org/proceedings/80/slides/v6ops-12.pdf>>.

[ISOCsites]

"IPv6 Enabled Websites",
<<http://www.worldipv6day.org/ipv6-enabled-websites>>.

[Anderson10]

Anderson, T., "Measuring and Combating IPv6 Brokenness", 2010,
<<http://ripe61.ripe.net/presentations/162-ripe61.pdf>>.

Authors' Addresses

Tim Chown
University of Southampton
Highfield
Southampton, Hampshire SO17 1BJ
United Kingdom

Email: tjc@ecs.soton.ac.uk

Mat Ford
Internet Society
Geneva,
Switzerland

Email: ford@isoc.org

Stig Venaas
Cisco Systems
Tasman Drive
San Jose, CA 95134
USA

Email: stig@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 12, 2012

W. Dec
R. Asati
Cisco
C. Congxiao
CERNET Center/Tsinghua
University
H. Deng
China Mobile
July 11, 2011

Stateless 4Via6 Address Sharing
draft-dec-stateless-4v6-02

Abstract

This document presents an overview of the characteristics of stateless 4V6 solutions, alongside a assessment of the issues attributes. The impact of translated or mapped tunnel transport modes is also presented in the broader context of other industry standard reference architectures and existing deployments.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1. Introduction	4
2. Terminology	4
3. Stateless 4V6 Technical and Architectural Overview	5
3.1. IPv4 address and algorithmic port indexing	7
3.2. 4V6 CE IPv6 Address and domain info	7
3.3. IPv6 Adaptation Function	8
3.3.1. 4V6 Mapped Tunnel Mode	8
3.3.2. 4V6 Translation mode	8
4. Comparison of 4V6 transport modes	9
4.1. General Characteristics of 4V6 modes	9
4.2. Mobile SP Architecture and 4V6 Applicability	11
4.2.1. 3GPP overview	12
4.2.2. 3GPP and 4V6 modes	14
4.3. Cable SP Architectures & S46 Applicability	17
4.3.1. PacketCable Introduction	18
4.3.2. PacketCable Construct - Classifier	19
4.3.3. PacketCable MultiMedia & 4V6 Modes	20
5. Overview of potential issues and discussion	21
5.1. Notion of Unicast Address	21
5.1.1. Overview	21
5.1.2. Discussion	21
5.2. Implementation on hosts	22
5.2.1. Overview	22
5.2.2. Discussion	22
5.3. 4V6 address and impact on other IPv6 hosts	22
5.3.1. Overview	22
5.3.2. Discussion	23
5.4. Impact on 4V6 CE based applications	23
5.4.1. Overview	23
5.4.2. Discussion	23
5.5. 4V6 interface	24
5.5.1. Overview	24
5.5.2. Discussion	24

5.6. Non TCP/UDP port based IP protocols - ICMP)	24
5.6.1. Overview	24
5.6.2. Discussion	24
5.7. Provisioning and Operational Systems	25
5.7.1. Overview	25
5.7.2. Discussion	25
5.8. Training & Education	26
5.8.1. Overview	26
5.8.2. Discussion	26
5.9. Security and Port Randomization	27
5.9.1. Overview	27
5.9.2. Discussion	27
5.10. Unknown Failure Modes	28
5.10.1. Overview	28
5.10.2. Discussion	28
5.11. Possible Impact on NAT66 use & design	29
5.11.1. Overview	29
5.11.2. Discussion	29
5.12. Port statistical multiplexing and monetization of port space	29
5.12.1. Overview	29
5.12.2. Discussion	30
5.13. Readdressing	30
5.13.1. Overview	30
5.13.2. Discussion	30
5.14. Ambiguity about communication between devices sharing an IP address.	31
5.14.1. Overview	31
5.14.2. Discussion	31
5.15. Other	32
5.15.1. Abuse Claims	32
5.15.2. Fragmentation and Traffic Asymmetry	32
5.15.3. Multicast Services	33
6. Conclusion	33
7. IANA Considerations	34
8. Security Considerations	34
9. Contributors and Acknowledgements	34
10. References	34
10.1. Normative References	34
10.2. Informative References	34
Authors' Addresses	36

1. Introduction

As network service providers move towards deploying IPv6 and IPv4 dual stack networks, and further on towards IPv6 only networks, a problem arises in terms of supporting residual IPv4 services, over an infrastructure geared for IPv6-only operations, and doing so in the context of IPv4 address depletion. This class of problem is referred to by the draft as the 4via6 problem, for which a stateless solution is desired driven by motivation as documented in [I-D.operators-softwire-stateless-4v6-motivation]. Solutions such as a 4rd [I-D.despres-softwire-4rd], [I-D.murakami-softwire-4v6-translation] and dIVI [I-D.xli-behave-divi] offer such stateless solutions, by using fully distributed NAT44 functionality located on end user CPEs, which allows the network operators' core to remain effectively stateless in terms of NAT44. The solutions, collectively called Stateless4V6, rely on the same IPv4 address being used by multiple CPEs, each with a different TCP/UDP port range, and are derived from the Address+Port (A+P) solution space [I-D.ymbk-aplusp]. Differences between the solutions come down to the mode of transport (translation or mapped tunneling), and the mapping algorithm used. This document looks at the issues that have been claimed as applying to A+P technology, in the specific context of the referenced solutions, and also analyzes the two modes of transport.

2. Terminology

Stateless4V6 domain: A domain is composed out of an arbitrary number of 4V6 CE and Gateway nodes that share a mapping relationship between an operator assigned IPv6 prefix and one or more IPv4 subnets along with all the applicable TCP/UDP ports, all mapped into the IPv6 address space. An 4V6 system can have multiple domains.

Stateless4V6 CE: A CPE node that implements 4V6 functionality including NAT44 which is provisioned by means of 4V6. The device interfaces to the SP network using native IPv6 and provides a NAT44 and IPv4-IPv6 adaptation service to the user.

Stateless4V6 Gateway A Service Provider node that implements the stateless 46 adaptation functionality for interfacing between the SP's IPv6 domain and an IPv4 domain in delivering end user IPv4 connectivity beyond the domain.

IPv4 Address sharing The notion of attributing the same IPv4 address by multiple CEs in an 4V6 domain.

Port-set: A set composed of unique TCP/UDP ports (ranges) associated to a IPv4 address. A single 4V6 CE is expected to have a single port-set for each IPv4 address.

Port-set-id: A numeric identifier of a given port set that is unique in a given 4V6 domain. A port-set-id is used to algorithmically determine the port-set members. The port-set-id is conveyed to CEs as part the CE's IPv6 addressing information, ie it is part of IPv6 subnet or address of a given CE, and its format places no restriction on the use of SLAAC or DHCP addressing.

CE-index: A numeric value, composed of a full or partial IPv4 address and optionally a port-set-id, which uniquely identifies a given CE in an 4V6 domain.

3. Stateless 4V6 Technical and Architectural Overview

This section presents the architectural and technical overview of a stateless 4v6 solution, and evidenced in whole or in part by various stateless 4via6 solution proposals such as 4rd, dIVI. Figure 1 depicts the overall architecture with two IPv4 user networks connected via 4via6 CPEs that share an IPv4 address. The goal of the system is to allow IPv4 user connectivity to the Public IPv4 network, across an operator's IPv6 network.

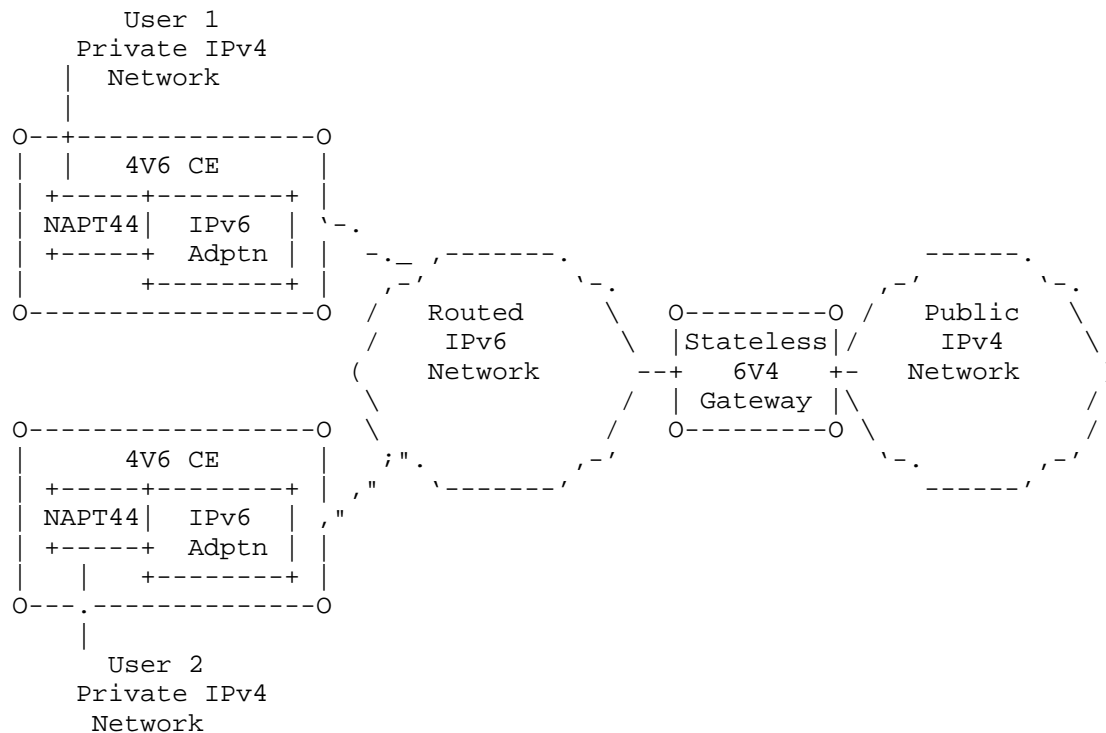


Figure 1 - Generalized Stateless 4V6 system

On IPv4 network user side, the routed IPv6 service provider network is demarcated with a 4V6 CE. The CPE externally has only a native IPv6 interface to the SP network, and a native IPv4 interface towards the end user network.

The IPv4 Internet is demarcated from the operator IPv6 network with one or more operator managed stateless 6V4 gateways that contain an IPv6 adaptation function (not detailed in the diagram) matching the one in the CE. Note: The stateless 6v4 gateway can be integrated into any existing network element (eg a core router, or an IP Edge).

Internally, the 4V6 CE is modelled as having a port restricted NAPT44 function coupled with a stateless IPv6 adaptation function that is able to ferry the end-user's IPv4 traffic across the IPv6 network, besides deriving 4V6 provisioning info from it. The NAPT44 function derives its IPv4 address, which may be shared with that of other users, and its unique Layer 4 (TCP/UDP) port range from the IPv6 address/prefix by means of an 4V6 algorithm and a port indexing schema. No DNS64 or IPv6 aware ALGs, nor IPv4 ALGs are used or assumed. Any IPv4 ALG functionality that the CPE may support, remain unaffected. The CPE is expected to act as a DNS resolver proxy,

using native DNS over IPv6 to the SP network.

The service provider is assumed to be operating all the necessary provisioning and accounting infrastructure to support a regular IPv6 deployment. Similarly, the network operator is assumed to have the ability to assign an IPv6 prefix or IPv6 address to a CPE, and log such an address assignment.

End user host's DO NOT implement any of the 4V6, or other address sharing technologies, nor are they addressed directly with a shared IPv4 address. End user IPv4 hosts connected to the CPE receive unique private addresses assigned by the CPE, and it is the CPE that is directly addressed by the shared IPv4 address.

The IPv6 Adaptation function is not multi homed to the same IPv6 network. The CE itself may be multi homed, but it only has one IPv6 addressed interface for the stateless 4via6 application and IPv6 adaptation function.

Although tangential to the discussion of stateless 4V6, it is useful to note that the CPE is expected to have a native IPv6 interface to the end user network, with any of the end user IPv6 hosts (single or dual stack) receiving IPv6 addresses from an IPv6 delegated prefix issued to the CPE.

3.1. IPv4 address and algorithmic port indexing

At the heart of the 4V6 solution, irrespective of mode of transport, lies the algorithm described in the specific solution drafts that allows the mapping of a shared IPv4 address and a TCP/UDP given port-set to a single IPv6 prefix or address. Notably, the 4V6 system allows both the shared IPv4 address use, as well as full non-shared IPv4 address use, all subject to the 4V6 domain configuration.

The S46 domain information required to compute the IPv4 address and correct port set is retrieved from the 4V6 prefix advertised to the CE, and pre-configured or statelessly acquired domain information.

3.2. 4V6 CE IPv6 Address and domain info

As presented in Section 2, IPv6 address of an 4V6 CE is composed out of the SP advertised IPv6 prefix (containing the CE-index) and an algorithmically computed appendix to complete the 128-bit address. This IPv6 address is *in addition* to any other IPv6 interface address that the CE configures or is configured with, including addresses a SLAAC address from the 4V6 prefix. One characteristics of the resulting IPv6 prefix or address is that it is for all intents and purposes a regular IPv6 prefix address that can be assigned to

any regular IPv6 host.

The IPv6 4V6 interface is reserved for the 4V6 application and the 4V6 IPv6 adaptation function will exclusively use this IPv6 address. This is because the 4V6 system supports stateless communication between the 4V6 CE and the 4V6 gateway only by means of packets sent to/from this address.

3.3. IPv6 Adaptation Function

The IPv6 adaptation function plays a key role in the 4V6 system, in statelessly allowing the IPv4 user payload to be transported across an IPv6 (only) network. Two modes of such a function are currently proposed and presented in the following subsections

3.3.1. 4V6 Mapped Tunnel Mode

This type of IPv6 adaptation function is adopted and described in [I-D.despres-softwire-4rd]. It operates by mapping the IPv4 addresses and the port Index derived from received IPv4 packets and their UDP/TCP payload into an IPv6 address, all in the context of an 4V6 domain. Then, the original IPv4 packet is sent across the IPv6 domain encapsulated as an IPv4inIPv6 packet. The IPv4 packet header is then statelessly extracted by the 4V6 CE or 4V6 gateway before transmission outside of the domain.

The figure below illustrates IPv4 packet transport in 4v6 Mapped Tunnel mode.

TBC

3.3.2. 4V6 Translation mode

This type of IPv6 adaptation function is adopted and described in [I-D.murakami-softwire-4v6-translation] and [I-D.xli-behave-divi]. It operates by mapping the IPv4 address and the port Index derived from received IPv4 packets and their UDP/TCP payload into an IPv6 address, all in the context of an 4V6 domain. Then, it is only the TCP/UDP payload of the original IPv4 packet is sent across the IPv6 domain as a regular TCP/UDP over IPv6 packet. The IPv4 packet header is then statelessly recreated by the 4V6 CE or 4V6 gateway before transmission outside of the domain. The operation of the 4V6 CE and gateway are very similar, if not identical, to that of stateless NAT64 as specified in RFC 6053.

The figure below illustrates IPv4 packet transport in 4v6 Translation mode.

TBC

4. Comparison of 4V6 transport modes

This section presents the an overview of the similarities and differences between an IPv4-IPv6 translation based 4V6 transport mode and one that utilizes IPv4-in-IPv6 mapped transport. The comparison takes into consideration a wider deployment view composed of functionality that is known to be in common use today, as well as more specific functions appearing in architectures defined by CableLabs, and 3GPP.

4.1. General Characteristics of 4V6 modes

The following table presents a comparison of the 4V6 transport modes, in terms of the base technology, and constrains, including also IPv4.

Item	4V6 Translation mode	4V6 Mapped Tunnel Mode
Base Technology	Port restricted NAPT44 with modified stateless NAT64	Port restricted NAPT44 with IPv4 in IPv6 mapped encapsulation
Location of NAPT44	CPE	CPE
IPv4 Forwarding paradigm	L3 + L4 lookup	L3 + L4 lookup
IPv6 Addressing Constraints	CE uses dedicated 4V6 suffix.	CE uses dedicated 4V6 suffix.
Type of IPv6 prefix/address announcement method supported	ICMPv6 (SLAAC), DHCPv6 (both IA_NA and IA_PD)	ICMPv6 (SLAAC), DHCPv6 (both IA_NA and IA_PD)
Can 4V6 IPv6 prefix be used by non 4V6 devices	Yes	Yes
IPv4 addressing constraints	Fixed sharing ratio per IPv4 address.	Fixed sharing ratio per IPv4 address.
TCP/UDP Port range constraint	Ports are statically allocated	Ports are statically allocated
Requires ALG64 or DNS64	No	No
Requires IPv6 DNS	Recommended	Recommended
4V6 CE Parameter provisioning methods (given suitable protocol extensions)	ICMPv6, Stateless DHCPv6, TR69	ICMPv6, Stateless DHCPv6, TR69.
IPv6 Domain Routing to CE based on:	Regular closest IP match to CE-IPv6 subnet	Regular closest IP match to CE-IPv6 subnet

IPv6 Domain Routing to S64 Gateway based on	IPv6 4V6 domain aggregate route	4V6 Gateway unicast/anycast address
-----	-----	-----
IPv4 Header Checksum recalculation required	Yes	No
-----	-----	-----
Supports non TCP/UDP Protocols	No*	No*
-----	-----	-----
Supports IPv4 fragmentation (without additional state)	No	No
-----	-----	-----
Requires IPv6 PMTU discovery/configuration	Yes	Yes
-----	-----	-----
Supports IPv4 Header Options	Yes - limited as per NAT64 [RFC6145]	Yes, except source route option
-----	-----	-----
Overhead in relation to average payload of a) ~550 bytes b) 1400 bytes).	a) 0% b) 0%	a) 4.36% b) 1.71%
-----	-----	-----
Supports non-shared IPv4 usage (ie whole IPv4 address assignment to a single device)	Yes	Yes
-----	-----	-----
Can support IPv4 to IPv6 host communication (for traffic not requiring ALGs)	As per [RFC6145] stateless NAT64 specification	No
-----	-----	-----

* Without specific ALGs. Non UDP/TCP protocols, like ICMP, can be supported with specific ALGs.

4.2. Mobile SP Architecture and 4V6 Applicability

This section presents the applicability and comparison of the 4V6 modes to current 3GPP architectures used by Mobile SP for delivering

all sorts of mobile services.

4.2.1. 3GPP overview

The 3rd Generation Partnership Project (3GPP) is a collaboration between groups of telecommunications associations, whose scope is to develop a globally applicable mobile phone systems and architectures based on service requirements. 3GPP standards are structured as Releases, each of which incorporates numerous individual standard documents. Currently, 3GPP Release 7 is the latest release in common practical deployment, with Release 8 being readied for deployment. Releases 9 and 10 are finalized, and work is underway on Release 11.

One of the major service requirement drivers of recent and ongoing 3GPP releases is the realization of services that deliver specific QoS, or user charging goals, all based on a policy system (eg tiered data rate or volume plans). Technically this translates to the Policy and Charging Control (PCC) framework, which in turn attributes specific functionality to nodes in the 3GPP architecture, such as the PDN-Gw and the PCRF. This functionality comprises both data-plane features (eg IP flow classification) as well as the interfaces/protocols between nodes (eg Diameter, and its specific 3GPP applications).

The 3GPP specifications allow both IPv4 and IPv6 traffic to be handled, and subject to operator defined handling and charging policies by means of applying suitable user traffic filters. Such filters are currently defined to be either IPv4 or IPv6, are applicable to user plane traffic, and are used in a variety of critical roles including the signalling of PDP contexts/EPC Bearers, as well as PCC signalling and interaction with applications.

The following table illustrates the impact of the 4V6 translation and tunnel transport modes respectively on the 3GPP architecture including PCC interfaces. In assessing the impact of these 4V6 transport modes a number of additional assumptions are taken:

- o The 3GPP system supports native IPv6 user traffic, as say per either of the E-UTRAN Release 8 or 9 specifications, using the relevant EPS bearer or PDP functionality.
- o The 4V6 gateway functionality is not part of the 3GPP core architecture (given that currently it is not scoped by a 3GPP Release). Instead, the 4V6 gateway is taken to be a stand alone component in the 3GPP network operator's core reachable via the SGi interface.

The above system, in the context of 3GPPs E-UTRAN architecture as

defined in [E-UTRAN] is shown in Figure 2

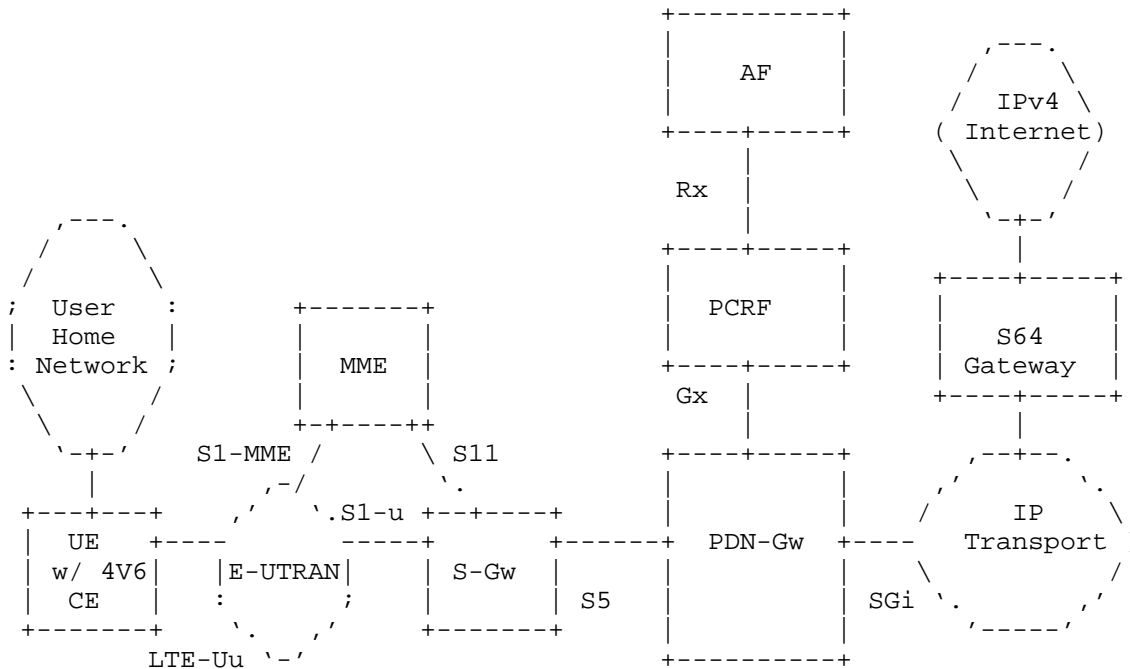


Figure 2 - 3GPP Architecture with 4V6

The main 3GPP system components, and terms are summarized as follows (the reader is referred to [E-UTRAN for a more detailed definition]:

UE The User Equipment, typically a phone or a 3G/4G capable Home Router (shown to incorporate 4V6 functionality)

E-UTRAN Evolved Universal Terrestrial Radio Access Network. The Radio Access network, composed on E-NodeB elements.

MME Mobility Management Entity. Responsible for user authentication, PDN/SGw selection. Does not interact with the user data plane

S-Gw Serving Gateway (function). Responsible for handling local mobility, (some) traffic accounting, traffic forwarding, bearer establishment.

PDN-Gw Packet Data Network Gateway (function). Responsible for per user IP traffic handling, incl. address assignment, filtering, QoS, accounting.

PCRF Policy And Charging Rules Function. Responsible for authorizing and applying policy rules, as well as binding them to user bearers.

Bearer The bearer represents a virtual connection, typically that between a UE and a PDN-Gw. The bearer is specified as an IP Fliter (in terms of IP address, port numbers) and is the object of policy rules. 3GPP, depending on Release and document, defines many terms that are used to refer to the same notion: PDP context, EPS Bearer.

AF Application Function. A functional element offering (higher level) applications that require dynamic policy and/or charging control over the user plane (bearer) behaviour. The AF can be seen as bridging the gap between applications and how they affect the IP data plane of a user.

S5 It provides user plane tunnelling and tunnel management between SGW and PDN-GW, using GTP or PMIPv6 as the network based mobility management protocol.

S1-u Provides user plane tunnelling and inter eNodeB path switching during handover between eNodeB and SGW, using the GTP-U protocol

SGi It is the interface between the PDN-GW and the packet data network. Packet data network may be an operator external public or private packet data network or an intra operator packet data network.

Gx Bearer and flow control interface between the user data-plane element (PDN-Gw) and the Policy System. A Diameter based interface with a suite of 3GPP applications

4.2.2. 3GPP and 4V6 modes

4V6 translated traffic appears for all intents and purposes as regular IPv6-user traffic to the 3GPP system and packet processing functions (eg the PDN-Gw). Hence, and based on the stated assumptions, any such 4V6 traffic can be handled using existing native IPv6 functionality defined by the core 3GPP specifications.

In contrast, 4V6 tunneled traffic requires additional data plane processing to get to the "real" user IPv4 payload and apply the desired functions. Such additional processing is currently not part

of the functionality covered by the 3GPP specifications. In view of this, and solely in relation to the 4V6 tunnel transport mode, two alternative hypotheses need to be placed in order to complete the comparison

i) that such IPv4 in IPv6 processing functionality will be supported as part of the existing EPS bearer functionality defined in E-UTRAN, perhaps as a dedicated EPS bearer (ie an additional virtual interface per subscriber). Or, that;

ii) a new 46 EPS bearer type (ie interface type) identification and signalling will be defined by the 3GPP architecture, which formalizes the v4inv6 relationship between the IPv4-user payload and the v6-user layers.

An apparent benefit of approach (ii) would be in allowing the system to clearly distinguish and expose to other systems v4-user traffic versus v6-user traffic, which is composed of v4inv6 and regular v6 traffic that a UE may generate. The former approach (i) is more convoluted given the ambiguity in distinguishing, and representing such a combination of v6-user and v6-user-bearer and v4-user traffic, all while keeping coherence in terms of the policy system. These two options are designated with ** in the table below.

Item	4V6 Translation Mode	4V6 Mapped Tunnel Mode
User Data Plane at the PDN-Gw (as per section 5.1.2 in [EUTRAN])	IPv6 over GTP-U over UDP over IP	IPv4 over IPv6 over GTP-U over UDP over IP
Gx (Diameter)	No discernible impact	Impacted: no way to express v4 over v6 in TFT Filter and Flow Descriptors
Rx (Diameter)	No discernible impact	Impacted: no way to express v4 over v6 in Media-Component-Description and, Flow-Description-AVP
S5 (GTP)	No impact	Impacted with new PDP/EPS Bearer type*
New 46 Bearer definition	Not required	Possibly required**
Secondary interface (dedicated bearer or secondary PDP) for 46 traffic	Not required	Possibly required**
PDN-Gw	No impact	New TFT capability, IP Gate functionality, changes to Gx, and likely changes to S5/S7 related to signalling the new bearer
SGw	No Impact	No discernible impact
PCRF	No impact for IPv6. Feature to map IPv4-IPv6 addresses needed only in case of IPv4-only applications.	Impacted for both IPv6 and IPv4-only applications and Gx applications utilizing flow control/charging

AF Application Function	No discernible impact	Flow based application control impacted
UE	S46 application	S46 application
LTE-Uu	No discernible impact	Likely changes required to support signalling of EPS bearer or PDP type
Lawful Intercept	No discernible impact	New rules for tunnel support

*A new PDP Type or EPS bearer signalling has a broader 3GPP system wide impact not fully covered here.

As the table illustrates, the 4V6 tunnel transport model appears to affect a significant number of 3GPP elements, when the intent is to realize a full suite of services. This observation appears to apply to any other carrier inserted tunneling technology (eg DS-lite). Hence, a substantial investment in 3GPP standard terms and in the evolution of deployed systems appears to be required.

In contrast the 4V6 translation mode bears none to no discernible impact on existing 3GPP Release 8/9 specifications and their deployments, while allowing the operator to realize the full set of services on 4V6, alongside any native IPv6 traffic, allowed for by these architectures. Hence, little beyond the addition of 4V6 components operating using translation mode appears to be required.

4.3. Cable SP Architectures & S46 Applicability

Cable SPs (commonly referred to as Multi System Operators (MSOs)) usually deliver video, data, and voice service over the cable and fiber access to residential and commercial customers. Many MSOs offer SLAs with various services by exploiting QoS not only in their IP/MPLS network, but also their access network.

The cable access network (now synonymous with Hybrid Fiber Coax (HFC)) is commonly enabled with Data Over Cable Service Interface Specifications (DOCSIS, a CableLabs standard) to facilitate the implementation of packet based services. In this paradigm, the HFC/DOCSIS access bandwidth is typically shared among a number of customers, hence, ensuring optimal service quality & experience per customer becomes extremely important for MSOs' success.

Cable SPs/MSOs ensure the optimal service quality of various advanced

& real-time multimedia services (such as IP telephony, multimedia conferencing, interactive gaming etc.) by utilizing "PacketCable" framework to enforce QoS on the HFC/DOCSIS access.

The next sub-section 4.3.1 provides a brief introduction to PacketCable, section 4.3.2 explains a key PacketCable construct - Classifier, and section 4.3.3 tabulates the impact of 4V6 modes to PacketCable enabled DOCSIS/IP services.

4.3.1. PacketCable Introduction

PacketCable, a CableLabs standard, defines a framework for ensuring the Quality of Service (QoS) on the HFC/DOCSIS Access. PacketCable specifications (including PacketCable Multi Media [PCMM] and PacketCable Dynamic QoS [PC-DQOS]) specify interoperable interface specifications to implement Dynamic QoS, Admission Control, Accounting, Policy, and Security functions.

As an example, the figure below illustrates PCMM [PCMM] architecture including a set of IP-based interfaces (referred to as pkt-mm-1 through 12) that leverage core QoS and policy management capabilities.

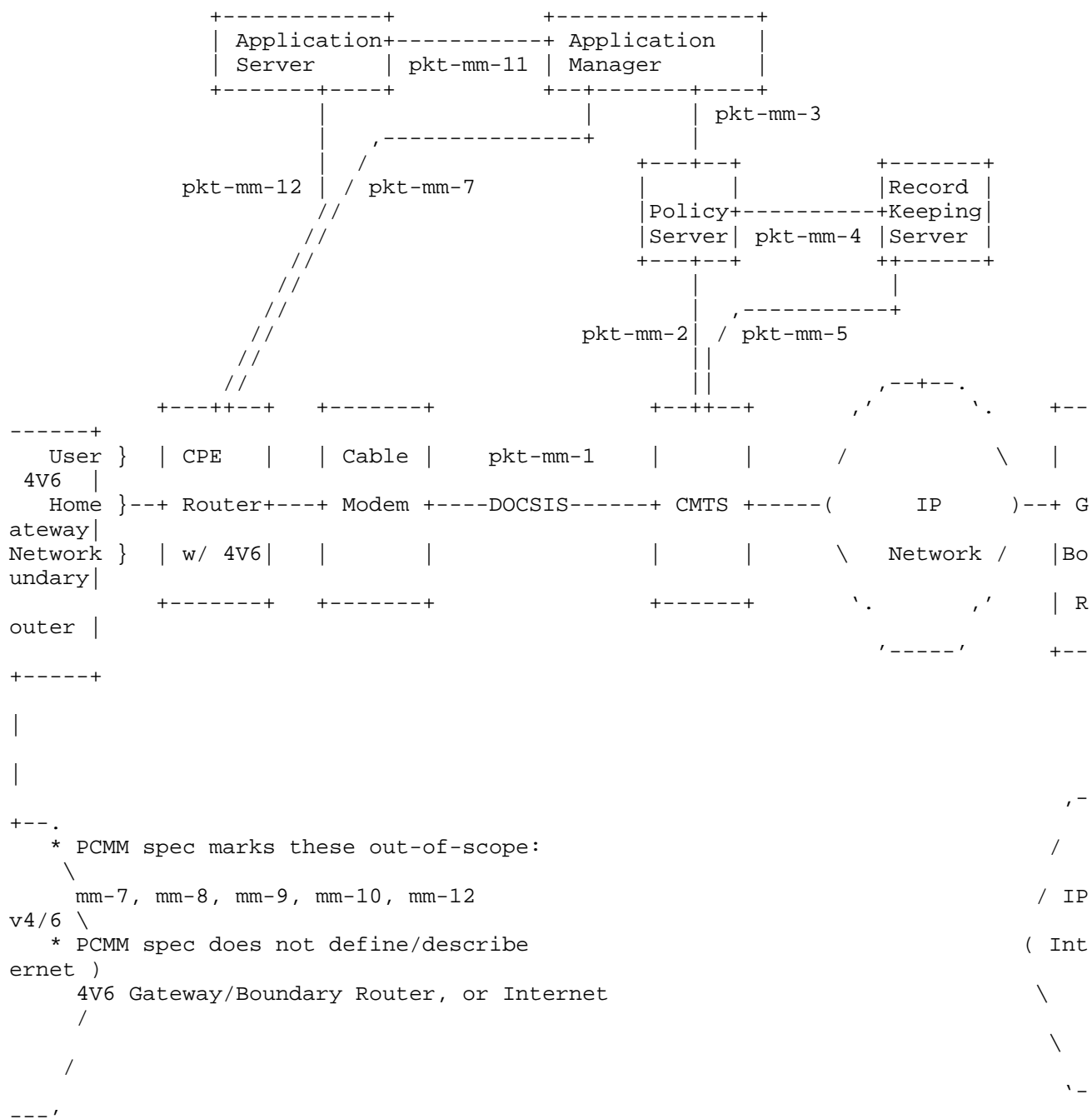


Figure 3 - PacketCable Multimedia Architecture (with 4V6)

The PacketCable framework is also critically important for MSOs to comply with government regulations for things such as E911 when they offer voice/telephony services.

4.3.2. PacketCable Construct - Classifier

PacketCable framework fundamentally relies on Cable Modem (CM) and Cable Modem Termination System (CMTS) to first qualify and then classify the appropriate IP traffic between them, for effective QoS enforcement. The framework requires the usage of "Classifier" for both qualification (in control plane) and classification (in data

plane).

For ex, PCMM specification [PCMM] mandates the usage of classifier in the control plane (i.e. 'Upstream Packet Classification Encoding' in pkt-mm-1 interface (DOCSIS) , whereas 'Multimedia Classifier Object' in pkt-mm-2 and pkt-mm-3 interfaces (COPS)) for conveying the attributes of an IP flow belonging to an application (telephony, say), and subsequently its usage in the data plane i.e. filter matching on the IP packets' layer2/3/4 headers prior to QoS treatment.

The PCMM specification mandates the 'classifier' to include Source and Destination IP addresses, DSCP/TOS, IP Protocol, Source and Destination ports for an IPv4 traffic flow, and Source and Destination IP addresses, TC, Next Header, Source and Destination ports for an IPv6 traffic flow. Hence, the CMTS and CM would construct their data-plane filter based on the classifier.

The PacketCable DQOS specification [PC-DQOS] mandates the 'protocol' (or next header) in IP header to be 17 (=UDP) in the classifier to accommodate voice RTPoUDPoIP traffic.

The above means that Cable SP/MSOs following these specification cannot benefit from the PacketCable framework if the IP traffic pertaining to application traffic that is encapsulated in another IP header (e.g. tunneling mode) on the DOCSIS access.

4.3.3. PacketCable MultiMedia & 4V6 Modes

4V6 translated traffic appears for all intents and purposes as regular IPv6-user traffic to the PacketCable framework (both control plane and data plane). Hence, it is likely that any such 4V6 traffic can be handled using native IPv6 functionality defined by the PacketCable (Multi Media) specifications and supported by IPv6 capable CMTS devices.

PCMM specification already allows for (and mandates) a minimum of four classifiers to be included in Gate-set. Hence, a Policy Server can communicate (via pkt-mm-2) both IPv4 and IPv6 classifier to the CMTS, which can use IPv6 classifier for constructing its filters, and use IPv4 classifier to convey to the CM via DOCSIS messages (pkt-mm-1)

In contrast, 4V6 tunneled traffic requires additional data plane processing to get to the "real" user IPv4 payload and apply the desired functions. Such additional processing is currently not part of the functionality covered by the PacketCable specifications, nor part of compliant implementations.

5. Overview of potential issues and discussion

This section summarizes the issues attributed to an A+P, or port restricted scheme, along with a discussion of applicability to the assumed system and possible resolutions. The summary of issues stem from [I-D.thaler-port-restricted-ip-issues] and associated discussions.

5.1. Notion of Unicast Address

5.1.1. Overview

The issue, referred to as the "definition of a unicast address", relates to the notion that in a shared IPv4 address system, multiple hosts will be visible as having a single IPv4 address outside of the system. This issue is a general characteristic of any NAPT44 based solution [I-D.ietf-intarea-shared-addressing-issues], including DS-Lite. However, a more specific aspect of this issue in the context of an address sharing system is the possibility that a single host having multiple interfaces will be assigned the same IPv4 address (with different port ranges) on each of its interfaces. It may also be that multiple hosts sharing an address find themselves on the same Layer 2 segment. Either would impede hosts from working within the notion of known host IP stack and protocol implementations.

5.1.2. Discussion

A number of the characteristics of the 4via6 solution architecture cause the issues not to be applicable, key of which is that there is no expectation for any kind of end hosts to be part of the shared IPv4 address system.

In the stateless 4via6 system, CPE nodes are assigned with a shared IPv4 address+port range by means of the unique IPv6 address, containing the embedded IPv4 address + port index, of that CPE node. The CPE node is in addition enabled to be running the port restricted NAPT44 function from the IPv6 derived address, a key characteristic of the solution. On the IPv6 plane, the IPv6 address of the CPE is practically indistinguishable from any "regular" IPv6 address, and in fact any host that is not aware of it conveying an embedded IPv4 address would be able to use this just like any other regular IPv6 address, ie the 4via6 solution uses standard IPv6 addressing. In terms of the IPv4 dimension, since the shared address and port index are never used to address native IPv4 nodes or hosts, but instead uniquely assigned to a single NAPT44 function that is part of the CPEs, all legacy or other IPv4 hosts are not exposed to the presented issues.

Going beyond the ascribed issue however, it appears desirable to have the 4via6 CPEs that are to be part of the shared system to be able to provide a hint to the network operator in terms of their special capability. Such a hint can be a DHCPv6 Option Request Option, which would be useful to allow the DHCPv6 sub-system to also inform the CPE of any other stateless 4via6 system parameters. A largely similar ORO option is currently being defined as part of [I-D.ietf-software-ds-lite-tunnel-option]

Recommendation: Define a suitable DHCPv6 ORO for conveying the 4via6 capability of a CPE.

5.2. Implementation on hosts

5.2.1. Overview

The issue, as presented, relates to the need for modifications on end hosts or devices to support a port constrained mechanism and the overall impossibility of realizing such modifications. Furthermore, host applications that attempt to bind to specific ports that are not part of the allowed port range will fail to do so and may also require modifications.

5.2.2. Discussion

As presented in Section 3 the solution assumes the use of a dedicated CPE implementing the 4via6 functionality within a port constrained mode and NAPT44. Granted, CPE nodes will require to implement new functionality such as the IPv6 adaptation function, that is likely alongside introducing native IPv6 support. However, any and all existing end user IPv4 devices (eg PCs, etc) will not be affected. Nor are such devices expected to behave in any way different from that of today, where they typically obtain a private rfc1918 address and multiplexed by a CPE using a NAPT44 function.

In summary, the assumed 4via6 solution requires a specific 4via6 CPE but does not require any IPv4 host stack changes.

5.3. 4V6 address and impact on other IPv6 hosts

5.3.1. Overview

The issue relates to the question of whether the operation of a regular IPv6, non 4V6 capable, host would be adversely impacted should it be assigned or auto-configured with an address from an S64 address or prefix pool.

5.3.2. Discussion

The 4V6 prefix is for all intents and purposes a regular IPv6 prefix, and as such can be announced/assigned to any IPv6 host which in turn can use derived addresses like any other IPv6 address. Thus, an 4V6 IPv6 domain can address non-4V6 devices, leaving such devices to operate as native IPv6.

There is however a restriction on the 4V6 CE devices. As described in Section 2, a 4V6 CE constructs itself the full 128 bit address from the concatenation of the IPv6 prefix, 4V6 domain information acquired statelessly, and a pre-determined or algorithmic interface-id. By definition, only one 4V6 CE can use the same IPv4 address and port index. Hence, while there is no exact limitation on the number of non 4V6 hosts that can be addressed from an 4V6 prefix, there is a limit of one 4V6 CE per 4V6 prefix. Using a 4V6 prefix to address network segments without 4V6 devices does diminish the efficiency of the IPv4 address sharing mechanism, in terms of using up port ranges onto segments that will not use them. This is naturally a deployment consideration which an operator can optimize.

5.4. Impact on 4V6 CE based applications

5.4.1. Overview

It has been claimed that applications implemented on the CE itself, eg a DNS resolver-client, may be impacted by the 4V6 functionality. In particular, a concern is that such applications would either need to be specially engineered to issue socket calls or extensive IP stack modifications made to support them.

5.4.2. Discussion

By definition the 4V6 CE is an IPv6 capable device, and any IPv6 capable applications will be able to use the native IPv6 stack (note: IPv6 interface selection, is discussed in section 5.5). As such, the concern raised does not apply to applications that can be expected to support IPv6, and instead only to IPv4-only applications running on the 4V6 CE.

The shared IPv4 address is intended to be used only by the 4V6 CE function. This shared IPv4 address does not need to be assigned to an interface on the 4V6 CE and thus a target for potential applications. Any such applications running on the 4V6 will use any of the other (likely private) IPv4 address on the CE, which then will be routed to the 4V6 function this is applied post routing for the packets generated by these applications.

5.5. 4V6 interface

5.5.1. Overview

A 4V6 CE will have a "self configured" 4V6 IPv6 interface address, alongside any other SLAAC or DHCPv6 derived addresses, potentially from the same prefix. This particular 4V6 address may be subject to specific filtering rules or restrictions by the operator, besides usage and filtering restrictions on the 4V6 CE. Also, for the 4V6 system to operate as intended, the 4V6 application on the CE must be restricted to using the specific 4V6 address when sourcing 4V6 packets. Also, the 4V6 CE needs to be set-up to correctly forward IPv4 traffic to the 4V6 application.

5.5.2. Discussion

While the method of creating the interface is implementation specific, the generic operating model that is envisaged is for the 4V6 application to create the 4V6 interface as a virtual interface with an IPv4 unnumbered address. The application would then install a default IPv4 route pointing to this virtual interface, which would be effectively see the 4V6 application acting as a network appliance on the forwarded traffic. In terms of IPv6 behaviour, the 4V6 application is expected to be set up to specify the use (binding) to the 4v6 IPv6 virtual interface.

5.6. Non TCP/UDP port based IP protocols - ICMP)

5.6.1. Overview

This issue relates to the inability of using regular ICMP messages to "ping" an end-host that has been addressed with a shared IPv4 address. The issue can be generalized one applicable to any IP protocol that is not TCP/UDP port based, and also in terms of the ability of using such protocols from end hosts that are assigned a shared IPv4 address.

5.6.2. Discussion

The inability to ping a CPE from the IPv4 Internet is shared by other IPv4 address sharing mechanisms such as DS-Lite. Thus, the issue is no better or worse in the case of the stateless 4via6 solution. The same can be said of end user hosts using other non UDP/TCP port based protocols from behind a NAT44 function, ie they will not function irrespective of address sharing or not.

As discussed in [I-D.ietf-intarea-shared-addressing-issues], all IP address sharing solutions break protocols which do not use transport

numbers. A mitigation solution is to utilize specific ALGs. For ICMP in particular, a mitigation solution would be to rewrite the "Identifier" and perhaps "Sequence Number" fields in the ICMP request, treating them as if they were port numbers.

As a conclusion, this issue can be partially mitigated, likely at par to centralized NAT solutions.

5.7. Provisioning and Operational Systems

5.7.1. Overview

The general claim of this issue is that a service providers' provisioning and accounting systems would need to [radically] evolve to deal with the notions of shared IPv4 addresses and port range constrains.

5.7.2. Discussion

The stateless 4via6 solution relies on a fully operational IPv6 network, which on the IPv6 plane fundamentally does not differ from a regular IPv6 network, and the stateless 4via6 solution may be seen as an IPv6 application - devices connecting to the network, need unique IPv6 addresses which the network is able to provide. In the 4via6 solution it happens that these unique IPv6 addresses embed an IPv4 address. Hence, additional system enhancements that the stateless 4via6 solution requires, over and above those simply needed to deploy and operate an IPv6 network, lie in the domain of supporting the provisioning of the IPv6 adaptation functionality of the CPEs. This may require the operator to use DHCPv6, or other provisioning methods such as IPv6CP, TR-69, in order to configure any relevant 4via6 service parameters to a CPE.

From an IPv4 perspective, an operator will likely want to have a management system capable of the assignment of IPv4 addresses to the shared pool, and tuning the re-use factor. In this, the solution exhibits no grossly different characteristics than those of any system with an operator managed NAT44 function where similar management capabilities need to be introduced.

One additional aspect of the stateless 4via6 solution needs to be highlighted. On a par basis this solution requires less per subscriber management, accounting and logging capabilities than centralized NAPT44 alternatives such as DS-Lite, due to the following:

- o The assignment of an IPv6 address that embeds a deterministic IPv4 address and port range removes the need for the operator to perform any NAPT44 binding logging, ie the task of determining which user had a given IPv4 address and port at a given time is simply a matter of determining who had the corresponding IPv6 address, rather than collecting large amounts of dynamic binding data.
- o There is no need for the operator to manage NAPT44 binding data access and retention.
- o Given the stateless nature of the 4via6 solution, all subscriber CPEs in an operator's domain can share exactly the same 4via6 service configuration, i.e. The operator does not need to be concerned with managing on a per user basis specific AFTR assignment and/or load balancing such users and throughout ensuring symmetric traffic flows throughout.
- o The location of the NAPT44 function on the user's CPE, allows easy and direct management of the port mappings by the end user removing a need for the operator to introduce PCP [I-D.wing-software-port-control-protocol] (or similar) protocols in on AFTRs, and on CPE devices. In effect the end user can retains control of any bindings, which could be via today's GUI, or UPnP IGDv2, or even PCP.
- o As and when needed, a stateless 4via6 solution readily supports the assignment of an unshared IPv4 address, and full port control by the end user. A similar capability with centralised NAPT44 solutions involve onerous management of per subscriber configurations on the operator's AFTR.

5.8. Training & Education

5.8.1. Overview

The issue claims a concern with the need for developers and support staff to be trained & educated in dealing with a port constrained systems.

5.8.2. Discussion

There appear to be at least two levels of looking at this issue in the stateless 4via6 context. On one level, it is perfectly true that developers and support staff will need to be trained with running/ supporting a native IPv6 network, that is now a basis of the solution. This however is an inherent aspect of deploying an IPv6 network and applications. On another level, support and developers

need to familiarized with the NAPT44 characteristics of the system, that are not different from those already known about such systems today. More specifically, there appears to be no such thing as a port unconstrained carrier grade NAPT44 system, in either tomorrow's stateless 4via6 or AFTR guises, or today's residential CPE NAPT44 implementations that have an inherent hard set translation limit (often 1024 translation, corresponding to a usage of 1024 ports). That application developers should be trained to be reasonably conservative in the usage of ports is thus not an issue of the stateless 4via6 solution, but pretty much of any NAPT44 based solution, even those in use today.

Another useful observation here is that the stateless 4via6 solution, actually allows an operator to retain existing troubleshooting procedures, given which today encompass CPE based NAPT44, rather than changing them radically to an AFTR. Furthermore, it is possible to alleviate any port-range constraints for users by allocating more generous port ranges without the need to manage such users configuration on active core network devices (eg AFTR).

5.9. Security and Port Randomization

5.9.1. Overview

Preserving port randomization [RFC6056] may be more or less difficult depending on the address sharing ratio (i.e., the size of the port space assigned to a CPE). Port randomization may be more difficult to achieve with a stateless solution than stateful solution. The CPE can only randomize the ports inside be assigned a fixed port range.

5.9.2. Discussion

The claim that a stateless 4via6 solution grossly affects security does not appear to be entirely accurate when considering the following i) the actual threat ii) a comparison to a centralized NAPT44 solution, at par value in terms of the number communicating of IP end-points (inside) utilizing the system iii) the ability to use native IPv6 as well as proposed extensions.

First and foremost, an 4V6 system is a native IPv6 system, which can use the IPv6 interfaces in a port-unrestricted mode, which means that the port restriction carries no security implications.

Assuming all other information has already been acquired, and also the presence of no exploits, the basic security of IP protocols lies in the computational cost of guessing a combination of a protocol field, eg a TCP sequence number or say a DNS transaction ID, multiplied by a the cost of a guess of a source port. Thus, for a

TCP brute force attack, the already substantial cost in guessing a sequence number (2^{32} possibilities) is multiplied by the port range guess cost. In this context a 1K port restricted 4V6 system, is costly enough for most practical purposes. More discussion to improve the robustness of TCP against Blind In-Window Attacks can be found at [RFC5961].

For a UDP/DNS brute force attack, the computational cost required to scan/generate the full range of 2^{32} possibilities, corresponding to a port unrestricted system is relatively low/easy - today's GPUs can do so in a few seconds. UDP/DNS can be said to be inherently vulnerable, with the solution residing in DNSec. A 4V6 port restricted system would indeed lower such a computational cost, but for practical purposes it will still be in the order of seconds. Moreover, for the case of UDP/DNS, the CPE is expected to use its DNS proxy resolver functionality, which acting on native IPv6, is totally unaffected by the port restriction, and thus any of the security claims. Other means than the (IPv4) source port randomization to provide protection against attacks should be used (e.g., use [I-D.vixie-dnsextdns0x20] to protect against DNS attacks, [RFC5961] to improve the robustness of TCP against Blind In-Window Attacks,

Beyond the above, in terms of the avoiding a fixed linear or single port range allocations, extensions to the 4V6 solution provides additional remedy, eg [I-D.bajko-pripaddrassign].

In general thus, with a 4V6 solution it is possible to realize a viable level of security, which for practical purposes offers does not and extending the 4V6 solution with . For some applications, like DNS, it is recommended that IPv6 be used. It remains an area of possible further proposals for optional port range randomization methods to be combined in a 4V6 solution.

5.10. Unknown Failure Modes

5.10.1. Overview

The issue purports that a system with a port constraints introduces new unknown failure modes, not known with NAT44 or NAPT44 systems, and in general is more complex than such a system.

5.10.2. Discussion

This claim does not appear to have objective technical arguments that can be discussed. A restricted port range system, such as the one assumed in this document, does not appear to have any more or less complexity than any of the other NAPT44 solutions against which the same issue has not been levelled. That is a statement that can be

made in consideration of each of those alternative solution network design (eg elaborate routing rules or topologies) and feature implementation complexities, which appear to be no better than that of a stateless 4via6 address port range system. Ultimately, system complexity is something best left adjudicated by the operators choosing to deploy one or the other of these IP based transition solutions.

5.11. Possible Impact on NAT66 use & design

5.11.1. Overview

The notion of a shared address with a constrained port range is seen as possibly bearing influence on use in future schemes involving NAT66, where IPv6 address sharing is in general deemed not to be desired (ie there is good reason to avoid PAT66).

5.11.2. Discussion

The authors do not propose, nor expect to see the IP address sharing characteristic applying to future NAT66/PAT66 discussions and specification. However, having said that it is useful to take a humble step back and consider the general aspect of causality in this context. The direct cause that brought about IPv4 shared address solutions to the fore was a shortage/exhaustion of a limited IPv4 address resource, alongside a failure of the community to migrate IPv4 networks to IPv6 in a timely manner. At the time of writing it is hard to imagine the same occurring with respect to IPv6 address resources, and hopefully the same set of causes will not be allowed to re-occur. This appears to be the only way to ensure that IPv6 address sharing effect does not come to be, as opposed to precluding such notions within the context of today's IPv4 world where the causality is rather clear.

5.12. Port statistical multiplexing and monetization of port space

5.12.1. Overview

An issue attributed to 4V6 solutions is that due to their characteristic of assigning a fixed amount of ports to participating system nodes, the overall pool of ports cannot be dynamically/statistically multiplexed.

A corollary of this claimed issue is the claim that port range constraints will lead to monetization by service providers of such port ranges, for example by charging users based on the number of ports assigned or creating some bronze, silver, gold type of port based service categories.

5.12.2. Discussion

The 4via6 address shared solution indeed limits the ability to "overload" ie statistically multiplex amongst users, the ports available of a given public IPv4 address. This can be seen as a trade off vs dynamic allocation and the need to log (large amounts) of NAT bindings. Furthermore, the solution is meant to be fundamentally a transitional one for supporting legacy IPv4 users till full migration to IPv6 can occur. As an example, even with a static allocation of ~1000 ports per shared IP user, it allows an operator to effectively multiply by ~64 the current IPv4 unrealizable address space. To put it into a network growth perspective, it allows an operator to support for some 10 years a steady 50% annual increase in users, without requiring new IPv4 addresses. This is likely an alluring (if unlikely) prospect for most, but it demonstrates the fact that even with static port allocations, IPv4 address sharing can go a long way for many operators.

CGN-based solutions, because they can dynamically assign ports, provide better IPv4 address sharing ratio than stateless solutions (i.e., can share the same IP address among a larger number of customers). For Service Providers who desire an aggressive IPv4 address sharing, a CGN-based solution is more suitable than the stateless. However, in case a CGN pre-allocates port ranges, for instance to alleviate traceability complexity it also reduces its port utilization efficiency.

5.13. Readdressing

5.13.1. Overview

Due to the port range encoding being part of the CPE's IPv6 address, any change in the range requires a re-configuration of the CPEs 4via6 address. This is said to be an issue given the impact that IP address changes have on existing traffic flows, as well as general IPv6 network routing

5.13.2. Discussion

It is true that under the assumed notions of the stateless 4via6 solution, IPv6 re-addressing is required to effect a change in terms of the shared IPv4 address or ports. Such changes can and are likely best done using dynamic address configuration methods such as DHCPv6, or alternatively out of band management tools, eg TR-69, especially when the 4via6 address can be derived from a delegated prefix. Using these, the impact of the address change does not translate to a neither a classic IPv6 host renumbering problem nor an unmanageable network renumbering problem. On the CPE, the change only affects the

4via6 address of the CPE and not any end user IPv6 hosts behind the CPE (which would likely continue to derive their IPv6 addresses from an unchanged delegated prefix). On the service provider network side, the change, if any, represents a network renumbering case which the operator can be reasonably expected to handle within their network numbering plan, especially given that the IPv6-prefix of the an IPv4-in-IPv6 address is summarizable.

An addressing change will impacting any existing IPv4 flows that are being NAT'ed by the CPE. This is also analogous to the today's practice of IPv4 address changes espoused by some operators, which while not being commendable, is established in the market. Nevertheless, as a means of alleviating such an impact it appears desirable for the solutions to investigate the viability of mechanisms that could allow for more graceful addressing changes.

To facilitate IPv6 summarization and operator appears to have two 4V6 deployment choices. When encoding IPv4 addresses in lower order address space bits that are subject to summarization, the operator would need to assign a modest dedicated IPv6 prefix (such as a /64) as an 4V6 IPv6 addressing sub-domain. Alternatively, without resorting to a separate 4V6 addressing sub-domain, an operator could allow for the IPv4 address embedding to be embedded in a high-order portion of the IPv6 domain address space, one that closely follows the IPv6 domain prefix. These two valid address subnetting and deployment options deserve better description in the solution specifications.

5.14. Ambiguity about communication between devices sharing an IP address.

5.14.1. Overview

A regular IPv4 destination based routed system inherently does not allow two devices to communicate while sharing the same IPv4 address, even if with different ports. Similarly, such a system does not allow on the basis of a IPv4 source address alone to perform address spoofing prevention. These two issues naturally render regular IPv4 based routed networks incapable of supporting a shared address solution.

5.14.2. Discussion

In terms of the IPv4 data plane of the 4via6 solution, the CPE and the stateless gateway components need to be modified in terms of their IPv4 forwarding behaviour. The CPE's NAPT44 function, must be capable of sending traffic towards the IPv6 adaptation function when the traffic is addressed to its (shared) IPv4 address but a different

port than the one assigned to the CPE. Similarly, the CPE's NAPT44 function must be capable of receiving traffic addressed from its (shared) IPv4 address but a different port than the one assigned to it.

On the IPv6 data plane the stateless 4via6 solution does not suffer from the issue by the nature of relying on regular IPv6 forwarding. Address-spoofing security can be realized on regular IPv6 devices plane, in a way which effectively does not allow a CPE to send IPv6 traffic from a source IPv6 address that it has not been assigned. The spoofing of IPv4 addresses can be prevented in this manner in 4via6 solution relying on translation (dIVI). Tunneling 4via6 solutions (4rd) require IPv6+IPv4 source address validation to be performed at the CPE and stateless gateway, by the IPv6 adaptation function.

The conceptual IPv6 adaptation function has many of its core principles already defined either as part of IPinIP tunneling or stateless NAT64 drafts. However additional work, such as defining the port indexing schemes, is needed and is at the heart of what needs to be covered in the individual solution drafts that fall under the stateless 4via6 family. Throughout, no legacy IPv4 end-systems are expected to implement these techniques.

5.15. Other

5.15.1. Abuse Claims

Because the IPv4 address is shared between several customers, and in order to meet the traceability requirement discussed in Section 12 of [I-D.ietf-intarea-shared-addressing-issues], Service Providers must store the assigned ports in addition to the IPv4 address.

If the remote server does not implement the recommendation detailed in [I-D.ietf-intarea-server-logging-recommendations], the Service Provider may be obliged to reveal the identity of all customers sharing the same IP address at a given time.

5.15.2. Fragmentation and Traffic Asymmetry

In order to deliver a fragmented IPv4 packet to its final destination, among those having the same IPv4 address, a dedicated procedure similar to the one defined in Section 3.5 of [RFC6146] is required to reassemble the fragments in order to look at the destination port number.

When several stateless IPv4/IPv6 interconnection nodes are deployed, and because of traffic asymmetry, situations where fragments are not

handled by the same stateless IPv4/IPv6 interconnection node may occur. Such context would lead to session breakdowns. As a mitigation, a solution would be to redirect fragments towards a given node which will be responsible for implementing the procedure documented in [RFC6146]. The redirection procedure is stateless.

As a conclusion, this issue can be mitigated.

5.15.3. Multicast Services

IPv4 service continuity must be guaranteed during the transition period, including the delivery of multicast-based services such as IPTV. Because only an IPv6 prefix will be provided to a CPE, dedicated functions are required to be enabled for the delivery of legacy multicast services to IPv4 receivers. This is critical since many of the current IPTV contents are likely to remain IPv4-formatted and there will remain legacy receivers (e.g., IPv4-only Set Top Boxes (STB)) that can't be upgraded or be easily replaced.

This issue is similar to the one encountered in the stateful case, and the same solution can be used to mitigate the issue (e.g., [I-D.qin-software-dslite-multicast]).

As a conclusion, this issue can be solved.

6. Conclusion

As per the discussion in this document, the authors believe that the set of issues specifically attributed to A+P based such as the stateless 4via6 solution with characteristics as per Section 3, either do not apply, or can be mitigated. In several aspects, a stateless 4V6 solution represents a reasonable trade off compared to alternatives in areas such as NAT logging, ease of deployment and operations, all of which are actually facilitated by such a solution.

In terms of the 4V6 transport mode, both translation and mapped tunnel appear to be viable and applicable to different contexts. The mapped tunnel mode appears desirable when the operator has no expectations of applying to the service any existing more elaborate traffic based services, such as defined by 3GPP or CableLabs. The translation based approach appears particularly attractive to operators who are concerned with integrating such traffic into a more elaborate suite of services, and minimizing the overhead (esp in relation to wireless transmission). The translation transport mode approach does appear to be free of critical problems, which have historically been found to affect stateful 46 translation based schemes that sought to work on the application layer.

7. IANA Considerations

This document does not raise any IANA considerations.

8. Security Considerations

This document does not introduce any security considerations over and above those already covered by the referenced solution drafts.

9. Contributors and Acknowledgements

The authors thank Dan Wing, Xing Li, Jan Zorz, Satoru Matsushima, Mohamed Boucadair, Qiong Sun, and Arkadiusz Kaliwoda for their review and feedback on the draft.

10. References

10.1. Normative References

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

10.2. Informative References

- [I-D.bajko-pripaddrassign]
Bajko, G., Savolainen, T., Boucadair, M., and P. Levis,
"Port Restricted IP Address Assignment",
draft-bajko-pripaddrassign-03 (work in progress),
September 2010.
- [I-D.despres-software-4rd]
Despres, R., "IPv4 Residual Deployment across IPv6-Service
networks (4rd) A NAT-less solution",
draft-despres-software-4rd-00 (work in progress),
October 2010.
- [I-D.ietf-intarea-server-logging-recommendations]
Durand, A., Gashinsky, I., Lee, D., and S. Sheppard,
"Logging recommendations for Internet facing servers",
draft-ietf-intarea-server-logging-recommendations-04 (work
in progress), April 2011.
- [I-D.ietf-intarea-shared-addressing-issues]
Ford, M., Boucadair, M., Durand, A., Levis, P., and P.
Roberts, "Issues with IP Address Sharing",

draft-ietf-intarea-shared-addressing-issues-05 (work in progress), March 2011.

[I-D.ietf-software-ds-lite-tunnel-option]

Hankins, D. and T. Mrugalski, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Option for Dual- Stack Lite", draft-ietf-software-ds-lite-tunnel-option-10 (work in progress), March 2011.

[I-D.ietf-software-dual-stack-lite]

Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual- Stack Lite Broadband Deployments Following IPv4 Exhaustion", draft-ietf-software-dual-stack-lite-11 (work in progress), May 2011.

[I-D.murakami-software-4v6-translation]

Murakami, T., Chen, G., Deng, H., Dec, W., and S. Matsushima, "4via6 Stateless Translation", draft-murakami-software-4v6-translation-00 (work in progress), July 2011.

[I-D.operators-software-stateless-4v6-motivation]

Boucadair, M., Matsushima, S., Lee, Y., Bonness, O., Borges, I., and G. Chen, "Motivations for Stateless IPv4 over IPv6 Migration Solutions", draft-operators-software-stateless-4v6-motivation-02 (work in progress), June 2011.

[I-D.qin-software-dslite-multicast]

Wang, Q., Qin, J., Boucadair, M., Jacquenet, C., and Y. Lee, "Multicast Extensions to DS-Lite Technique in Broadband Deployments", draft-qin-software-dslite-multicast-04 (work in progress), June 2011.

[I-D.thaler-port-restricted-ip-issues]

Thaler, D., "Issues With Port-Restricted IP Addresses", draft-thaler-port-restricted-ip-issues-00 (work in progress), February 2010.

[I-D.vixie-dnsext-dns0x20]

Vixie, P. and D. Dagon, "Use of Bit 0x20 in DNS Labels to Improve Transaction Identity", draft-vixie-dnsext-dns0x20-00 (work in progress), March 2008.

[I-D.wing-software-port-control-protocol]

Wing, D., Penno, R., and M. Boucadair, "Pinhole Control

Protocol (PCP)",
draft-wing-software-port-control-protocol-02 (work in
progress), July 2010.

[I-D.xli-behave-divi]

Bao, C., Li, X., Zhai, Y., and W. Shang, "dIVI: Dual-
Stateless IPv4/IPv6 Translation", draft-xli-behave-divi-03
(work in progress), July 2011.

[I-D.ymbk-aplusp]

Bush, R., "The A+P Approach to the IPv4 Address Shortage",
draft-ymbk-aplusp-10 (work in progress), May 2011.

[RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's
Robustness to Blind In-Window Attacks", RFC 5961,
August 2010.

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

[RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-
Protocol Port Randomization", BCP 156, RFC 6056,
January 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.

Authors' Addresses

Wojciech Dec
Cisco
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands

Email: wdec@cisco.com

Rajiv Asati
Cisco
Raleigh, NC
USA

Phone:
Fax:
Email: rajiva@cisco.com
URI:

Congxiao Bao
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing, 100084
CN

Phone: +86 10-62785983
Fax:
Email: congxiao@cernet.edu.cn
URI:

Hui Deng
China Mobile
Beijing,
CN

Phone:
Fax:
Email: denghui@chinamobile.com
URI:

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 16, 2012

W. Dec
R. Asati
Cisco
C. Congxiao
CERNET Center/Tsinghua
University
H. Deng
China Mobile
M. Boucadair
France Telecom
October 14, 2011

Stateless 4Via6 Address Sharing
draft-dec-stateless-4v6-04

Abstract

This document presents an overview of the characteristics of stateless 4V6 solutions, alongside a assessment of the issues attributes. The impact of translated or mapped tunnel transport modes is also presented in the broader context of other industry standard reference architectures and existing deployments.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 16, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	4
2. Terminology	4
3. Stateless 4V6 Technical and Architectural Overview	5
3.1. IPv4 address and algorithmic port indexing	7
3.2. 4V6 CE IPv6 Address and domain info	7
3.3. IPv6 Adaptation Function	8
3.3.1. 4V6 Stateless Tunneling Mode	8
3.3.2. 4V6 Stateless Translation mode	9
4. Comparison of 4V6 transport modes	9
4.1. General Characteristics of 4V6 modes	9
4.2. Mobile SP Architecture and 4V6 Applicability	12
4.2.1. 3GPP overview	13
4.2.2. 3GPP and 4V6 modes	15
4.3. Cable SP Architectures & 4V6 Applicability	18
4.3.1. PacketCable Introduction	18
4.3.2. PacketCable Construct - Classifier	20
4.3.3. 4V6 Modes Impact on PacketCable	20
5. Overview of potential issues and discussion	21
5.1. Notion of Unicast Address	21
5.1.1. Overview	21
5.1.2. Discussion	22
5.2. Implementation on hosts	22
5.2.1. Overview	22
5.2.2. Discussion	23
5.3. 4V6 address and impact on other IPv6 hosts	23
5.3.1. Overview	23
5.3.2. Discussion	23
5.4. Impact on 4V6 CE based applications	24
5.4.1. Overview	24
5.4.2. Discussion	24
5.5. 4V6 interface	24
5.5.1. Overview	24

5.5.2. Discussion	24
5.6. Non TCP/UDP port based IP protocols - ICMP)	25
5.6.1. Overview	25
5.6.2. Discussion	25
5.7. Provisioning and Operational Systems	25
5.7.1. Overview	25
5.7.2. Discussion	25
5.8. Training & Education	27
5.8.1. Overview	27
5.8.2. Discussion	27
5.9. Security and Port Randomization	28
5.9.1. Overview	28
5.9.2. Discussion	28
5.10. Unknown Failure Modes	28
5.10.1. Overview	28
5.10.2. Discussion	28
5.11. Possible Impact on NAT66 use & design	29
5.11.1. Overview	29
5.11.2. Discussion	29
5.12. Port statistical multiplexing and monetization of port space	29
5.12.1. Overview	29
5.12.2. Discussion	29
5.13. Readdressing	30
5.13.1. Overview	30
5.13.2. Discussion	30
5.14. Ambiguity about communication between devices sharing an IP address.	31
5.14.1. Overview	31
5.14.2. Discussion	31
5.15. Other	32
5.15.1. Abuse Claims	32
5.15.2. Fragmentation and Traffic Asymmetry	32
5.15.3. Multicast Services	33
6. Conclusion	33
7. IANA Considerations	33
8. Security Considerations	33
9. Contributors and Acknowledgements	34
10. References	34
10.1. Normative References	34
10.2. Informative References	34
Authors' Addresses	36

1. Introduction

As network service providers move towards deploying IPv6 and IPv4 dual stack networks, and further on towards IPv6 only networks, a problem arises in terms of supporting residual IPv4 services, over an infrastructure geared for IPv6-only operations, and doing so in the context of IPv4 address depletion. This class of problem is referred to by the draft as the 4via6 problem, for which a stateless solution is desired driven by motivation as documented in [I-D.operators-software-stateless-4v6-motivation]. Solutions such as a 4rd [I-D.despres-software-4rd], [I-D.murakami-software-4v6-translation], and [I-D.xli-behave-divi-pd], as well as dIVI [I-D.xli-behave-divi] offer such stateless solutions, by using fully distributed NAT44 functionality located on end user CPEs, which allows the network operators' core to remain effectively stateless in terms of NAT44. The solutions, collectively called Stateless4V6, rely on the same IPv4 address being used by multiple CPEs, each with a different TCP/UDP port range, and are derived from the Address+Port (A+P) solution space [I-D.ymbk-aplusp]. Differences between the solutions come down to the mode of transport (translation or mapped tunneling), and the mapping algorithm used. This document looks at the issues that have been claimed as applying to A+P technology, in the specific context of the referenced solutions, and also analyzes the two modes of transport.

2. Terminology

Stateless4V6 domain: A domain is composed out of an arbitrary number of 4V6 CE and Gateway nodes that share a mapping relationship between an operator assigned IPv6 prefix and one or more IPv4 subnets along with all the applicable TCP/UDP ports, all mapped into the IPv6 address space. An 4V6 system can have multiple domains.

Stateless4V6 CE: A CPE node that implements 4V6 functionality including NAT44 which is provisioned by means of 4V6. The device interfaces to the SP network using native IPv6 and a IPv4-IPv6 adaptation service.

Stateless4V6 Gateway A Service Provider node that implements the stateless 46 adaptation functionality for interfacing between the SP's IPv6 domain and an IPv4 domain in delivering end user IPv4 connectivity beyond the domain.

IPv4 Address sharing The notion of attributing the same IPv4 address by multiple CEs in an 4V6 domain.

Port-set: A set composed of unique TCP/UDP ports (ranges) associated to a IPv4 address. A single 4V6 CE is expected to have a single port-set for each IPv4 address.

Port-set-id: A numeric identifier of a given port set that is unique in a given 4V6 domain. A port-set-id is used to algorithmically determine the port-set members. The port-set-id is conveyed to CEs as part the CE's IPv6 addressing information, ie it is part of IPv6 subnet or address of a given CE, and its format places no restriction on the use of SLAAC or DHCP addressing.

CE-index: A numeric value, composed of a full or partial IPv4 address and optionally a port-set-id, which uniquely identifies a given CE in an 4V6 domain.

3. Stateless 4V6 Technical and Architectural Overview

This section presents the architectural and technical overview of a stateless 4v6 solution, and evidenced in whole or in part by various stateless 4via6 solution proposals such as 4rd, dIVI. Figure 1 depicts the overall architecture with two IPv4 user networks connected via 4via6 CPEs that share an IPv4 address. The goal of the system is to allow IPv4 user connectivity to the Public IPv4 network, across an operator's IPv6 network.

A key characteristic of the system, and a major differentiator with respect to previous solutions, is that translation state is only (ever) present on the CE, with the rest of the system performing stateless transport. This stateless transport applies to both the mapped-tunnel and translated modes, as described in the dedicated sections.

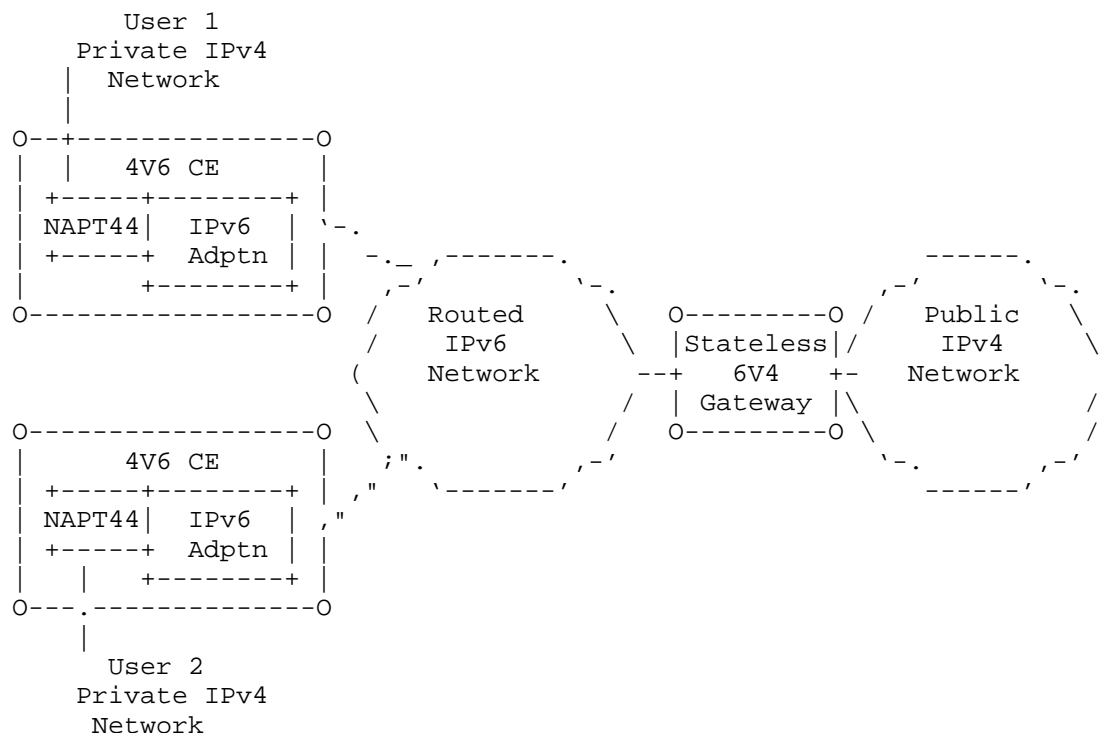


Figure 1 - Generalized Stateless 4V6 system

On IPv4 network user side, the routed IPv6 service provider network is demarcated with a 4V6 CE. The CPE externally has only a native IPv6 interface to the SP network, and a native IPv4 interface towards the end user network.

The IPv4 Internet is demarcated from the operator IPv6 network with one or more operator managed stateless 6V4 gateways that contain an IPv6 adaptation function (not detailed in the diagram) matching the one in the CE. Note: The stateless 6v4 gateway can be integrated into any existing network element (eg a core router, or an IP Edge).

Internally, the 4V6 CE is modelled as having a port restricted NAPT44 function coupled with a stateless IPv6 adaptation function that is able to ferry the end-user's IPv4 traffic across the IPv6 network, besides deriving 4V6 provisioning info from it. The NAPT44 function derives its IPv4 address, which may be shared with that of other users, and its unique Layer 4 (TCP/UDP) port range from the IPv6 address/prefix by means of an 4V6 algorithm and a port indexing schema. Any IPv4 ALG functionality that the CPE may support, remain unaffected. The CPE is expected to act as a DNS resolver proxy, using native DNS over IPv6 to the SP network.

Two forms of the IPv6 adaptation function are: i) 4v6 stateless tunneling ii) 4v6 stateless translation, each described in further in this document.

The service provider is assumed to be operating all the necessary provisioning and accounting infrastructure to support a regular IPv6 deployment. Similarly, the network operator is assumed to have the ability to assign an IPv6 prefix or IPv6 address to a CPE, and log such an address assignment.

End user host's DO NOT implement any of the 4V6, or other address sharing technologies, nor are they addressed directly with a shared IPv4 address. End user IPv4 hosts connected to the CPE receive unique private addresses assigned by the CPE, and it is the CPE that is directly addressed by the shared IPv4 address.

Although tangential to the discussion of stateless 4V6, it is useful to note that the CPE is expected to have a native IPv6 interface to the end user network, with any of the end user IPv6 hosts (single or dual stack) receiving IPv6 addresses from an IPv6 delegated prefix issued to the CPE.

3.1. IPv4 address and algorithmic port indexing

At the heart of the 4V6 solution, irrespective of mode of transport, lies the algorithm described in the specific solution drafts that allows the mapping of a shared IPv4 address and a TCP/UDP given port-set to a single IPv6 prefix or address. Notably, the 4V6 system allows both the shared IPv4 address use, as well as full non-shared IPv4 address use, all subject to the 4V6 domain configuration.

The S46 domain information required to compute the IPv4 address and correct port set is retrieved from the 4V6 prefix advertised to the CE, and pre-configured or statelessly acquired domain information.

3.2. 4V6 CE IPv6 Address and domain info

As presented in Section 2, IPv6 address of an 4V6 CE is composed out of the SP advertised IPv6 4V6 prefix, containing the CE-index, and an algorithmically computed appendix to complete the 128-bit address. This IPv6 address is *in addition* to any other IPv6 interface address that the CE configures or is configured with, including a SLAAC address from the 4V6 prefix or any IPv6 address source. One characteristics of the resulting IPv6 prefix or address is that it is for all intents and purposes a regular IPv6 prefix address that can be assigned to any regular IPv6 host.

The IPv6 4V6 interface is reserved for the 4V6 application and the

4V6 IPv6 adaptation function will exclusively use this IPv6 address. This is because the 4V6 system supports stateless communication between the 4V6 CE and the 4V6 gateway only by means of packets sent to/from this address.

3.3. IPv6 Adaptation Function

The IPv6 adaptation function plays a key role in the 4V6 system, in statelessly allowing the IPv4 user payload to be transported across an IPv6 (only) network. Two modes of such a function are currently proposed and presented in the following subsections

3.3.1. 4V6 Stateless Tunneling Mode

This type of IPv6 adaptation function is adopted and described in [I-D.despres-softwire-4rd].

The 4V6 gateway operates in the IPv4->IPv6 direction by mapping all or part of the IPv4 destination address and the port Index derived from the UDP/TCP payload into an IPv6 CE destination address. The resulting packet is sent using IPv4inIPv6 encapsulation to the CE, sourced from the 4V6's gateway IPv6 address, where the original IPv4 packet is extracted and passed to the stateful NAPT44 function.

The 4V6 CE operates in the IPv4->IPv6 direction, for traffic bound to the IPv4 internet, by encapsulating the IPv4 packet in an IPv6 header using IPv4inIPv6 encapsulation, and sending the resulting packet to the (well known) unicast address of the 4V6 gateway. There the IPv4 packet is extracted and forwarded.

The the original IPv4 packet addressing information is only partially visible on the IPv6 data plane, and the original Layer 4 information is only visible as part of the encapsulated IPv4 payload packet.

The figure below illustrates the CE model of a 4v6 Mapped Tunnel mode.

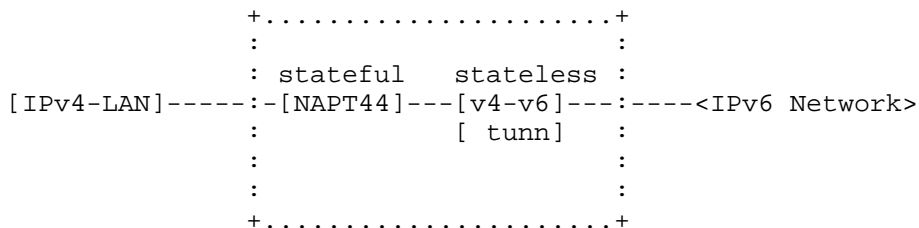


Figure 2 - 4v6 CE model with Tunnel mode

3.3.2. 4V6 Stateless Translation mode

This type of IPv6 adaptation function is adopted and described in [I-D.murakami-software-4v6-translation], I-D.xli-behave-divi-pd, and [I-D.xli-behave-divi]. The 4V6 translation mode transport operates by means of stateless NAT46 [RFC6145] extended to map the the TCP/UDP port index algorithmically derived from received IPv4 packets into an IPv6 address suffix, in the IPv6 header, besides the full IPv4 mapped representation of the original IPv4 address information. The resulting packet is then sent across the IPv6 domain as an IPv6 packet - this IPv6 packet, besides mapping the original original IPv4 address information into a determinate IPv6 format, also places the Layer 4 and packet content directly after the IPv6 header, as any regular IPv6 with TCP/UDP packet. This IPv6 packet is thus capable of being processed by regular IPv6 network elements or servers in the IPv6 domain. At either end of the IPv6 domain, the IPv4 packet header is statelessly recreated, by the 4v6 CE or gateway, again using exactly the same NAT64 process as in [RFC6145].

The figure below illustrates the IPv6 4v6 Stateless Translation model of a 4v6 CE.

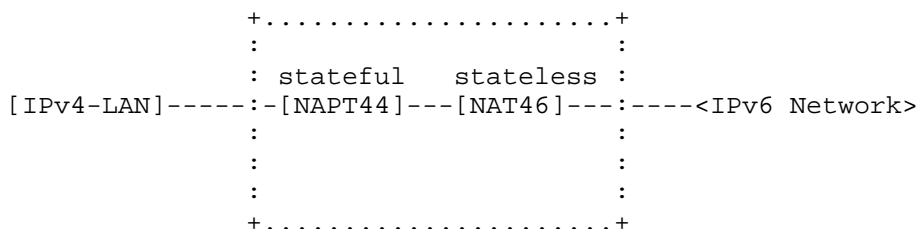


Figure 3 - 4v6 CE model with stateless NAT64

4. Comparison of 4V6 transport modes

This section presents the an overview of the similarities and differences between an IPv4-IPv6 translation based 4V6 transport mode and one that utilizes IPv4-in-IPv6 tunnelling. The comparison takes into consideration a wider deployment view composed of functionality that is known to be in common use today.

4.1. General Characteristics of 4V6 modes

The following table presents a comparison of the 4V6 transport modes, in terms of the base technology, and constrains, including also IPv4.

Item	4V6 Translation mode	4V6 Tunnel Mode
Base Technology	Port restricted NAPT44 with modified stateless NAT64 on CPE and Gateway	Port restricted NAPT44 with IPv4 in IPv6 mapped tunneling on CPE and Gateway
Location of stateful NAPT44 function	CPE	CPE
IPv4 Forwarding paradigm	L3 + L4 lookup	L3 + L4 lookup
IPv6 Addressing Constraints	CE uses 4V6 suffix.	CE uses 4V6 suffix.
Type of IPv6 prefix/address announcement method supported	ICMPv6 (SLAAC), DHCPv6 (both IA_NA and IA_PD)	ICMPv6 (SLAAC), DHCPv6 (both IA_NA and IA_PD)
Can the 4V6 IPv6 prefix be used by non 4V6 devices?	Yes	Yes
IPv4 addressing constraints	Fixed sharing ratio per IPv4 address.	Fixed sharing ratio per IPv4 address.
TCP/UDP Port range constraint	Ports are statically allocated	Ports are statically allocated
Requires ALG64 or DNS64	No	No
Requires IPv6 DNS on CPE	Recommended	Recommended
4V6 CE Parameter provisioning methods (assuming suitable protocol extensions)	ICMPv6, Stateless DHCPv6, TR69	ICMPv6, Stateless DHCPv6, TR69.

IPv6 Domain Routing to CE based on:	Regular closest IP match to CE-IPv6 subnet	Regular closest IP match to CE-IPv6 subnet
-----	-----	-----
IPv6 Domain Routing to 4V6 Gateway based on	IPv6 4V6 domain aggregate route	4V6 Gateway unicast/anycast address
-----	-----	-----
IPv4 Header Checksum recalculation required	Yes	No
-----	-----	-----
Supports non TCP/UDP Protocols	No*	No*
-----	-----	-----
ICMPv4 Limitations	No ICMPv4 from "outside the domain". Internal ICMPv4-v6 translation as per [RFC6145]	No ICMPv4 from "outside the domain".
-----	-----	-----
ICMPv5 identifier NAT/Markup needed	Yes	Yes
-----	-----	-----
Supports IPv4 fragmentation (without additional state)	No	No
-----	-----	-----
Requires IPv6 PMTU discovery/configuration	Yes	Yes
-----	-----	-----
Supports IPv4 Header Options	No - as per NAT64 [RFC6145]	Yes (use of source route option is constrained)
-----	-----	-----
TCP/UDP Checksum recalculation	Yes - depending on suffix, as per NAT64 [RFC6145]	No
-----	-----	-----
Supports UDP null checksum	Yes/Configurable - as per NAT64 [RFC6145]	Yes
-----	-----	-----
Transparency to DF bit	Yes, configurable as per [RFC6145]	Yes
-----	-----	-----

Supports IPv4 Fragmentation	Partial (no fragments from outside the domain)	Partial (no fragments from outside the domain)
-----	-----	-----
Transparency to IPv4 TOS	Yes, configurable as per [RFC6145]	Yes
-----	-----	-----
Overhead in relation to original average payload on IPv6 of a) ~550 bytes b) 1400 bytes).	a) 0% b) 0%	a) 4.36% b) 1.71%
-----	-----	-----
Supports non-shared IPv4 usage (ie whole IPv4 address assignment to a single device)	Yes	Yes
-----	-----	-----
Can support IPv4 to IPv6 host communication (for traffic not requiring ALGs)	Yes - As per [RFC6145] stateless NAT64 specification	No
-----	-----	-----
Changes to network element provisioning tool(s)**	Yes - Mapping IPv4 to IPv6 addresses	Yes - Enabling IPv4inIPv6 functionality
-----	-----	-----

* Without specific ALGs. Non UDP/TCP protocols, like ICMP, can be supported with specific ALGs.

**Network (feature) provisioning tools/applications need to be 4V6 aware. With the translation technique, the tool needs to enable the operator to map IPv4 addresses to IPv6 addresses as dictated by the 4V6 domain. With the tunneling technique, the tool needs to allow the operator to enable IPv4 (inIPv6) functionality and modify its characteristics.

4.2. Mobile SP Architecture and 4V6 Applicability

This section presents the applicability and comparison of the 4V6 modes to current 3GPP architectures used by Mobile SP for delivering all sorts of mobile services.

4.2.1. 3GPP overview

The 3rd Generation Partnership Project (3GPP) is a collaboration between groups of telecommunications associations, whose scope is to develop a globally applicable mobile phone systems and architectures based on service requirements. 3GPP standards are structured as Releases, each of which incorporates numerous individual standard documents. Currently, 3GPP Release 7 is the latest release in common practical deployment, with Release 8 being readied for deployment. Releases 9 and 10 are finalized, and work is underway on Release 11.

One of the major service requirement drivers of recent and ongoing 3GPP releases is the realization of services that deliver specific QoS, or user charging goals, all based on a policy system (eg tiered data rate or volume plans). Technically this translates to the Policy and Charging Control (PCC) framework, which in turn attributes specific functionality to nodes in the 3GPP architecture, such as the PDN-Gw and the PCRF. This functionality comprises both data-plane features (eg IP flow classification) as well as the interfaces/protocols between nodes (eg Diameter, and its specific 3GPP applications).

The 3GPP specifications allow both IPv4 and IPv6 traffic to be handled, and subject to operator defined handling and charging policies by means of applying suitable user traffic filters. Such filters are currently defined to be either IPv4 or IPv6, are applicable to user plane traffic, and are used in a variety of critical roles including the signalling of PDP contexts/EPC Bearers, as well as PCC signalling and interaction with applications.

The following table illustrates the impact of the 4V6 translation and tunnel transport modes respectively on the 3GPP architecture including PCC interfaces. In assessing the impact of these 4V6 transport modes a number of additional assumptions are taken:

- o The 3GPP system supports native IPv6 user traffic, as say per either of the E-UTRAN Release 8 or 9 specifications, using the relevant EPS bearer or PDP functionality.
- o The 4V6 gateway functionality is not part of the 3GPP core architecture (given that currently it is not scoped by a 3GPP Release). Instead, the 4V6 gateway is taken to be a stand alone component in the 3GPP network operator's core reachable via the SGi interface.

The above system, in the context of 3GPPs E-UTRAN architecture as defined in [E-UTRAN] is shown in Figure 2

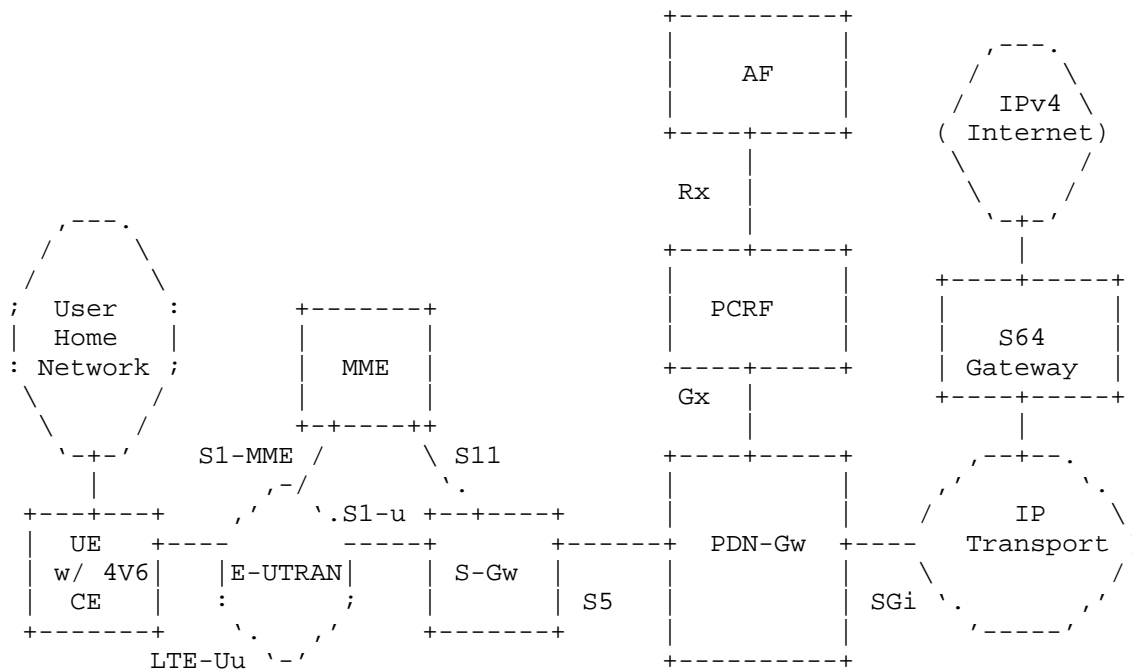


Figure 2 - 3GPP Architecture with 4V6

The main 3GPP system components, and terms are summarized as follows (the reader is referred to [E-UTRAN for a more detailed definition]:

UE The User Equipment, typically a phone or a 3G/4G capable Home Router (shown to incorporate 4V6 functionality)

E-UTRAN Evolved Universal Terrestrial Radio Access Network. The Radio Access network, composed on E-NodeB elements.

MME Mobility Management Entity. Responsible for user authentication, PDN/SGw selection. Does not interact with the user data plane

S-Gw Serving Gateway (function). Responsible for handling local mobility, (some) traffic accounting, traffic forwarding, bearer establishment.

PDN-Gw Packet Data Network Gateway (function). Responsible for per user IP traffic handling, incl. address assignment, filtering, QoS, accounting.

PCRF Policy And Charging Rules Function. Responsible for authorizing and applying policy rules, as well as binding them to user bearers.

Bearer The bearer represents a virtual connection, typically that between a UE and a PDN-Gw. The bearer is specified as an IP Filter (in terms of IP address, port numbers) and is the object of policy rules. 3GPP, depending on Release and document, defines many terms that are used to refer to the same notion: PDP context, EPS Bearer.

AF Application Function. A functional element offering (higher level) applications that require dynamic policy and/or charging control over the user plane (bearer) behaviour. The AF can be seen as bridging the gap between applications and how they affect the IP data plane of a user.

S5 It provides user plane tunnelling and tunnel management between SGW and PDN-GW, using GTP or PMIPv6 as the network based mobility management protocol.

S1-u Provides user plane tunnelling and inter eNodeB path switching during handover between eNodeB and SGW, using the GTP-U protocol

SGi It is the interface between the PDN-GW and the packet data network. Packet data network may be an operator external public or private packet data network or an intra operator packet data network.

Gx Bearer and flow control interface between the user data-plane element (PDN-Gw) and the Policy System. A Diameter based interface with a suite of 3GPP applications

4.2.2. 3GPP and 4V6 modes

4V6 translated traffic appears for all intents and purposes as regular IPv6-user traffic to the 3GPP system and packet processing functions (eg the PDN-Gw). Hence, and based on the stated assumptions, any such 4V6 traffic can be handled using existing native IPv6 functionality defined by the core 3GPP specifications.

In contrast, 4V6 tunneled traffic requires additional data plane processing to get to the "real" user IPv4 payload and apply the desired functions. Such additional processing is currently not part of the functionality covered by the 3GPP specifications. In view of this, and solely in relation to the 4V6 tunnel transport mode, two alternative hypotheses need to be placed in order to complete the comparison

i) that such IPv4 in IPv6 processing functionality will be supported as part of the existing EPS bearer functionality defined in E-UTRAN, perhaps as a dedicated EPS bearer (ie an additional virtual interface per subscriber). Or, that;

ii) a new 46 EPS bearer type (ie interface type) identification and signalling will be defined by the 3GPP architecture, which formalizes the v4inv6 relationship between the IPv4-user payload and the v6-user layers.

An apparent benefit of approach (ii) would be in allowing the system to clearly distinguish and expose to other systems v4-user traffic versus v6-user traffic, which is composed of v4inv6 and regular v6 traffic that a UE may generate. The former approach (i) is more convoluted given the ambiguity in distinguishing, and representing such a combination of v6-user and v6-user-bearer and v4-user traffic, all while keeping coherence in terms of the policy system. These two options are designated with ** in the table below.

Item	4V6 Translation Mode	4V6 Mapped Tunnel Mode
User Data Plane at the PDN-Gw (as per section 5.1.2 in [EUTRAN])	IPv6 over GTP-U over UDP over IP	IPv4 over IPv6 over GTP-U over UDP over IP
Gx (Diameter)	No discernible impact	Impacted: no way to express v4 over v6 in TFT Filter and Flow Descriptors
Rx (Diameter)	No discernible impact	Impacted: no way to express v4 over v6 in Media-Component-Description and, Flow-Description-AVP
S5 (GTP)	No impact	Impacted with new PDP/EPS Bearer type*
New 46 Bearer definition	Not required	Possibly required**

Secondary interface (dedicated bearer or secondary PDP) for 4G traffic	Not required	Possibly required**
-----	-----	-----
PDN-Gw	No impact	New TFT capability, IP Gate functionality, changes to Gx, and likely changes to S5/S7 related to signalling the new bearer
-----	-----	-----
SGW	No Impact	No discernible impact
-----	-----	-----
PCRF	No impact for IPv6. Feature to map IPv4-IPv6 addresses needed only in case of IPv4-only applications.	Impacted for both IPv6 and IPv4-only applications and Gx applications utilizing flow control/charging
-----	-----	-----
AF Application Function	No discernible impact	Flow based application control impacted
-----	-----	-----
UE	4V6 application	4V6 application
-----	-----	-----
LTE-Uu	No discernible impact	Likely changes required to support signalling of EPS bearer or PDP type
-----	-----	-----
Lawful Intercept	No discernible impact	New rules for tunnel support
-----	-----	-----

*A new PDP Type or EPS bearer signalling has a broader 3GPP system wide impact not fully covered here.

As the table illustrates, the 4V6 tunnel transport model appears to affect a significant number of 3GPP elements, when the intent is to realize a full suite of services. This observation appears to apply to any other carrier inserted tunneling technology (eg DS-lite). Hence, a substantial investment in 3GPP standard terms and in the evolution of deployed systems appears to be required.

In contrast the 4V6 translation mode bears none to no discernible impact on existing 3GPP Release 8/9 specifications and their deployments, while allowing the operator to realize the full set of services on 4V6, alongside any native IPv6 traffic, allowed for by these architecture. Hence, little beyond the addition of 4V6 components operating using translation mode appears to be required.

4.3. Cable SP Architectures & 4V6 Applicability

Cable SPs (commonly referred to as Multi System Operators (MSOs)) usually deliver video, data, and voice service over the cable and fiber access to residential and commercial customers. Many MSOs offer SLAs with various services by exploiting QoS not only in their IP/MPLS network, but also their access network.

The cable access network (now synonymous with Hybrid Fiber Coax (HFC)) is commonly enabled with Data Over Cable Service Interface Specifications (DOCSIS, a CableLabs standard) to facilitate the implementation of packet based services. In this paradigm, the HFC/DOCSIS access bandwidth is typically shared among a number of customers, hence, ensuring optimal service quality & experience per customer becomes extremely important for MSOs' success.

Cable SPs/MSOs ensure the optimal service quality of various advanced & real-time multimedia services (such as IP telephony, multimedia conferencing, interactive gaming etc.) by utilizing "PacketCable" framework to enforce QoS on the HFC/DOCSIS access.

The next sub-section 4.3.1 provides a brief introduction to PacketCable, section 4.3.2 explains a key PacketCable construct - Classifier, and section 4.3.3 tabulates the impact of 4V6 modes to PacketCable enabled DOCSIS/IP services.

4.3.1. PacketCable Introduction

PacketCable, a CableLabs standard, defines a framework for ensuring the Quality of Service (QoS) on the HFC/DOCSIS Access. PacketCable specifications (e.g. PacketCable 1.0, PacketCable Multi Media [PCMM], PacketCable Dynamic QoS [PC-DQOS], PacketCable 2.0) specify interoperable interface specifications for executing QoS, Admission Control, Accounting, Policy, and Security functions on Cable Modem (CM) and Cable Modem Termination System (CMTS), as/when needed. They all require DOCSIS 1.1 or later versions.

The PacketCable framework is also critically important for MSOs to comply with government regulations for things such as E911 when they offer voice/telephony services, Lawful Intercept (LI) etc.

The figure below illustrates one of PacketCable variants i.e. PCMM [PCMM] architecture, as an example, that defines a set of IP-based interfaces (referred to as pkt-mm-1 through 12) pertaining to core QoS and policy management capabilities.

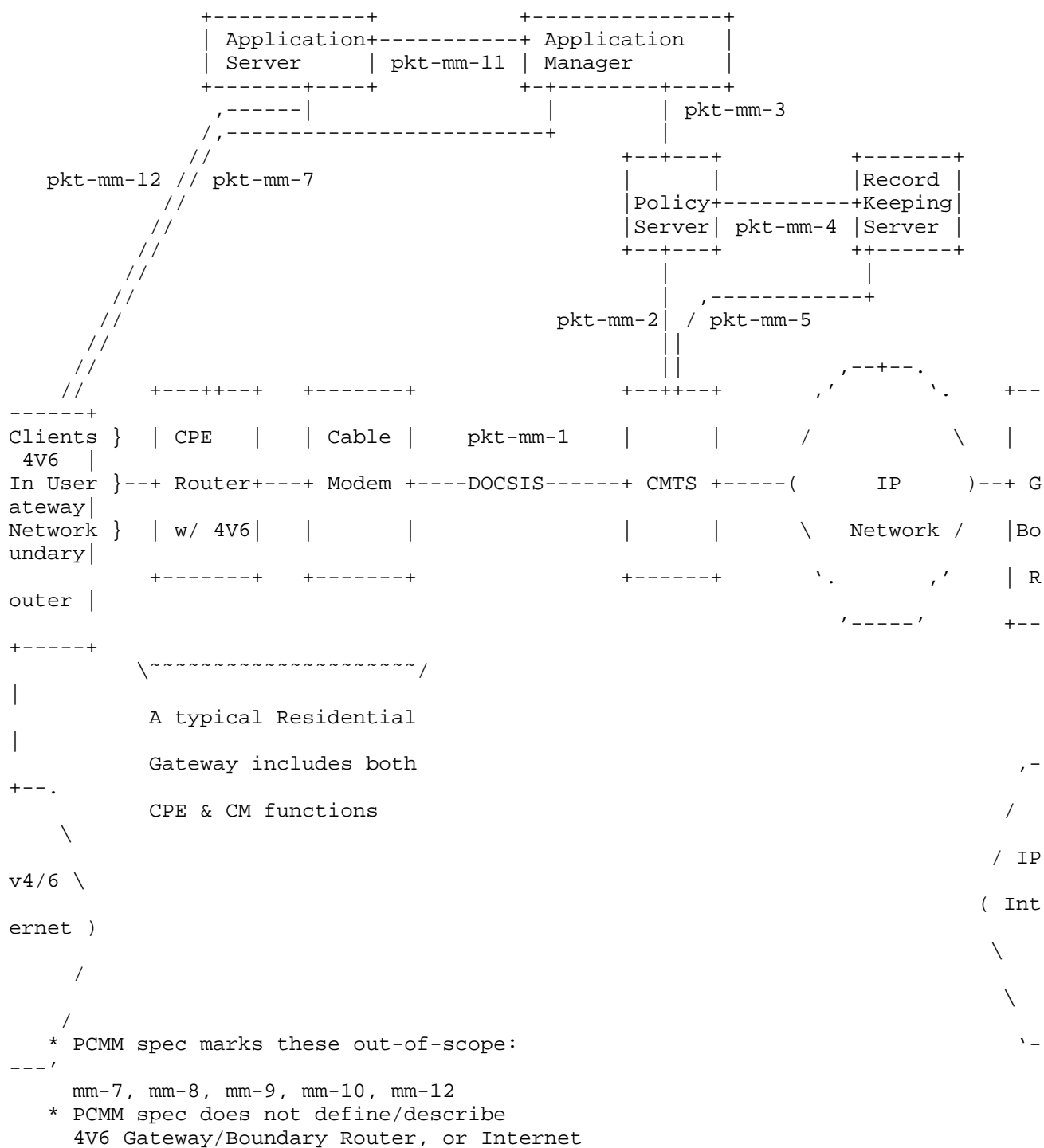


Figure 3 - PacketCable Multimedia Architecture (with 4V6)

Dec, et al.

Expires April 16, 2012

[Page 19]

4.3.2. PacketCable Construct - Classifier

PacketCable framework fundamentally relies on Cable Modem (CM) and Cable Modem Termination System (CMTS) to first qualify and then classify the appropriate IP traffic between them, for effective QoS enforcement. The framework requires the usage of "Classifier" for both qualification (in control plane) and classification (in data plane).

Taking PCMM specification [PCMM] again as an example, PCMM mandates the usage of classifier in the control plane (i.e. 'Upstream Packet Classification Encoding' in pkt-mm-1 interface (DOCSIS) , whereas 'Multimedia Classifier Object' in pkt-mm-2 and pkt-mm-3 interfaces (COPS)) for conveying the attributes of an IP flow belonging to an application (telephony, say), and subsequently its usage in the data plane i.e. filter matching on the IP packets' layer2/3/4 headers prior to QoS treatment.

The PCMM specification mandates the 'classifier' to include Source and Destination IP addresses, DSCP/TOS, IP Protocol, Source and Destination ports for an IPv4 traffic flow received by the CMTS, and similarly, Source and Destination IP addresses, TC, Next Header, Source and Destination ports for an IPv6 traffic flow received by the CMTS.

Similar to PCMM, PacketCable DQOS specification [PC-DQOS] also mandates the usage of classifier in the control plane (DSx messaging). In particular, PC-DQOS mandates the classifier definition to have 'protocol' (or next header) in IP header to be 17 (=UDP) along with specific Source and Destination ports (and Source and Destination IP addresses, optionally) so as to accommodate voice RTPoUDPoIP traffic.

In summary, the CMTS (and CM) construct their data-plane filter based on the 'classifier' information.

4.3.3. 4V6 Modes Impact on PacketCable

In 4V6 Tunnel mode, the 4V6 tunneled traffic requires additional data plane processing to get to the "real" user IPv4 payload and apply the desired functions. Such additional processing is currently not part of the functionality covered by the PacketCable specifications, nor part of compliant implementations.

In 4V6 Translation Mode, the 4V6 translated traffic appears for all intents and purposes as regular IPv6-user traffic to the PacketCable framework (both control plane and data plane). Hence, it is likely that any such 4V6 traffic can be handled using native IPv6

functionality e.g. classifier as defined by the PacketCable specifications and supported by CMTS and CM.

Taking PCMM specification as an example, it is worth noting that PCMM already allows for (and mandates) a minimum of four classifiers to be included in Gate-set. Hence, a Policy Server can communicate (via pkt-mm-2) both IPv4 and IPv6 classifier to the CMTS, which can use IPv6 classifier for constructing its data-plane filters (for DownStream processing), and convey IPv4 classifier to the CM via DOCSIS messages (pkt-mm-1) for any Upstream Processing. So, the 4V6 Translation Mode would work out in current implementations/deployment reasonably well.

Separately, it is likely that the CPE Router would be engaged in serving IPv4 multicast content to IPv6 receivers (and vice versa) in future, requiring 'translation' function.

In summary, while 4V6 Translation mode can work with the existing PacketCable framework, 4V6 Tunnel mode can not.

5. Overview of potential issues and discussion

This section summarizes the issues attributed to an A+P, or port restricted scheme, along with a discussion of applicability to the assumed system and possible resolutions. The summary of issues stem from [I-D.thaler-port-restricted-ip-issues] and associated discussions.

5.1. Notion of Unicast Address

5.1.1. Overview

The issue, referred to as the "definition of a unicast address", relates to the notion that in a shared IPv4 address system, multiple hosts will be visible as having a single IPv4 address outside of the system. This issue is a general characteristic of any NAT44 based solution [I-D.ietf-intarea-shared-addressing-issues], including DS-Lite. However, a more specific aspect of this issue in the context of an address sharing system is the possibility that a single host having multiple interfaces will be assigned the same IPv4 address (with different port ranges) on each of its interfaces. It may also be that multiple hosts sharing an address find themselves on the same Layer 2 segment. Either would impede hosts from working within the notion of known host IP stack and protocol implementations.

5.1.2. Discussion

A number of the characteristics of the 4via6 solution architecture cause the issues not to be applicable, key of which is that there is no expectation for any kind of end hosts to be part of the shared IPv4 address system.

In the stateless 4via6 system, CPE nodes are assigned with a shared IPv4 address+port range by means of the unique IPv6 address, containing the embedded IPv4 address + port index, of that CPE node. The CPE node is in addition enabled to be running the port restricted NAPT44 function from the IPv6 derived address, a key characteristic of the solution. On the IPv6 plane, the IPv6 address of the CPE is practically indistinguishable from any "regular" IPv6 address, and in fact any host that is not aware of it conveying an embedded IPv4 address would be able to use this just like any other regular IPv6 address, ie the 4via6 solution uses standard IPv6 addressing. In terms of the IPv4 dimension, since the shared address and port index are never used to address native IPv4 nodes or hosts, but instead uniquely assigned to a single NAPT44 function that is part of the CPEs, all legacy or other IPv4 hosts are not exposed to the presented issues.

Going beyond the ascribed issue however, it appears desirable to have the 4via6 CPEs that are to be part of the shared system to be able to provide a hint to the network operator in terms of their special capability. Such a hint can be a DHCPv6 Option Request Option, which would be useful to allow the DHCPv6 sub-system to also inform the CPE of any other stateless 4via6 system parameters. A largely similar ORO option is currently being defined as part of [I-D.ietf-software-ds-lite-tunnel-option]

Recommendation: Define a suitable DHCPv6 ORO for conveying the 4via6 capability of a CPE.

5.2. Implementation on hosts

5.2.1. Overview

The issue, as presented, relates to the need for modifications on end hosts or devices to support a port constrained mechanism and the overall impossibility of realizing such modifications. Furthermore, host applications that attempt to bind to specific ports that are not part of the allowed port range will fail to do so and may also require modifications.

5.2.2. Discussion

As presented in Section 3 the solution assumes the use of a dedicated CPE implementing the 4via6 functionality within a port constrained mode and NAPT44. Granted, CPE nodes will require to implement new functionality such as the IPv6 adaptation function, that is likely alongside introducing native IPv6 support. However, any and all existing end user IPv4 devices (eg PCs, etc) will not be affected. Nor are such devices expected to behave in any way different from that of today, where they typically obtain a private rfc1918 address and multiplexed by a CPE using a NAPT44 function.

In summary, the assumed 4via6 solution requires a specific 4via6 CPE but does not require any IPv4 host stack changes.

5.3. 4V6 address and impact on other IPv6 hosts

5.3.1. Overview

The issue relates to the question of whether the operation of a regular IPv6, non 4V6 capable, host would be adversely impacted should it be assigned or auto-configured with an address from an S64 address or prefix pool.

5.3.2. Discussion

The 4V6 prefix is for all intents and purposes a regular IPv6 prefix, and as such can be announced/assigned to any IPv6 host which in turn can use derived addresses like any other IPv6 address. Thus, an 4V6 IPv6 domain can address non-4V6 devices, leaving such devices to operate as native IPv6.

There is however a restriction on the 4V6 CE devices. As described in Section 2, a 4V6 CE constructs itself the full 128 bit address from the concatenation of the IPv6 prefix, 4V6 domain information acquired statelessly, and a pre-determined or algorithmic interface-id. By definition, only one 4V6 CE can use the same IPv4 address and port index. Hence, while there is no exact limitation on the number of non 4V6 hosts that can be addressed from an 4V6 prefix, there is a limit of one 4V6 CE per 4V6 prefix. Using a 4V6 prefix to address network segments without 4V6 devices does diminish the efficiency of the IPv4 address sharing mechanism, in terms of using up port ranges onto segments that will not use them. This is naturally a deployment consideration which an operator can optimize.

5.4. Impact on 4V6 CE based applications

5.4.1. Overview

It has been claimed that applications implemented on the CE itself, eg a DNS resolver-client, may be impacted by the 4V6 functionality. In particular, a concern is that such applications would either need to be specially engineered to issue socket calls or extensive IP stack modifications made to support them.

5.4.2. Discussion

By definition the 4V6 CE is an IPv6 capable device, and any IPv6 capable applications will be able to use the native IPv6 stack (note: IPv6 interface selection, is discussed in section 5.5). As such, the concern raised does not apply to applications that can be expected to support IPv6, and instead only to IPv4-only applications running on the 4V6 CE.

The shared IPv4 address is intended to be used only by the 4V6 CE function. This shared IPv4 address does not need to be assigned to an interface on the 4V6 CE and thus a target for potential applications. Any such applications running on the 4V6 will use any of the other (likely private) IPv4 address on the CE, which then will be routed to the 4V6 function this is applied post routing for the packets generated by these applications.

5.5. 4V6 interface

5.5.1. Overview

A 4V6 CE will have a "self configured" 4V6 IPv6 interface address, alongside any other SLAAC or DHCPv6 derived addresses, potentially from the same prefix. This particular 4V6 address may be subject to specific filtering rules or restrictions by the operator, besides usage and filtering restrictions on the 4V6 CE. Also, for the 4V6 system to operate as intended, the 4V6 application on the CE must be restricted to using the specific 4V6 address when sourcing 4V6 packets. Also, the 4V6 CE needs to be set-up to correctly forward IPv4 traffic to the 4V6 application.

5.5.2. Discussion

While the method of creating the interface is implementation specific, the generic operating model that is envisaged is for the 4V6 application to create the 4V6 interface as a virtual interface with an IPv4 unnumbered address. The application would then install a default IPv4 route pointing to this virtual interface, which would

be effectively see the 4V6 application acting as a network appliance on the forwarded traffic. In terms of IPv6 behaviour, the 4V6 application is expected to be set up to specify the use (binding) to the 4v6 IPv6 virtual interface.

5.6. Non TCP/UDP port based IP protocols - ICMP)

5.6.1. Overview

This issue relates to the inability of using regular ICMP messages to "ping" an end-host that has been addressed with a shared IPv4 address. The issue can be generalized one applicable to any IP protocol that is not TCP/UDP port based, and also in terms of the ability of using such protocols from end hosts that are assigned a shared IPv4 address.

5.6.2. Discussion

The inability to ping a CPE from the IPv4 Internet is shared by other IPv4 address sharing mechanisms such as DS-Lite. Thus, the issue is no better or worse in the case of the stateless 4via6 solution. The same can be said of end user hosts using other non UDP/TCP port based protocols from behind a NAT44 function, ie they will not function irrespective of address sharing or not.

As discussed in [I-D.ietf-intarea-shared-addressing-issues], all IP address sharing solutions break protocols which do not use transport numbers. A mitigation solution is to utilize specific ALGs. For ICMP in particular, a mitigation solution would be to rewrite the "Identifier" and perhaps "Sequence Number" fields in the ICMP request, treating them as if they were port numbers.

As a conclusion, this issue can be partially mitigated, likely at par to centralized NAT solutions.

5.7. Provisioning and Operational Systems

5.7.1. Overview

The general claim of this issue is that a service providers' provisioning and accounting systems would need to [radically] evolve to deal with the notions of shared IPv4 addresses and port range constraints.

5.7.2. Discussion

The stateless 4via6 solution relies on a fully operational IPv6 network, which on the IPv6 plane fundamentally does not differ from a

regular IPv6 network, and the stateless 4via6 solution may be seen as an IPv6 application - devices connecting to the network, need unique IPv6 addresses which the network is able to provide. In the 4via6 solution it happens that these unique IPv6 addresses embed an IPv4 address. Hence, additional system enhancements that the stateless 4via6 solution requires, over and above those simply needed to deploy and operate an IPv6 network, lie in the domain of supporting the provisioning of the IPv6 adaptation functionality of the CPEs. This may require the operator to use DHCPv6, or other provisioning methods such as IPv6CP, TR-69, in order to configure any relevant 4via6 service parameters to a CPE.

From an IPv4 perspective, an operator will likely want to have a management system capable of the assignment of IPv4 addresses to the shared pool, and tuning the re-use factor. In this, the solution exhibits no grossly different characteristics than those of any system with an operator managed NAT44 function where similar management capabilities need to be introduced.

One additional aspect of the stateless 4via6 solution needs to be highlighted. On a par basis this solution requires less per subscriber management, accounting and logging capabilities than centralized NAPT44 alternatives such as DS-Lite, due to the following:

- o The assignment of an IPv6 address that embeds a deterministic IPv4 address and port range removes the need for the operator to perform any NAPT44 binding logging, ie the task of determining which user had a given IPv4 address and port at a given time is simply a matter of determining who had the corresponding IPv6 address, rather than collecting large amounts of dynamic binding data.
- o There is no need for the operator to manage NAPT44 binding data access and retention.
- o Given the stateless nature of the 4via6 solution, all subscriber CPEs in an operator's domain can share exactly the same 4via6 service configuration, i.e. The operator does not need to be concerned with managing on a per user basis specific AFTR assignment and/or load balancing such users and throughout ensuring symmetric traffic flows throughout.
- o The location of the NAPT44 function on the user's CPE, allows easy and direct management of the port mappings by the end user removing a need for the operator to introduce PCP [I-D.wing-software-port-control-protocol] (or similar) protocols in on AFTRs, and on CPE devices. In effect the end user can

retains control of any bindings, which could be via today's GUI, or UPnP IGDv2, or even PCP.

- o As and when needed, a stateless 4via6 solution readily supports the assignment of an unshared IPv4 address, and full port control by the end user. A similar capability with centralised NAPT44 solutions involve onerous management of per subscriber configurations on the operator's AFTR.

5.8. Training & Education

5.8.1. Overview

The issue claims a concern with the need for developers and support staff to be trained & educated in dealing with a port constrained systems.

5.8.2. Discussion

There appear to be at least two levels of looking at this issue in the stateless 4via6 context. On one level, it is perfectly true that developers and support staff will need to be trained with running/supporting a native IPv6 network, that is now a basis of the solution. This however is an inherent aspect of deploying an IPv6 network and applications. On another level, support and developers need to be familiarized with the NAPT44 characteristics of the system, that are not different from those already known about such systems today. More specifically, there appears to be no such thing as a port unconstrained carrier grade NAPT44 system, in either tomorrow's stateless 4via6 or AFTR guises, or today's residential CPE NAPT44 implementations that have an inherent hard set translation limit (often 1024 translation, corresponding to a usage of 1024 ports). That application developers should be trained to be reasonably conservative in the usage of ports is thus not an issue of the stateless 4via6 solution, but pretty much of any NAPT44 based solution, even those in use today.

Another useful observation here is that the stateless 4via6 solution, actually allows an operator to retain existing troubleshooting procedures, given which today encompass CPE based NAPT44, rather than changing them radically to an AFTR. Furthermore, it is possible to alleviate any port-range constraints for users by allocating more generous port ranges without the need to manage such users configuration on active core network devices (eg AFTR).

5.9. Security and Port Randomization

5.9.1. Overview

Preserving port randomization [RFC6056] may be more or less difficult depending on the address sharing ratio (i.e., the size of the port space assigned to a CPE). Port randomization may be more difficult to achieve with a stateless solution than stateful solution. The CPE can only randomize the ports inside be assigned a fixed port range.

5.9.2. Discussion

The difference in the random port selection range may be significant in practice and using port-restricted systems without any measures (like random port selection in draft-bajko-pripaddrassign-03) is one of the trade-offs of the mechanism. It should be however noted that even full port unrestricted systems, today, rarely implement random port selection from the full port range, as such the difference is largely theoretical, again viewed from today's perspective. Only with a longer term prospect of devices/hosts adopting random port selection according to RFC 6056 the NAT-based port-restricted mechanisms, will degrade security to a certain extent.

5.10. Unknown Failure Modes

5.10.1. Overview

The issue purports that a system with a port constraints introduces new unknown failure modes, not known with NAT44 or NAPT44 systems, and in general is more complex than such a system.

5.10.2. Discussion

This claim does not appear to have objective technical arguments that can be discussed. A restricted port range system, such as the one assumed in this document, does not appear to have any more or less complexity than any of the other NAPT44 solutions against which the same issue has not been levelled. That is a statement that can be made in consideration of each of those alternative solution network design (eg elaborate routing rules or topologies) and feature implementation complexities, which appear to be no better than that of a stateless 4via6 address port range system. Ultimately, system complexity is something best left adjudicated by the operators choosing to deploy one or the other of these IP based transition solutions.

5.11. Possible Impact on NAT66 use & design

5.11.1. Overview

The notion of a shared address with a constrained port range is seen as possibly bearing influence on use in future schemes involving NAT66, where IPv6 address sharing is in general deemed not to be desired (ie there is good reason to avoid PAT66).

5.11.2. Discussion

The authors do not propose, nor expect to see the IP address sharing characteristic applying to future NAT66/PAT66 discussions and specification. However, having said that it is useful to take a humble step back and consider the general aspect of causality in this context. The direct cause that brought about IPv4 shared address solutions to the fore was a shortage/exhaustion of a limited IPv4 address resource, alongside a failure of the community to migrate IPv4 networks to IPv6 in a timely manner. At the time of writing it is hard to imagine the same occurring with respect to IPv6 address resources, and hopefully the same set of causes will not be allowed to re-occur. This appears to be the only way to ensure that IPv6 address sharing effect does not come to be, as opposed to precluding such notions within the context of today's IPv4 world where the causality is rather clear.

5.12. Port statistical multiplexing and monetization of port space

5.12.1. Overview

An issue attributed to 4V6 solutions is that due to their characteristic of assigning a fixed amount of ports to participating system nodes, the overall pool of ports cannot be dynamically/statistically multiplexed.

A corollary of this claimed issue is the claim that port range constraints will lead to monetization by service providers of such port ranges, for example by charging users based on the number of ports assigned or creating some bronze, silver, gold type of port based service categories.

5.12.2. Discussion

The 4via6 address shared solution indeed limits the ability to "overload" ie statistically multiplex amongst users, the ports available of a given public IPv4 address. This can be seen as a trade off vs dynamic allocation and the need to log (large amounts) of NAT bindings. Furthermore, the solution is meant to be

fundamentally a transitional one for supporting legacy IPv4 users till full migration to IPv6 can occur. As an example, even with a static allocation of ~1000 ports per shared IP user, it allows an operator to effectively multiply by ~64 the current IPv4 unrealizable address space. To put it into a network growth perspective, it allows an operator to support for some 10 years a steady 50% annual increase in users, without requiring new IPv4 addresses. This is likely an alluring (if unlikely) prospect for most, but it demonstrates the fact that even with static port allocations, IPv4 address sharing can go a long way for many operators.

CGN-based solutions, because they can dynamically assign ports, provide better IPv4 address sharing ratio than stateless solutions (i.e., can share the same IP address among a larger number of customers). For Service Providers who desire an aggressive IPv4 address sharing, a CGN-based solution is more suitable than the stateless. However, in case a CGN pre-allocates port ranges, for instance to alleviate traceability complexity it also reduces its port utilization efficiency.

5.13. Readdressing

5.13.1. Overview

Due to the port range encoding being part of the CPE's IPv6 address, any change in the range requires a re-configuration of the CPE's 4via6 address. This is said to be an issue given the impact that IP address changes have on existing traffic flows, as well as general IPv6 network routing

5.13.2. Discussion

It is true that under the assumed notions of the stateless 4via6 solution, IPv6 re-addressing is required to effect a change in terms of the shared IPv4 address or ports. Such changes can and are likely best done using dynamic address configuration methods such as DHCPv6, or alternatively out of band management tools, eg TR-69, especially when the 4via6 address can be derived from a delegated prefix. Using these, the impact of the address change does not translate to a neither a classic IPv6 host renumbering problem nor an unmanageable network renumbering problem. On the CPE, the change only affects the 4via6 address of the CPE and not any end user IPv6 hosts behind the CPE (which would likely continue to derive their IPv6 addresses from an unchanged delegated prefix). On the service provider network side, the change, if any, represents a network renumbering case which the operator can be reasonably expected to handle within their network numbering plan, especially given that the IPv6-prefix of the an IPv4-in-IPv6 address is summarizable.

An addressing change will impact any existing IPv4 flows that are being NAT'ed by the CPE. This is also analogous to the today's practice of IPv4 address changes espoused by some operators, which while not being commendable, is established in the market. Nevertheless, as a means of alleviating such an impact it appears desirable for the solutions to investigate the viability of mechanisms that could allow for more graceful addressing changes.

To facilitate IPv6 summarization and operator appears to have two 4V6 deployment choices. When encoding IPv4 addresses in lower order address space bits that are subject to summarization, the operator would need to assign a modest dedicated IPv6 prefix (such as a /64) as an 4V6 IPv6 addressing sub-domain. Alternatively, without resorting to a separate 4V6 addressing sub-domain, an operator could allow for the IPv4 address embedding to be embedded in a high-order portion of the IPv6 domain address space, one that closely follows the IPv6 domain prefix. These two valid address subnetting and deployment options deserve better description in the solution specifications.

5.14. Ambiguity about communication between devices sharing an IP address.

5.14.1. Overview

A regular IPv4 destination based routed system inherently does not allow two devices to communicate while sharing the same IPv4 address, even if with different ports. Similarly, such a system does not allow on the basis of a IPv4 source address alone to perform address spoofing prevention. These two issues naturally render regular IPv4 based routed networks incapable of supporting a shared address solution.

5.14.2. Discussion

In terms of the IPv4 data plane of the 4via6 solution, the CPE and the stateless gateway components need to be modified in terms of their IPv4 forwarding behaviour. The CPE's NAPT44 function, must be capable of sending traffic towards the IPv6 adaptation function when the traffic is addressed to its (shared) IPv4 address but a different port than the one assigned to the CPE. Similarly, the CPE's NAPT44 function must be capable of receiving traffic addressed from its (shared) IPv4 address but a different port than the one assigned to it.

On the IPv6 data plane the stateless 4via6 solution does not suffer from the issue by the nature of relying on regular IPv6 forwarding. Address-spoofing security can be realized on regular IPv6 devices

plane, in a way which effectively does not allow a CPE to send IPv6 traffic from a source IPv6 address that it has not been assigned. The spoofing of IPv4 addresses can be prevented in this manner in 4via6 solution relying on translation (dIVI). Tunneling 4via6 solutions (4rd) require IPv6+IPv4 source address validation to be performed at the CPE and stateless gateway, by the IPv6 adaptation function.

The conceptual IPv6 adaptation function has many of its core principles already defined either as part of IPinIP tunneling or stateless NAT64 drafts. However additional work, such as defining the port indexing schemes, is needed and is at the heart of what needs to be covered in the individual solution drafts that fall under the stateless 4via6 family. Throughout, no legacy IPv4 end-systems are expected to implement these techniques.

5.15. Other

5.15.1. Abuse Claims

Because the IPv4 address is shared between several customers, and in order to meet the traceability requirement discussed in Section 12 of [I-D.ietf-intarea-shared-addressing-issues], Service Providers must store the assigned ports in addition to the IPv4 address.

If the remote server does not implement the recommendation detailed in [I-D.ietf-intarea-server-logging-recommendations], the Service Provider may be obliged to reveal the identity of all customers sharing the same IP address at a given time.

5.15.2. Fragmentation and Traffic Asymmetry

In order to deliver a fragmented IPv4 packet to its final destination, among those having the same IPv4 address, a dedicated procedure similar to the one defined in Section 3.5 of [RFC6146] is required to reassemble the fragments in order to look at the destination port number.

When several stateless IPv4/IPv6 interconnection nodes are deployed, and because of traffic asymmetry, situations where fragments are not handled by the same stateless IPv4/IPv6 interconnection node may occur. Such context would lead to session breakdowns. As a mitigation, a solution would be to redirect fragments towards a given node which will be responsible for implementing the procedure documented in [RFC6146]. The redirection procedure is stateless.

As a conclusion, this issue can be mitigated.

5.15.3. Multicast Services

IPv4 service continuity must be guaranteed during the transition period, including the delivery of multicast-based services such as IPTV. Because only an IPv6 prefix will be provided to a CPE, dedicated functions are required to be enabled for the delivery of legacy multicast services to IPv4 receivers. This is critical since many of the current IPTV contents are likely to remain IPv4-formatted and there will remain legacy receivers (e.g., IPv4-only Set Top Boxes (STB)) that can't be upgraded or be easily replaced.

This issue is similar to the one encountered in the stateful case, and the same solution can be used to mitigate the issue (e.g., [I-D.qin-software-dslite-multicast]).

As a conclusion, this issue can be solved.

6. Conclusion

As per the discussion in this document, the authors believe that the set of issues specifically attributed to A+P based such as the stateless 4via6 solution with characteristics as per Section 3, either do not apply, or can be mitigated. In several aspects, a stateless 4V6 solution represents a reasonable trade off compared to alternatives in areas such as NAT logging, ease as of deployment and operations, all of which are actually facilitated by such a solution.

In terms of the 4V6 transport mode, both translation and mapped tunnel appear to be share the same key characteristics, but applicable to different contexts. The mapped tunnel mode appears desirable when the operator has no expectations of applying any more elaborate traffic based services, and/or concerned about the loss of IP Options or the use of NAT64 technology. The translation based approach appears particularly attractive to operators who are concerned about integrating traffic into a more elaborate suite of services based on regular IPv6 data-plane functionality, as opposed to specific IPinIP data plane functionality.

7. IANA Considerations

This document does not raise any IANA considerations.

8. Security Considerations

This document does not introduce any security considerations over and

above those already covered by the referenced solution drafts.

9. Contributors and Acknowledgements

The authors thank Dan Wing, Nejc Skoberne, Remi Depres, Xing Li, Jan Zorz, Satoru Matsushima, Mohamed Boucadair, Qiong Sun, and Arkadiusz Kaliwoda for their reviews and draft input.

10. References

10.1. Normative References

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

10.2. Informative References

[I-D.bajko-pripaddrassign]

Bajko, G., Savolainen, T., Boucadair, M., and P. Levis, "Port Restricted IP Address Assignment", draft-bajko-pripaddrassign-03 (work in progress), September 2010.

[I-D.despres-softwire-4rd]

Despres, R., "IPv4 Residual Deployment across IPv6-Service networks (4rd) A NAT-less solution", draft-despres-softwire-4rd-00 (work in progress), October 2010.

[I-D.ietf-intarea-server-logging-recommendations]

Durand, A., Gashinsky, I., Lee, D., and S. Sheppard, "Logging recommendations for Internet facing servers", draft-ietf-intarea-server-logging-recommendations-04 (work in progress), April 2011.

[I-D.ietf-intarea-shared-addressing-issues]

Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", draft-ietf-intarea-shared-addressing-issues-05 (work in progress), March 2011.

[I-D.ietf-softwire-ds-lite-tunnel-option]

Hankins, D. and T. Mrugalski, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Option for Dual- Stack Lite", draft-ietf-softwire-ds-lite-tunnel-option-10 (work in progress), March 2011.

- [I-D.ietf-softwire-dual-stack-lite]
Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", draft-ietf-softwire-dual-stack-lite-11 (work in progress), May 2011.
- [I-D.murakami-softwire-4v6-translation]
Murakami, T., Chen, G., Deng, H., Dec, W., and S. Matsushima, "4via6 Stateless Translation", draft-murakami-softwire-4v6-translation-00 (work in progress), July 2011.
- [I-D.operators-softwire-stateless-4v6-motivation]
Boucadair, M., Matsushima, S., Lee, Y., Bonness, O., Borges, I., and G. Chen, "Motivations for Stateless IPv4 over IPv6 Migration Solutions", draft-operators-softwire-stateless-4v6-motivation-02 (work in progress), June 2011.
- [I-D.qin-softwire-dslite-multicast]
Wang, Q., Qin, J., Boucadair, M., Jacquenet, C., and Y. Lee, "Multicast Extensions to DS-Lite Technique in Broadband Deployments", draft-qin-softwire-dslite-multicast-04 (work in progress), June 2011.
- [I-D.thaler-port-restricted-ip-issues]
Thaler, D., "Issues With Port-Restricted IP Addresses", draft-thaler-port-restricted-ip-issues-00 (work in progress), February 2010.
- [I-D.vixie-dnsexst-dns0x20]
Vixie, P. and D. Dagon, "Use of Bit 0x20 in DNS Labels to Improve Transaction Identity", draft-vixie-dnsexst-dns0x20-00 (work in progress), March 2008.
- [I-D.wing-softwire-port-control-protocol]
Wing, D., Penno, R., and M. Boucadair, "Pinhole Control Protocol (PCP)", draft-wing-softwire-port-control-protocol-02 (work in progress), July 2010.
- [I-D.xli-behave-divi]
Bao, C., Li, X., Zhai, Y., and W. Shang, "dIVI: Dual-Stateless IPv4/IPv6 Translation", draft-xli-behave-divi-03 (work in progress), July 2011.

- [I-D.xli-behave-divi-pd]
Li, X., Bao, C., Dec, W., Asati, R., Xie, C., and Q. Sun,
"dIVI-pd: Dual-Stateless IPv4/IPv6 Translation with Prefix
Delegation", draft-xli-behave-divi-pd-01 (work in
progress), September 2011.
- [I-D.ymbk-aplusp]
Bush, R., "The A+P Approach to the IPv4 Address Shortage",
draft-ymbk-aplusp-10 (work in progress), May 2011.
- [RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's
Robustness to Blind In-Window Attacks", RFC 5961,
August 2010.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X.
Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052,
October 2010.
- [RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-
Protocol Port Randomization", BCP 156, RFC 6056,
January 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.

Authors' Addresses

Wojciech Dec
Cisco
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands

Email: wdec@cisco.com

Rajiv Asati
Cisco
Raleigh, NC
USA

Phone:
Fax:
Email: rajiva@cisco.com
URI:

Congxiao Bao
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing, 100084
CN

Phone: +86 10-62785983
Fax:
Email: congxiao@cernet.edu.cn
URI:

Hui Deng
China Mobile
Beijing,
CN

Phone:
Fax:
Email: denghui@chinamobile.com
URI:

Mohamed Boucadair
France Telecom
France

Phone:
Fax:
Email: mohamed.boucadair@orange-ftgroup.com
URI:

V6OPS
Internet Draft
Intended status: Informational
Expires: January 9, 2012

X.Deng
T.Zheng
M.Boucadair
L.Wang
France Telecom
X.Huang
Q.Zhao
Y.Ma
BUPT
July 8, 2011

Implementing AplusP in the provider's IPv6-only network
draft-deng-v6ops-aplusp-experiment-results-01.txt

Status of this Memo

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

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

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

This Internet-Draft will expire on January 9, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

This memo describes an implementation of A+P in a provider's IPv6-only network. It provides details of the implementation, network elements, configurations and test results as well. Besides traditional port range A+P, a scattered port sets flavour of A+P is also implemented and verified for the sake of distributing incoming ports among customers in a more discrete way. The test results consist of the application compatibility test, UPnP extension for A+P, port usage and BitTorrent behaviour with A+P.

This memo focuses on the IPv6 flavor of A+P.

Table of Contents

1. Introduction	3
2. Terminology	3
3. Implementation environment	4
3.1. Environment Overview	4
3.2. Implementation and Configuration of A+P	5
3.2.1. IPv4-Embedded IPv6 Address Format For A+P CPE.	5
3.2.2. DHCPv6 Configurations	6
3.2.3. Avoiding Fragmentation	6
3.3. Implementing scattered Port Sets for A+P	7
3.3.1. Scattered Port Sets allocation mechanism	7
3.3.2. IPv4-Embedded IPv6 Address Format for Scattered Port Sets A+P CPE	10
3.3.3. Customize a scattered Ports Set A+P NAT on Linux	10
4. Application Tests and Experiments in A+P Environment	11
4.1. A+P Impacts on Applications	12
4.2. UPnP extension experiment	13
4.3. Port Usage of Applications	14
4.4. BitTorrent Behaviour in A+P	16
5. Security Considerations	17
6. IANA Considerations	17
7. Conclusion	17
8. References	18
8.1. Normative References	18
8.2. Informative References	18
9. Acknowledgments	19

1. Introduction

A+P [draft-ymbk-aplusp-09] is a technique to share IPv4 addresses during the IPv6 transition period without requiring a NAT function in the provider's network. The main idea of A+P is treating some bits from the port number in the TCP/UDP header as additional end point

identifiers to extend the address field, thereby leaving a range of ports available to applications. This feature facilitates migration of networks to IPv6-only while offering the IPv4 connectivity services to customers, because the IPv4 address and the significant bits from the port range can be encoded in an IPv6 address and therefore transporting IPv4 traffic over IPv6 network by stateless IPv6 routing.

We have implemented A+P in a residential ADSL access network, where IPv6-only access network is provided over PPPoE. In this document, we describe the implementation environment including A+P IPv6 prefix format and network elements configurations, and results of application tests as well. The document focuses on the implementation of the SMAP function specified in [draft-ymbk-aplusp-09]:

- o Implement DHCPv6 options to retrieve an IPv4-embedded IPv6 address and a port range.
- o Support of those DHCPv6 options in both the DHCPv6 server side and the DHCPv6 client side.
- o Support of those DHCPv6 options in both the DHCPv6 server side and the DHCPv6 client side.

For extensive application tests results in A+P environment, please refer to [draft-boucadair-behave-bittorrent-portrange-02] and [draft-boucadair-port-range-01].

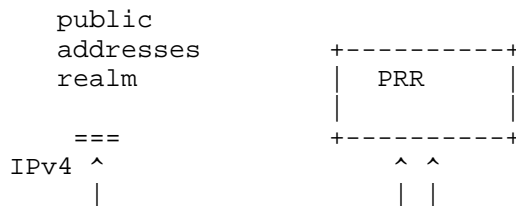
2. Terminology

This document makes use of the following terms:

- o PRR: Port Range Router
- o A+P CPE: A+P aware Customer Premise Equipment

3. Implementation environment

3.1. Environment Overview



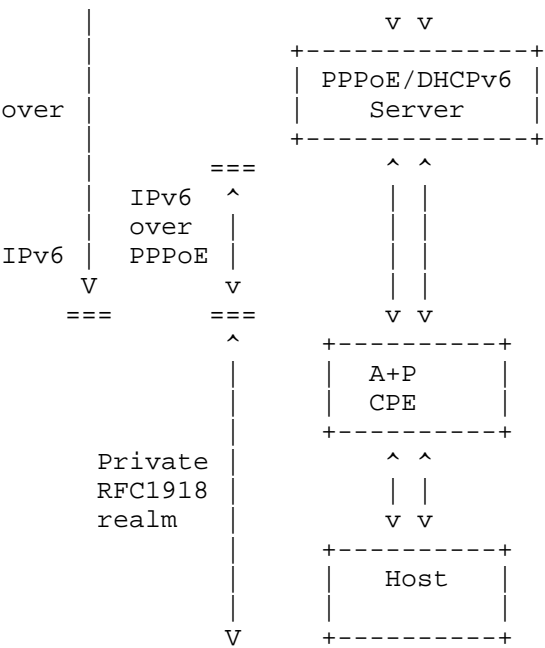


Figure 1 : Implementation Environment

We had developed both A+P home gate way function and Port Range Router (PRR) function on Linux platform and ported the home gate way function to a Linksys wrt 54G CPE, on which an openwrt 2.6.32 (based on Linux kernel) is running.

Figure 2 shows the Parameters of A+P CPE. IPv6 is provisioning over PPPoE to CPE while DHCPv6 server offers IPv6 prefix and A+P

parameters by extended options defined in [draft-boucadair-dhcpv6-shared-address-option].

Model	CPU Speed (MHz)	Flash (MB)	RAM (MB)	Wireless NIC	Wireless Standard	Wired Ports
Linksys WRT54GS	200	8	32	Broadcom (integrated)	11g	5

Figure 2 :Parameters of A+P CPE

3.2. Implementation and Configuration of A+P

Aplusp CPE, using Netfilter framework, the IPv4 port restricted NAT operation performed by CPE has been implemented by simply rules through iptables tool on Linux. After the port restricted NAT operation, the IPv4 packets are sent to a TUN interface which is described as a virtual network interface in Linux. Using the IPv4-Embedded IPv6 address format defined in section 3.2.1, an IPv4-in-IPv6 encapsulation/decapsulation is performed by the TUN interface handler.

PRR, located in the interconnection point of the IPv6 network and IPv4 network, has been implemented with two main functions: 1) IPv4-in-IPv6 encapsulation/decapsulation; Like CPE, TUN driver is also used in PRR to achieve function IPv4-in-IPv6 encapsulation/decapsulation. 2) destination port based routing function, which is responsible for routing the IPv4 traffic originated from the IPv4 Internet to the Port Range restricted A+P CPE. Destination port based routing is implemented by generating IPv6 destination address, pre-assigned from IPv4 address and port range to each CPE, according to IPv4-Embedded IPv6 address format defined in section 3.2.1.

3.2.1. IPv4-Embedded IPv6 Address Format For A+P CPE

31bits	1bit	32bits	8 bits	16bits	4bits	1bit	1bit	1bit	1bit	32 bits
AplusP Prefix	flag 0	Public IPv4 Address	EUI64	port Range	Port Range Size	flag 1	flag 2	flag 3	flag 4	Public IPv4 Address

Figure 3 :IPv4-Embedded IPv6 address format

flag0: Is this address used by CPE or PRR?

flag1: Is address shared?

flag2: Is length of invariable present?

flag3: Is port range identifying sub network?

flag4: Reserved?

To facilitate test and experiment on AplusP solution, recently, we are considering release this AplusP implementation under open source license. For more implementation details, please refer to [Implementing A+P]

3.2.2. DHCPv6 Configurations

DHCPv6 options defined in [draft-boucadair-dhcpv6-shared-address-option] have been implemented. These options allow to configure a shared address together with a port range using DHCPv6.

3.2.3. Avoiding Fragmentation

Normally the TCP protocol stack will employ Maximum Segment Size (MSS) negotiation and/or Path Maximum Transmission Unit Discovery (PMTUD) to determine

the maximum packet size, and then try to send as large as possible datagram to achieve better throughput. However the IPv4-in-IPv6 encapsulation and the PPPoE header is very likely to cause a larger packet that exceeds the maximum MTU of the wire, and result in undesired fragmentation processing and decrease transmission

efficiency.

A simple solution is to enable iptables on A+P CPE to modify the MSS value of TCP session, using the command like "iptables -t mangle -A FORWARD -p tcp --tcp-flags SYN,RST SYN -j TCPMSS --set-mss DESIRED_MSS_VALUE". Here the DESIRED_MSS_VALUE is taken into account of common size of IPv4 header without options, common size of TCP header and size of basic IPv6 header and PPPoE header as well.

3.3. Implementing scattered Port Sets for A+P

3.3.1. Scattered Port Sets allocation mechanism

As described in [I-D.ietf-intarea-shared-addressing-issues], a bulk of incoming ports can be reserved as a centralized resource shared by all subscribers using a given restricted IPv4 address. In order to distribute incoming ports as scattered as possible among subscribers sharing the same restricted IPv4 address, other than allocating a continuous range of ports to per subscriber, a solution to distribute bulks of non-continuous ports among subscribers, which also takes port randomization of CPE NAT into account, because port randomization is one protection among others against blind attacks, is elaborated thereby.

On every restricted IPv4 address, according to port set size N, $\log_2(N)$ bits are randomly chose as subscribers identification bits(s bit) among 1st and 16th bits. Take a sharing ration 1:32 for example, Figure 4 shows an example of 5bits (2nd, 5th, 7th, 9th, 11th) being chose as s bit.

1st	2nd	3rd	4th	5th	6th	7th	8th
0	s	0	0	s	0	s	0
9th	10th	11th	12th	13th	14th	15th	16th
s	0	s	0	0	0	0	0

Figure 4 : An s bit selection example (on a sharing ration 1:32 address).

Subscriber ID pattern is then formed by setting all the s bits to 1

and other trivial bits to 0. Figure 5 illustrates an example of subscriber ID pattern which follows the s bit selection of figure 4. Note that the subscriber ID pattern can be different, ensured by the random s bit selection, per restricted IP address no matter whether the sharing ratio varies.

1st	2nd	3rd	4th	5th	6th	7th	8th
0	1	0	0	1	0	1	0

9th	10th	11th	12th	13th	14th	15th	16th
1	0	1	0	0	0	0	0

Figure 5 : A subscriber ID pattern example (on a sharing ration 1:32 address).

Subscribers ID value is then assigned by setting subscriber ID pattern bits (s bits shown in figure 4) to a unique customer value and setting other trivial bits to 1. An example of subscriber ID value, having a subscriber ID pattern shown in the figure 5 and a customer value 0, is shown in the figure 6.

1st	2nd	3rd	4th	5th	6th	7th	8th
1	0	1	1	0	1	0	1

9th	10th	11th	12th	13th	14th	15th	16th
0	1	0	1	1	1	1	1

Figure 6 : A subscriber ID value example (customer value: 0)

Subscriber ID pattern and subscriber ID value together uniquely defines a restricted port set (Non-contiguous port sets or a contiguous port range, depends on Subscriber ID pattern and subscriber ID value) on a restricted IP address.

Pseudo-code shown in the figure 7 describes how to use subscriber ID pattern and subscriber ID value to implement a random ephemeral port selection function within the defined restricted port sets on a customer NAT.

```
do{
    restricted_next_ephemeral = (random()|subscriber_ID_pattern)
                                & subscriber_ID_value;

    if(five-tuple is unique)
        return restricted_next_ephemeral;
}
```

Figure 7 : Random ephemeral port selection within the restricted port set

3.3.2. IPv4-Embedded IPv6 Address Format for Scattered Port Sets A+P CPE

31bits	1bit	32bits	8bits	16bits	4bits	1bit	1bit	1bit	1bit	32bits
AplusP Prefix	flag 0	Public IPv4 Address	EUI64	SID_ Value	Reser -ved	flag 1	flag 2	flag 3	flag 4	Public IPv4 Address

Figure 8 :IPv4-Embedded IPv6 address format

SID Value: Subscriber_ID_Value, which is unique for per subscriber sharing a given restricted IPv4 address. and has been allocated to each subscriber.

flag0: Is this address used by CPE or PRR?

flag1: Is address shared?

flag2: Is length of invariable present?

flag3: Is port range identifying sub network?

flag4: Reserved?

PRR maintains a mapping table, which consists of restricted IPv4 address and its Subscriber ID Pattern. To form an IPv6 destination address for incoming packet, PRR could find the right SID Pattern according to a destination IPv4 address, and then apply a simple operation shown in the figure 9.

$$\text{SID_Value} = \text{Destination_Port} \mid (\sim \text{SID_Pattern}).$$

Figure 9 :PRR calculates SID Value

3.3.3. Customize a scattered Ports Set A+P NAT on Linux

With a linux kernel 2.6.32.36, only one line of linux kernel code is changed, as shown in the figure5, and the same IPtables command line interface is used with the only one change of semantic that the original starting of port range becomes SID_Value and the ending port of a port range becomes SID_Pattern. The command line with iptables to configure a scattered Ports Set A+P is illustrated in the figure 11.

```
bool nf_nat_proto_unique_tuple(...)
...
//The Original code:
// *portptr = htons(min + off % range_size);
// was changed to:
*portptr = htons((ntohs(off) | min ) & max );
...
```

Figure 10:Function of finding a unique 5-tuple for a scattered port sets A+P NAT

```

iptables -t nat -A POSTROUTING -o eth0 -p tcp -j SNAT --to-source
a.b.c.d: SID_Value-SID_Pattern --random

iptables -t nat -A POSTROUTING -o eth0 -p udp -j SNAT --to-source
a.b.c.d: SID_Value-SID_Pattern --random

```

Figure 11: IPtables commands for a scattered ports set A+P NAT

4. Application Tests and Experiments in A+P Environment

A set of well-known applications have been tested in this IPv6 flavor of A+P environment to access A+P impacts on them. The test results show that IPv6 flavor of A+P has the same impacts on applications as IPv4 flavor A+P does [draft-boucadair-port-range-01]. Web browsing (IE and Firefox), Email (Outlook), Instant message(MSN),Skype, Google Earth work normally with A+P. For more details, please refer to [draft-boucadair-port-range-01].

4.1. A+P Impacts on Applications

Application	A+P impacts
IE	None
Firefox	None
FTP(Passive mode)	None
FTP(Active mode)	require opening port forwarding
Skype	None
Outlook	None
Google Earth	None
BitComet	UPnP extensions may be required, when listening port is out of A+P range; other minor effects(see section 4.4)
uTorrent	UPnP extensions may be required, when listening port is out of A+P range; other minor effects(see section 4.4)

Live Messenger	None
----------------	------

Figure 12:Aplusp impacts on applications

For P2P (Peer-to-Peer) applications, when some of them listening on specific port to expect inbound connection, it is likely to fail due to the listening port is out of A+P port range. Some UPnP extensions may be required to make P2P applications work properly with A+P. Other minor effects of A+P are discussed in section 4.4.

4.2. UPnP extension experiment

To make P2P application work properly with port restricted NAT , we have designed extensions including new variables, new errorcodes as well as new actions to UPnP 1.0, and have them implemented with [Emule], [open source UPnP SDK 1.0.4 for Linux] and [Linux UPnP IGD 0.92].

In figure 5, a new error code is proposed for the existing "AddPortMapping" action to explicitly indicate the situation that the requested external port is out of range.

ErrorCode	errorDescription	Description
728	ExternalPortOutOfRange	The external port is out of the port range assigned to this external interface

Figure 13:New ErrorCode for "AddPortMapping" action

New state variables have been introduced to reflect the valid port range. The definitions of these state variables are shown in figure 6.

Variable Name	Req. or Opt.	Data Type	Allowed Value	Default Value	Eng. Units
PortRangeLow	0	ui2	>=0	0	N/A
PortRangeHigh	0	ui2	<=65535	65535	N/A

Figure 14: New state variables for port range

Correspondingly, new actions, GetPortRangeLow and GetPortRangeHigh, defined to retrieve port range information are illustrated in figure 7. An IP address should be provided as argument to invoke the new actions, for the port range is associated with a specific IP address.

Action Name	Argument	Dir.	Related StateVariable
GetPortRangeLow	NewExternal IPAddress	IN	ExternalIPAddress
	NewPortRange Low	OUT	PortRangeLow
GetPortRangeHigh	NewExternal IPAddress	IN	ExternalIPAddress
	NewPortRange High	OUT	PortRangeHigh

Figure 15: New actions for port range

Please refer to [UPnP Extension] for more details of UPnP extension experiment in A+P.

4.3. Port Usage of Applications

Port consumptions of applications not only impact the deployment factor (i.e., port range size) for AplusP solution but also play an important role in determining the port limitation of per customer on

AFTR for Dual-Stack Lite.

Therefore we have also developed and deployed a Service Probe in our IPv6 network, which use IPv6 TCP socket to ask AplusP CPE for NAT session usage, and store AplusP NAT statistics in a Mysql database for further analysis of application behaviors in terms of port and session consumptions.

In figure 8, the maximum port usage of each application is the peak number of port consumption per second during the whole communication process. The duration time represents the total time from the first NAT binding entry being established to the last one being destroyed.

Application	Test case	Maximum port usage	Duration (seconds)
IE	browsing a news website	20-25	200
	browsing a video website	40-50	337
Firefox	browsing a news website	25-30	240
	browsing a video website	80-90	230
Chrome	browsing a news website	50-60	340
	browsing a video website	80-90	360
Android Chrome	browsing a news website	40-50	300
	browsing a video website	under 10	160
Google Earth	locating a place	30-35	240
Android Google Earth	locating a place	10-15	240
Skype	make a call	under 10	N/A
BitTorrent	downloading a file	200	N/A

Figure 16: Port usage of applications

4.4. BitTorrent Behaviour in A+P

[draft-boucadair-behave-bittorrent-portrange] provides an exhaustive testing report about the behaviour of BitTorrent in an A+P architecture. [draft-boucadair-behave-bittorrent-portrange] describes the main behavior of BitTorrent service in an IP shared address environment. Particularly, the tests have been carried out on a testbed implementing [ID.boucadair-port-range] solution. The results are, however, valid for all IP shared address based solutions.

Two limitations were experienced. The first limitation occurs when two clients sharing the same IP address want to simultaneously retrieve the SAME file located in a SINGLE remote peer. This limitation is due to the default BitTorrent configuration on the remote peer which does not permit sending the same file to multiple ports of the same IP address. This limitation is mitigated by the fact that clients sharing the same IP address can exchange portions with each other, provided the clients can find each other through a common tracker, DHT, or Peer Exchange. Even if they can not, we observed that the remote peer would begin serving portions of the file automatically as soon as the other client (sharing the same IP address) finished downloading. This limitation is eliminated if the remote peer is configured with `bt.allow_same_ip == TRUE`.

The second limitation occurs when a client tries to download a file located on several seeders, when those seeders share the same IP address. This is because the clients are enforcing `bt.allow_same_ip` parameter to `FALSE`. The client will only be able to connect to one sender, among those having the same IP address, to download the file (note that the client can retrieve the file from other seeders having distinct IP addresses). This limitation is eliminated if the local client is configured with `bt.allow_same_ip == TRUE`, which is somewhat likely as those clients will directly experience better throughput by changing their own configuration.

Mutual file sharing between hosts having the same IP address has been checked. Indeed, machines having the same IP address can share files with no alteration compared to current IP architectures.

5. Security Considerations

TBD

6. IANA Considerations

This document includes no request to IANA.

7. Conclusion

Despite A+P introduces some impacts on existence applications, issues of P2P applications due to the port restricted NAT have been resolved by UPnP extension experiment in our test bed, and other issues are shared by other IP address sharing solutions. Therefore, from our work, it has been proved that deploying A+P in the Service Provider's IPv6 network during IPv6 transition period is feasible.

8. References

8.1. Normative References

[Implementing A+P]

Xiaoyu ZHAO., "Implementing Public IPv4 Sharing in IPv6 Environment", ICCGI 2010

[UPnP Extension]

Xiaoyu ZHAO., "UPnP Extensions for Public IPv4 Sharing in IPv6 Environment", ICNS 2010

8.2. Informative References

[1] Faber, T., Touch, J. and W. Yue, "The TIME-WAIT state in TCP and Its Effect on Busy Servers", Proc. Infocom 1999 pp. 1573-1583.

[Fab1999] Faber, T., Touch, J. and W. Yue, "The TIME-WAIT state in TCP and Its Effect on Busy Servers", Proc. Infocom 1999 pp. 1573-1583.

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

[draft-ymbk-aplusp-09]

R. Bush., " The A+P Approach to the IPv4 Address Shortage", draft-ymbk-aplusp-09 (work in progress), February 17, 2011.

[draft-boucadair-dhcpv6-shared-address-option]

M. Boucadair., "Dynamic Host Configuration Protocol (DHCPv6) Options for Shared IP Addresses Solutions", draft-

boucadair-dhcpv6-shared-address-option-01 (work in progress), December 21, 2009

[draft-boucadair-port-range-01]

"IPv4 Connectivity Access in the Context of IPv4 Address Exhaustion", draft-boucadair-port-range-01(work in progress), January 30, 2009

[Emule]

<http://www.emule-project.net/>. [Accessed October 26, 2009]

[UPnP SDK 1.0.4 for Linux]

<http://upnp.sourceforge.net/>. [Accessed October 26, 2009].

[Linux UPnP IGD 0.92].

<http://linuxigd.sourceforge.net/>. [Accessed October 26, 2009].

[draft-boucadair-behave-bittorrent-portrange]

M. Boucadair., "Behaviour of BitTorrent service in an IP Shared Address Environment", draft-boucadair-behave-bittorrent-portrange-02.txt

9. Acknowledgments

The experiments and tests described in this document have been explored, developed and implemented with help from Zhao Xiaoyu, Eric Burgey and JACQUENET Christian.

Thanks to Jan Zorz for comments.

Authors' Addresses

Xiaohong Deng
France Telecom
Hai dian district, 100190, Beijing,
China

Email: xiaohong.deng@orange-ftgroup.com

Mohamed BOUCADAIR
France Telecom
Rennes, 35000 France

Email: mohamed.boucadair@orange-ftgroup.com

Lan Wang
France Telecom
Hai dian district, 100190, Beijing, China

Email: lan.wang@orange-ftgroup.com

Tao Zheng
France Telecom
Hai dian district, 100190, Beijing, China

Email: tao.zheng@orange-ftgroup.com

Xiaohong Huang
Beijing University of Post and Telecommunication
Email: huangxh@bupt.edu.cn

Qin Zhao
Beijing University of Post and Telecommunication
Email: zhaoqin.bupt@gmail.com

Yan MA
Beijing University of Post and Telecommunication
Email: mayan@bupt.edu.cn

6man Working Group
Internet Draft
Intended status: Standards track
Expires: January, 2012

N. Elkins
Inside Products
L. Kratzke
IBM
M. Ackermann
BCBS of Michigan
July 2011

IPv6 Diagnostic Header
draft-elkins-6man-ipv6-diagnostic-header-00.txt

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

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

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

The IPv4 main header contained a 16-bit IPID for fragmentation and reassembly. This field was commonly used by network diagnosticians for tracking packets on multi-tier networks. In IPv6, the IPID has been moved to the Fragment header. A new Diagnostic header for IPv6 which can be sent with every packet as a part of the Destination Options header with a 64-bit IPID is defined in this document.

Table of Contents

1. Introduction	2
2. Conventions used in this document	3
3. Applicability	3
4. IPv6 Diagnostics Header Format	3
4.1. Destination Options Header	4
4.2. Diagnostic Header Option	5
4.3. Implementation Considerations	6
5. Backward Compatibility	6
6. Security Considerations	6
7. IANA Considerations	6
10. References	6
10.1. Normative References	7
10.2. Informative References	7
11. Acknowledgments	8

1. Introduction

In IPv4, the 16 bit Identification field is located at an offset of 4 bytes into the IPv4 header and is described in RFC791 [RFC791]. In IPv6, it is a 32 bit field contained in the Fragment header defined by section 4.5 of RFC2460 [RFC2460]. Unfortunately, unless fragmentation is being done by the source node, the packet will not contain this Fragment header, and therefore will have no Identification field.

The intended purpose of the IPv4 Identification (ID) field is to enable fragmentation and reassembly, and as currently specified is required to be unique within the maximum segment lifetime (MSL) on all datagrams. The MSL is often 2 minutes.

In practice, the IPID field is used for more than fragmentation. Some TCP stacks have the same IPID counter for all connections; some have an IPID counter on a per connection basis. Each time a TCP stack sends out a packet, the IPID is incremented by one (or sometimes 2).

During network diagnostics, packet traces may be taken at multiple places along the path or at the source and destination. Then, packets can be matched by looking at the IPID. Obviously, the time at each device will differ according to the clock on that device; so another metric is required. This method of taking multiple traces along the path is frequently used on large multi-tier networks to see where the packet loss or packet corruption is happening.

Having said that, a known problem with the uniqueness of the IPv4 ID is that since the field is only 16 bits, then for high speed devices, wrapping will occur. As discussed in RFC4963 [RFC4963] and draft-ietf-intarea-ipv4-id-update-02.txt [Draft-ipv4-id], if the uniqueness requirement were strictly enforced, all connections would be limited to a maximum speed of 6.4 Mbps. Clearly, this uniqueness requirement is widely ignored.

In IPv6, the IPID field, which is in the Fragment header, has been increased to 32 bits. This may need to be reconsidered as data rates increase but if the IPID is used for fragmentation and reassembly alone, the requirement for uniqueness within the MSL period is generally not an issue today. RFC4963 [RFC4963] discusses the issue of reassembly errors at high data rates for IPv4 with a 16-bit counter.

However, for its de-facto diagnostic mode usage, an IPID needs to be available whether or not fragmentation occurs, and needs to be unique in the context of the entire session, and across all the connections controlled by the TCP stack. The problem of 32 bit counters is known and has been resolved in areas such as SNMP counters by creating 64-bit counters as described in RFC2233 [RFC2863].

This document will address a way to make the IPv6 IPID field available and unique for its valuable diagnostic usage. A Destination Options header is proposed which may be sent by TCP stacks in diagnostic mode. The implementation MAY provide the option of either turning on the Diagnostic Header for all connections or turn it on just for specific connections. The more sophisticated usage of this header would be to send it for a single connection only.

2. Conventions used in this document

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 [RFC2119].

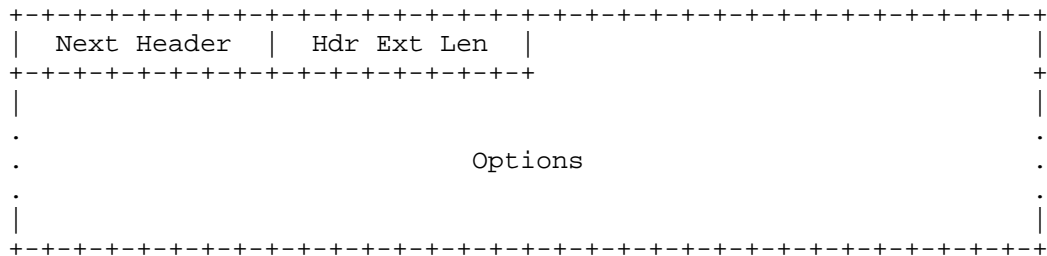
3. Applicability

The base IPv6 standard, RFC2460, [RFC2460] allows the use of extension headers including destination options in order to encode optional destination information in an IPv6 packet. Extended diagnostic information such as this MUST be sent by implementations upon request. The proposed Diagnostic Options header is an implementation of the destination options header.

4. IPv6 Diagnostics Header Format

4.1 Destination Options Header

The Destination Options header is used to carry optional information that need be examined only by a packet's destination node(s). The Destination Options header is identified by a Next Header value of 60 in the immediately preceding header, and has the following format:



Next Header	8-bit selector. Identifies the type of header immediately following the Destination Options header. Uses the same values as the IPv4 Protocol field [RFC-1700 et seq.].
Hdr Ext Len	8-bit unsigned integer. Length of the Destination Options header in 8-octet units, not including the first 8 octets.
Options	Variable-length field, of length such that the complete Destination Options header is an integer multiple of 8 octets long. Contains one or more TLV-encoded options.

Figure 1: Destination Options Extension header layout

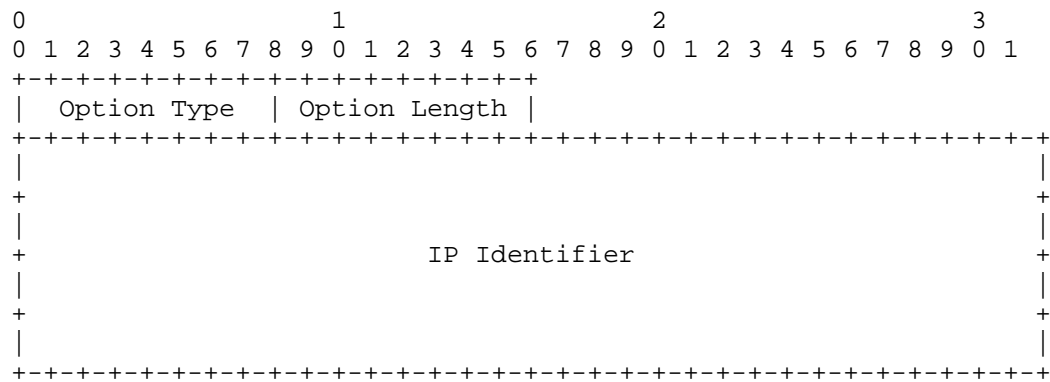
The desired action for a destination node who does not recognize this option is to ignore the header and continue processing the packet normally.

According to RFC2460 [RFC2460], if the Destination Options header Option Type has the value 00 in its highest-order two bits, the receiving device should skip over this option and continue processing the header.

4.2. IPv6 Diagnostic Header Option

The IPv6 Diagnostic Header option is carried by the Destination Option extension header (Next Header value = 60). It is used in a packet sent by a node to facilitate diagnostics by informing the recipient and passive viewers of the packet such as packet capture facilities of the packet's IP Identifier.

The IPv6 Diagnostic Header option is encoded in type-length-value (TLV) format as follows:



Option Type

nnn = 0xXX [To be assigned by IANA] [RFC2780]

Option Length

8-bit unsigned integer. Length of the option, in octets, excluding the Option Type and Option Length fields. This field MUST be set to 64.

IP Identifier

The IP Identifier of the packet for 64 bits.

The alignment requirement for the IP Identifier option is $8n+6$.

The two highest-order bits of the Option Type field are encoded to indicate specific processing of the option; for the IP Identifier option, these two bits MUST be set to 00. This indicates the following processing requirements:

- o 00 - skip over this option and continue processing the header.
- o The data within the option cannot change en route to the packet's final destination.

The IPv6 Diagnostic Header option MUST be placed as follows:

- o After the routing header, if that header is present
- o Before the Fragment Header, if that header is present
- o Before the AH Header or ESP Header, if either one of those headers are present.

For each IPv6 packet header, the IPv6 Diagnostic Header MUST NOT appear more than once. However, an encapsulated packet MAY contain a separate IP Identifier option associated with each encapsulating IP header.

The inclusion of a IPv6 Diagnostic Header in a packet affects the receiving node's processing of only this single packet. No state is created or modified in the receiving node as a result of receiving a IPv6 diagnostic Header in a packet.

4.3. Implementation Considerations

In implementation, a TCP stack may send this additional header for all connections or, in a more sophisticated usage, a single connection only.

We suggest that initiation of this header be done in a 'Debug on', 'Debug off' mode. That is, a diagnostician may decide that this header is required for a certain timeframe or for a certain set of packets after a network problem is encountered. The diagnostician may then issue a command to the TCP stack indicating that addition of the IP Identifier header should now begin. This is the 'Debug on' state. After a certain amount of time, then 'Debug off' should be issued as a command. Alternatively, the TCP stack may have a fixed time (for example, 5 minutes) after which debug mode will automatically be turned off.

5. Backward Compatibility

The scheme proposed in this document is backward compatible with all the currently defined IPv6 extension headers. According to RFC2460 [RFC2460], if the destination node does not recognize this option, it should skip over this option and continue processing the header.

6. Security Considerations

There are no security considerations.

7. IANA Considerations

A Destination Options value of XXX is pending IANA action. [RFC2780]

10. References

10.1. Normative References

[RFC791] Postel, J., "Internet Protocol", RFC 791 / STD 5, September 1981.

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

[RFC2863] K. McCloghrie, K., Kastenholz, F. "The Interfaces Group MIB", RFC 2863, June 2000.

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

[RFC2780] Bradner, S., Paxson, V. "IANA Allocation Guidelines For Values In the Internet Protocol and Related Headers", RFC 2780, March 2000.

See also:

<http://www.iana.org/assignments/ipv6-parameters>

10.2. Informative References

[RFC4963] Heffner, J., Mathis, M., Chandler, B., "IPv4 Reassembly Errors at High Data Rates", RFC 4963, July 2007.

[Draft-ipv4-id] Touch, J., "Updated Specification of the IPv4 ID Field", draft-ietf-intarea-ipv4-id-update-02.txt, March 2011

11. Acknowledgments

The authors would like to thank Fred Baker, Bill Jouris, Jose Isidro and James Ashton for their reviews and suggestions that made this document better.

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

Authors' Addresses

Nalini Elkins
Inside Products, Inc.
36A Upper Circle
Carmel Valley, CA
United States

Phone: +1 831 659 8360
Email: nalini.elkins@insidethestack.com

Lawrence Kratzke
IBM
8121 Glenbrittle Way
Raleigh, NC 27615
United States

Phone: +1 800-876-8801
Email: kratzke@us.ibm.com

Michael Ackermann
Blue Cross Blue Shield of Michigan
P.O. Box 2888
Detroit, Michigan 48231
United States

Phone: +1 310 460 4080
Email: mackermann@bcbsmi.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 31, 2011

W. Kumari
Google
I. Gashinsky
Yahoo!
J. Jaeggli
Zynga
June 29, 2011

Operational Neighbor Discovery Problems and Enhancements.
draft-gashinsky-v6nd-enhance-00

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 can be vulnerable to denial of service attacks 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 that scan networks for inventory and other purposes. 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 problem in detail and suggests possible implementation improvements as well as operational mitigation techniques that can in some cases to protect against such attacks. It also discusses possible modifications to the traditional [RFC4861] neighbor discovery protocol itself.

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 December 31, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

1. Introduction	4
1.1. Applicability	4
2. The Problem	4
3. Terminology	5
4. Background	6
5. Neighbor Discovery Overview	7
6. Operational Mitigation Options	7
6.1. Filtering of unused address space.	8
6.2. Appropriate Subnet Sizing.	8
6.3. Routing Mitigation.	8
6.4. Tuning of the NDP Queue Rate Limit.	9
7. Recommendations for Implementors.	9
7.1. Prioritize NDP Activities	10
7.2. Queue Tuning.	11
7.3. NDP Protocol Gratuitous NA	11
7.4. ND cache priming and refresh	12
8. IANA Considerations	13
9. Security Considerations	13
10. Acknowledgements	13
11. References	14
11.1. Normative References	14
11.2. Informative References	14
Appendix A. Text goes here.	14
Authors' Addresses	14

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 and suggests possible implementation improvements as well as operational mitigation techniques that can in some cases protect against such attacks. It also discusses possible modifications to the traditional [RFC4861] neighbor discovery protocol itself.

The RFC series documents generally describe on-the-wire 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 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 in 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 consideration 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 addresses. In contrast, the default IPv6 subnet size is a /64, a number so large it covers literally billions of billions of addresses, the overwhelming number of which will be unassigned. Consequently, simplistic implementations of Neighbor Discovery can be vulnerable to denial of service attacks whereby they 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 that scan networks for inventory and other purposes. 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 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 process of testing for the (non)existence of neighbors can induce a denial of service condition, where the number of Neighbor Discovery requests overwhelms the implementation's capacity to process them, exhausts available memory, replaces existing in-use mappings with incomplete entries that will never be completed, etc. The result can be network disruption, where existing traffic may be impacted, and devices that join the net find that address resolutions fails.

In order to alleviate 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 potential liabilities 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. Forwarding steps 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 the routing protocols, 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. The forwarding plane accesses the Neighbor Cache on every forwarded packet. Thus 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. Whenever 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 actions. In many implementations it 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, It 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 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).

An attacker sending the appropriate packets to addresses on a given subnet 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 unhappy and therefore forwarding unusable.

This behavior has been observed across multiple platforms and implementations.

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 send the packet to. 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.

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 stipulated that some of these options are "kludges", and are 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 portion. 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 nullrouting the subnet carrying a more specific /122 route.

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. Appropriate 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 subnet for inter-router links to only have 2 addresses. It is worth noting that this practice is common in IPv4 networks, partly to protect against the harmful effects of ARP flooding attacks.

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.

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, although, as with most tuning knobs (especially those that deal with rate limiting), you may be "completing the attack". 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) and, under attack conditions you may be unable to resolve "legitimate" addresses sooner than if you had just the the knob alone.

It is worth noting that this technique is only worth investigation if the device has separate queue for resolution of unknown addresses versus maintenance of existing entries.

7. Recommendations for Implementors.

The section provides some recommendations to implementors of IPv4 Neighbor Discovery.

At a high-level, implementors should program defensively. That is, they should assume that intruders 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.

1. It is critical to respond to Neighbor Solicitations for one's own address, especially when 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 may begin to react (negatively) to perceived connectivity problems, causing additional undesirable ripple effects.

4. Activities related to the sending and receiving of Router Advertisements also impact address resolutions. [XXX say more?]

5. 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 subnet queue limits at a rate below that of the global queue limit might limit the damage to the neighbor discovery process to the taret network.

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 also means that it will take the router longer to learn legitimate destinations. 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).

7.3. NDP Protocol Gratuitous NA

Per RFC 4861, section 7.2.5 and 7.2.6 [RFC4861] requires that unsolicited neighbor advertisements result in the receiver setting it's neighbor cache entry to STALE, kicking off the resolution of the neighbor using neighbor solicitation. If the link layer address in an unsolicited neighbor advertisement matches that of the existing ND cache entry, routers SHOULD retain the existing entry updating it's status with regards to LRU retention policy.

Hosts MAY be configured to send unsolicited Neighbor advertisement at a rate set at the discretion of the operators. The rate SHOULD be appropriate to the sizing of ND cache parameters and the host count on the subnet. An unsolicited NA rate parameter MUST NOT be enabled by default. The unsolicted rate interval as interpreted by hosts must jitter the value for the interval between transmissions. Hosts receiving a neighbor solicitation requests from a router following each of three subsequent gratuitous NA intervals MUST revert to RFC

4861 behavior.

Implementation of new behavior for unsolicited neighbor advertisement would make it possible under appropriate circumstances to greatly reduce the dependence on the neighbor solicitation process for retaining existing ND cache entries.

This may impact the detection of one-way reachability.

It is understood that this section may need to be moved into a separate document -- it is (currently) provided here for discussion purposes.

7.4. ND cache priming and refresh

With all of the above recommendations implemented, it should be possible to survive a "scan attack" with very little impact to the network, however, adding new hosts to the network (and the sending of traffic to them) may still be negatively impacted. Traffic to those new hosts would have to go through the unknown Neighbor Resolution queue, which is where the attack traffic would end up as well. A solution to this would be that any new host that joins the network would "announce" itself, and be added to the cache, therefore not requiring packets destined to it to go through the unknown NDP queue. This could be done by sending a ping packet to the all-routers multicast address, which would then trigger the router's own neighbor resolution process, which should be in a different queue than other packets.

All attempts should be made to keep these addresses in cache, since any eviction of legitimate hosts from the cache could potentially place resolutions for them into the same queue as the attack traffic. At present, [RFC4861] states that there should be MAX_UNICAST_SOLICIT (3) attempts, RETRANS_TIMER 1 second apart, so if there is an interruption in the network or control plane processing for longer than 3 seconds during the refresh, the entry would be evicted from the ND Cache. Any network event which takes longer than 3 seconds to converge (UDLD, STP, etc may take 30+ seconds) while under an attack, would result in ND cache eviction. If an entry is evicted during a scan, connectivity could be lost for an extended period of time.

NDP refresh timers could be revised as suggested in draft-nordmark-6man-impatient-nud-00 [1] and SHOULD have a configurable value for MAX_UNICAST_SOLICIT and RETRANS_TIMER, and include capabilities for binary/exponential backoff.

A suggested algorithm, which retains backward compatibility with [RFC4861] is: operator configurable values for MAX_UNICAST_SOLICIT,

RETRANS_TIMER, and a way to set adaptive back-off multiple, similar to ipv4 -- call it BACKOFF_MULTIPLE), so that we could implement:

```
next_retrans =  
($BACKOFF_MULTIPLE^$solicit_attempt_num)*$RETRANS_TIMER + jittered  
value.
```

The recommended behavior is to have 5 attempts, with timing spacing of 0 (initial request), 1 second later, 3 seconds later, then 9, and then 27, which represents:

```
MAX_UNICAST_SOLICIT=5
```

```
RETRANS_TIMER=1 (default)
```

```
BACKOFF_MULTIPLE=3
```

If BACKOFF_MULTIPLE=1 (which should be the default value), and MAX_UNICAST_SOLICIT=3, you would get the backwards-compatible RFC behavior, but operators should be able to adjust the values as necessary to insure that they are sufficiently aggressive about retaining ND entries in cache.

An Implementation following this algorithm would if the request was not answered at first due for example to a transitory condition, retry immediately, and then back off for progressively longer periods. This would allow for a reasonably fast resolution time when the transitory condition clears.

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 techniques create additional security or DOS exposure

10. Acknowledgements

The authors would like to thank Ron Bonica, Troy Bonin, John Jason Brzozowski, Randy Bush, Vint Cerf, Jason Fesler Erik Kline, Jared

Mauch, Chris Morrow and Suran De Silva. Special thanks to Thomas Narten 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.
- [RFC4398] Josefsson, S., "Storing Certificates in the Domain Name System (DNS)", RFC 4398, March 2006.
- [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

- [RFC4255] Schlyter, J. and W. Griffin, "Using DNS to Securely Publish Secure Shell (SSH) Key Fingerprints", RFC 4255, January 2006.

URIs

- [1] <<http://tools.ietf.org/html/draft-nordmark-6man-impatient-nud-00>>

Appendix A. Text goes here.

TBD

Authors' Addresses

Warren Kumari
Google

Email: warren@kumari.net

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

Email: igor@yahoo-inc.com

Joel
Zynga
111 Evelyn
Sunnyvale, CA
USA

Email: jjaeggli@zynga.com

IPv6 Operations Working Group (v6ops)
Internet-Draft
Intended status: Informational
Expires: December 10, 2011

F. Gont
UK CPNI
June 8, 2011

IPv6 Router Advertisement Guard (RA-Guard) Evasion
draft-gont-v6ops-ra-guard-evasion-01

Abstract

The IPv6 Router Advertisement Guard (RA-Guard) mechanism is commonly employed to mitigate attack vectors based on forged ICMPv6 Router Advertisement messages. Many existing IPv6 deployments rely on RA-Guard as the first line of defense against the aforementioned attack vectors. This document describes possible ways in which current RA-Guard implementations can be circumvented, and discusses possible mitigations.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

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 December 10, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Router Advertisement Guard (RA Guard) Evasion Vulnerability	4
2.1. Attack Vector based on IPv6 Extension Headers	4
2.2. Attack vector based on IPv6 fragmentation	4
3. Mitigations	8
4. Other Implications	9
5. Security Considerations	10
6. Acknowledgements	11
7. References	12
7.1. Normative References	12
7.2. Informative References	12
Appendix A. Changes from previous versions of the draft (to be removed by the RFC Editor before publication of this document as a RFC	13
A.1. Changes from draft-gont-v6ops-ra-guard-evasion-00	13
Appendix B. Assessment tools	14
Appendix C. Advice and guidance to vendors	15
Author's Address	16

1. Introduction

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

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

As part of the project "Security Assessment of the Internet Protocol version 6 (IPv6)" [CPNI-IPv6], we have devised two techniques for circumventing the RA-Guard protection, which are described in the following sections of this document. These techniques, and the corresponding tools to assess their effectiveness, had so far been made available only to vendors, in the hopes that they could implement counter-measures before they were publicly disclosed. However, since there has been some public discussion about these issues, it was deemed as appropriate to publish the present document.

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

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

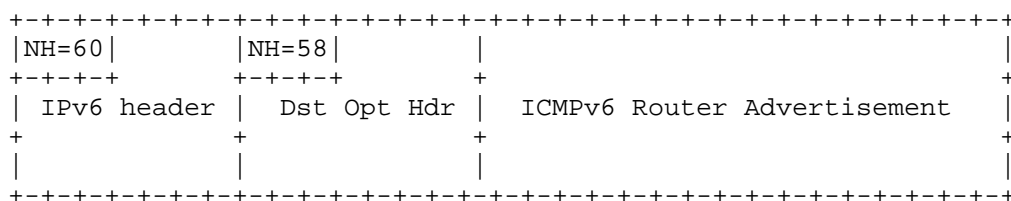
2. Router Advertisement Guard (RA Guard) Evasion Vulnerability

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

2.1. Attack Vector based on IPv6 Extension Headers

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

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



2.2. Attack vector based on IPv6 fragmentation

While the attack vector described in Section 2.1 may be effective with implementations that fail to process the entire header chain, it can easily be mitigated by an RA-Guard implementation, since all the information needed to identify ICMPv6 Router Advertisement messages is present in the attack packets.

This section presents a different attack vector, which aims at making it virtually impossible for a layer-2 device to identify ICMPv6 Router Advertisements by leveraging the IPv6 Fragment Header. The basic idea behind this attack vector is that if the forged ICMPv6 Router Advertisement is fragmented into at least two fragments, the layer-2 device implementing "RA-Guard" would be unable to identify the attack packet, and would thus fail to block it.

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

Original packet:

```

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

```

First fragment:

```

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

```

Second fragment:

```

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

```

It should be noted that the "Hdr Ext Len" field of the Destination

Options Header is present in the first fragment (rather than the second). Therefore, it would be impossible for a device processing only the second fragment to locate the ICMPv6 header contained in that fragment, since it is unknown how many bytes should be "skipped" to get to the next header following the Destination Options Header.

Thus, by leveraging the use of the Fragment Header together with the use of the Destination Options header, the attacker is able to conceal the type and contents of the ICMPv6 message he is sending (an ICMPv6 Router Advertisement in this example). Unless the layer-2 device were to implement IPv6 fragment reassembly, it would be impossible for the device to identify the ICMPv6 type of the message.

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

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

Original packet:

```

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

```

First fragment:

```

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

```

Second fragment:

```

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

```

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

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

3. Mitigations

The most effective and efficient mitigation for the RA-Guard evasion vulnerability discussed in this document would be to prohibit the use of IPv6 Extension Headers in Neighbor Discovery messages, as proposed in [I-D.gont-6man-nd-extension-headers].

Nevertheless, an administrator might want to mitigate these vulnerabilities by deploying more advanced filtering. The following filtering rules could be implemented as part of an "RA-Guard" implementation, such that the vulnerabilities discussed in this document can be mitigated:

- o When trying to identify an ICMPv6 Router Advertisement message, follow the IPv6 header chain, enforcing a limit on the maximum number of Extension Headers that is allowed for each packet. If such limit is exceeded, block the packet.
- o If the layer-2 device is unable to identify whether the packet is an ICMPv6 Router Advertisement message or not (i.e., the packet is a fragment, and the necessary information is missing), and the IPv6 Source Address of the packet is a link-local address or the unspecified address (::), block the packet.
- o In all other cases, pass the packet as usual.

This filtering policy assumes that host implementations require that the IPv6 Source Address of ICMPv6 Router Advertisement messages be a link-local address, and that they discard the packet if this check fails, as required by the current IETF specifications [RFC4861]. Unfortunately, it should be noted that the aforementioned filtering policy might be inefficient to implement (if at all possible), and might also result (at least in theory) in false positives.

4. Other Implications

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

5. Security Considerations

This document describes a number of techniques to circumvent a mechanism known as "RA-Guard", which many organizations deploy as a "first line of defense" against attacks based on forged Router Advertisements.

The most effective and efficient mitigation for these attacks would be to prohibit the use of IPv6 extension headers (as proposed by [I-D.gont-6man-nd-extension-headers]), such that the RA-Guard protection cannot be easily circumvented. However, since this mitigation requires an update to existing implementations, in the short term some network administrators might want to mitigate these issues by implemented the more advanced filtering policy described in Section 3.

6. Acknowledgements

The author would like to thank Karl Auer, Robert Downie, David Farmer, Marc Heuse, and Arturo Servin, for providing valuable comments on earlier versions of this document.

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

7. References

7.1. 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.

7.2. Informative References

- [RFC6104] Chown, T. and S. Venaas, "Rogue IPv6 Router Advertisement Problem Statement", RFC 6104, February 2011.
- [RFC6105] Levy-Abegnoli, E., Van de Velde, G., Popoviciu, C., and J. Mohacsi, "IPv6 Router Advertisement Guard", RFC 6105, February 2011.
- [I-D.gont-6man-nd-extension-headers]
Gont, F. and U. CPNI, "Security Implications of the Use of IPv6 Extension Headers with IPv6 Neighbor Discovery", draft-gont-6man-nd-extension-headers-00 (work in progress), May 2011.
- [CPNI-IPv6]
Gont, F., "Security Assessment of the Internet Protocol version 6 (IPv6)", UK Centre for the Protection of National Infrastructure, (to be published).
- [NDPMon] "NDPMon - IPv6 Neighbor Discovery Protocol Monitor", <<http://ndpmon.sourceforge.net/>>.
- [rafixd] "rafixd", <<http://www.kame.net/dev/cvsweb2.cgi/kame/kame/kame/rafixd/>>.
- [ramond] "ramond", <<http://ramond.sourceforge.net/>>.
- [THC-IPV6]
"THC-IPV6", <<http://www.thc.org/thc-ipv6/>>.

Appendix A. Changes from previous versions of the draft (to be removed by the RFC Editor before publication of this document as a RFC)

A.1. Changes from draft-gont-v6ops-ra-guard-evasion-00

- o Minor editorial changes
- o The discussion of the challenge represented by a combination of fragmentation and Destination Options headers was improved/clarified.
- o In Section 2.2, in the illustration of the second variant of the attack (fragmentation combined with two Destination Options headers), the figure corresponding to the "first fragment" was corrected.
- o Clarified the filtering rules in Section 3.

Appendix B. Assessment tools

CPNI has produced assessment tools, which have not yet been made publicly available. If you think that you would benefit from these tools to assess the security of your network or of your RA-Guard implementation, we might be able to provide a copy of the tools (please contact Fernando Gont at fernando@gont.com.ar).

[THC-IPV6] is a publicly-available set of tools that implements some of the techniques described in this document.

Appendix C. Advice and guidance to vendors

Vendors are urged to contact CSIRTUK (csirt@cpni.gsi.gov.uk) if they think they may be affected by the issues described in this document. As the lead coordination centre for these issues, CPNI is well placed to give advice and guidance as required.

CPNI works extensively with government departments and agencies, commercial organisations and the academic community to research vulnerabilities and potential threats to IT systems especially where they may have an impact on Critical National Infrastructure's (CNI).

Other ways to contact CPNI, plus CPNI's PGP public key, are available at <http://www.cpni.gov.uk>.

Author's Address

Fernando Gont
Centre for the Protection of National Infrastructure

Email: fernando@gont.com.ar
URI: <http://www.gont.com.ar>

V6OPS WG
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2012

S. Gundavelli
Cisco
July 3, 2011

Reserved IPv6 Interface Identifier for Proxy Mobile IPv6
draft-gundavelli-v6ops-pmipv6-address-reservations-00.txt

Abstract

Proxy Mobile IPv6 [RFC5213] requires all the mobile access gateways to use a fixed link-local and link-layer addresses on any of its access links that it shares with the mobile nodes. This was intended to ensure a mobile node does not detect any change with respect to its layer-3 attachment even after it roams from one mobile access gateway to another. In the absence of any reserved addresses for this use, it requires coordination across vendors and the manual configuration of these addresses on all the mobility elements in a Proxy Mobile IPv6 domain. This document attempts to simplify this operational requirement by making reservation for special addresses that can be used for this purpose.

Status of this Memo

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

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

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

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

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions and Terminology	3
2.1. Conventions	3
2.2. Terminology	4
3. IANA Considerations	4
4. Security Considerations	4
5. Acknowledgements	4
6. References	4
6.1. Normative References	4
6.2. Informative References	5
Author's Address	5

1. Introduction

Proxy Mobile IPv6 [RFC5213] is a network-based mobility management protocol that enables IP mobility support for a mobile node without requiring its participation in any mobility-related signaling. The mobility elements in the network ensure that the mobile node does not detect any change with respect to its layer-3 attachment even after it roams from one mobile access gateway to another and changes its point of attachment in the network. All the mobile access gateways in a Proxy Mobile IPv6 use a fixed link-local address and link-layer address on any of its access links that they share with the mobile nodes. It essentially ensures a mobile node after performing an handoff does not detect any link change.

The base Proxy Mobile IPv6 [RFC5213] all though required the use of a fixed link-local and a fixed layer-layer address, it did not reserve any specific addresses for this purpose and this is proving to be a operational challenge in deployments involving multi-vendor equipment. To address this problem, this specification makes the following two reservations. The mobility elements in the Proxy Mobile IPv6 domain MAY choose to use these fixed addresses.

This specification reserves an IPv6 interface identifier for Proxy Mobile IPv6 [RFC5213]. The reserved IPv6 interface identifier can be used by all the mobile access gateways in a Proxy Mobile IPv6 domain on any of its access links that it shares with the mobile node. The mobile access gateway can use this reserved IPv6 interface identifier for generating the link-local address that it uses in the Neighbor Discovery [RFC4861] related communication with the mobile node. This fixed identifier needs to be reserved from the registry, "Reserved IPv6 Interface Identifiers".

Furthermore, this specification also reserves a IANA Ethernet unicast address for Proxy Mobile IPv6 use. This reserved link-layer address can be used by the mobile access gateway in a Proxy Mobile IPv6 domain, in the layer-2 header for all packet communication with the mobile node.

2. Conventions and Terminology

2.1. Conventions

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

2.2. Terminology

All the mobility related terms used in this document are to be interpreted as defined in the base Proxy Mobile IPv6 specifications [RFC5213], [RFC5844]. All the IPv6 addressing related terminology is to be interpreted as specified in [RFC4291].

3. IANA Considerations

This document requires the following two IANA actions.

- o Action-1: This specification reserves an IPv6 interface identifier for Proxy Mobile IPv6 [RFC5213]. This fixed identifier needs to be reserved from the registry, "Reserved IPv6 Interface Identifiers".
- o Action-2: This specification reserves a IANA Ethernet unicast address for Proxy Mobile IPv6. This address needs to be reserved from the block. "IANA Ethernet Address block - Unicast Use".

4. Security Considerations

There are no additional security considerations known at this point of time, beyond what are identified in [RFC5213] and [RFC5453].

5. Acknowledgements

The author would like to thank Jari Arkko and Dave Thaler for all the discussions around the use of fixed link-local and link-layer address, during the standardization of Proxy Mobile IPv6 [RFC5213].

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.
- [RFC5213] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.

[RFC5453] Krishnan, S., "Reserved IPv6 Interface Identifiers", RFC 5453, February 2009.

6.2. Informative References

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

[RFC5844] Wakikawa, R. and S. Gundavelli, "IPv4 Support for Proxy Mobile IPv6", RFC 5844, May 2010.

Author's Address

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

Email: sgundave@cisco.com

V6OPS WG
Internet-Draft
Updates: 5213 (if approved)
Intended status: Standards Track
Expires: June 17, 2012

S. Gundavelli
Cisco
December 15, 2011

Reserved IPv6 Interface Identifier for Proxy Mobile IPv6
draft-gundavelli-v6ops-pmipv6-address-reservations-06.txt

Abstract

Proxy Mobile IPv6 [RFC5213] requires all the mobile access gateways to use a fixed link-local and link-layer addresses on any of its access links that it shares with the mobile nodes. This was intended to ensure a mobile node does not detect any change with respect to its layer-3 attachment even after it roams from one mobile access gateway to another. In the absence of any reserved addresses for this use, it requires coordination across vendors and the manual configuration of these addresses on all the mobility elements in a Proxy Mobile IPv6 domain. This document attempts to simplify this operational requirement by making reservation for special addresses that can be used for this purpose and it also updates RFC 5213.

Status of this Memo

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

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

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

This Internet-Draft will expire on June 17, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

Table of Contents

1. Introduction	3
2. Conventions & Terminology	4
2.1. Conventions	4
2.2. Terminology	4
3. IANA Considerations	4
4. Security Considerations	4
5. Acknowledgements	5
6. References	5
6.1. Normative References	5
6.2. Informative References	5
Author's Address	5

1. Introduction

Proxy Mobile IPv6 [RFC5213] is a network-based mobility management protocol that enables IP mobility support for a mobile node without requiring its participation in any mobility-related signaling. The mobility elements in the network ensure that the mobile node does not detect any change with respect to its layer-3 attachment even after it roams from one mobile access gateway to another and changes its point of attachment in the network. All the mobile access gateways in a Proxy Mobile IPv6 use a fixed link-local address and a fixed link-layer address on any of its access links that they share with the mobile nodes. This essentially ensures a mobile node after performing an handoff does not detect any change with respect to the IP network configuration.

Although the base Proxy Mobile IPv6 specification [RFC5213] requires the use of a fixed link-local and a fixed link-layer address, it did not reserve any specific addresses for this purpose and this is proving to be a operational challenge in deployments involving multi-vendor equipment. To address this problem, this specification makes the following two reservations.

1. This specification reserves one single Ethernet unicast address, (IANA-TBD1), for the use of Proxy Mobile IPv6. This reserved link-layer address SHOULD be used by the mobile access gateway in a Proxy Mobile IPv6 domain, on all of the access links that it shares with the mobile nodes. The protocol configuration variable, FixedMAGLinkLayerAddressOnAllAccessLinks [RFC5213], SHOULD be set to this reserved address. The mobile access gateway can use this address in all packet communication with the mobile node on the access links. Considerations from [RFC5342] apply with respect to the use of Ethernet parameters in IETF protocols. This address is allocated from the registry, "IANA Ethernet Address block - Unicast Use".
2. This specification reserves an IPv6 interface identifier, (IANA-TBD2). This interface identifier is a modified EUI-64 interface identifier generated from the allocated Ethernet unicast address (IANA-TBD1). The reserved IPv6 interface identifier SHOULD be used by all the mobile access gateways in a Proxy Mobile IPv6 domain on all of the access links that it shares with the mobile nodes. The protocol configuration variable, FixedMAGLinkLocalAddressOnAllAccessLinks [RFC5213], SHOULD be set to the link-local address generated using this reserved IPv6 interface identifier. The mobile access gateway can use this link-local address generated using this reserved IPv6 interface identifier in all Neighbor Discovery [RFC4861] related communication with the mobile node.

2. Conventions & Terminology

2.1. Conventions

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

2.2. Terminology

All the mobility related terms used in this document are to be interpreted as defined in the base Proxy Mobile IPv6 specifications [RFC5213], [RFC5844]. All the IPv6 addressing related terminology is to be interpreted as specified in [RFC4291].

3. IANA Considerations

This document requires the following two IANA actions.

- o Action-1: This specification reserves one single Ethernet unicast address, (IANA-TBD1), for Proxy Mobile IPv6. This address needs to be reserved from the block. "IANA Ethernet Address block - Unicast Use".
- o Action-2: This specification reserves an IPv6 interface identifier (IANA-TBD2) for Proxy Mobile IPv6 [RFC5213] from the registry, "Reserved IPv6 Interface Identifiers" [RFC5453]. This interface identifier is a modified EUI-64 interface identifier generated from the allocated Ethernet unicast address (IANA-TBD1) as specified in Appendix A of [RFC4291].

4. Security Considerations

All the security considerations specified in [RFC5213], and [RFC5844] continue to apply to the mobility elements in a Proxy Mobile IPv6 domain, when enabled to conform to this specification. Specifically, the issues related to the use of fixed link-local and link-layer address documented in section 6.9.3 of the base Proxy Mobile IPv6 specification are equally relevant here. In some sense, the reservations made in this specification results in the use of the same set of link-local and link-layer address values beyond a single Proxy Mobile IPv6 domain, thereby expanding the scope of the existing problem related to asserting ownership on the configured addresses from a single domain to multi-domain. Future work may be needed to address these issues.

5. Acknowledgements

The author would like to thank Jari Arkko and Dave Thaler for all the discussions around the use of fixed link-local and link-layer address, during the standardization of Proxy Mobile IPv6 [RFC5213]. The authors would also like to thank Tero Kivinen, Donald Eastlake 3rd, Stephen Farrell, Suresh Krishnan, Margaret Wasserman, Thomas Narten, Basavaraj Patil and Eric Voit for their reviews and participations in the discussions related to this document.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.
- [RFC5213] Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., and B. Patil, "Proxy Mobile IPv6", RFC 5213, August 2008.
- [RFC5453] Krishnan, S., "Reserved IPv6 Interface Identifiers", RFC 5453, February 2009.

6.2. Informative References

- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC5342] Eastlake, D., "IANA Considerations and IETF Protocol Usage for IEEE 802 Parameters", BCP 141, RFC 5342, September 2008.
- [RFC5844] Wakikawa, R. and S. Gundavelli, "IPv4 Support for Proxy Mobile IPv6", RFC 5844, May 2010.

Author's Address

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

Email: sgundave@cisco.com

v6ops Working Group
Internet-Draft
Intended status: Informational
Expires: January 5, 2012

N. Hilliard
INEX
July 4, 2011

A Discard Prefix for IPv6
draft-hilliard-v6ops-ipv6-discard-prefix-00

Abstract

Remote triggered black hole filtering describes a method of mitigating against denial-of-service attacks by selectively discarding traffic based on source or destination address. This document explains why a unique IPv6 prefix should be formally assigned by IANA for the purpose of facilitating IPv6 remote triggered black hole filtering.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. A Discard Prefix for IPv6	3
3. Operational Implications	4
4. IANA Considerations	4
5. Security Considerations	4
6. References	4
6.1. Normative References	4
6.2. Informative References	4
Author's Address	5

1. Introduction

Remote triggered black hole (RTBH) filtering describes a class of methods of blocking IP traffic to or from a specific destination on a network. These methods operate by setting the next-hop address of an IP packet with a specified source or destination address to be a unicast prefix which is wired locally or remotely to a router's discard or null interface. Typically, this information is propagated throughout an autonomous system using a dynamic routing protocol. By deploying RTBH systems across a network, traffic to or from specific destinations may be selectively black-holed in a manner which is efficient, scalable and straightforward to implement. For IPv4, some networks configure RTBH installations using [RFC1918] address space or the address blocks reserved for documentation in [RFC5737].

However RTBH configurations are not documentation, but operationally important features of many public-facing production networks. Furthermore, [RFC3849] specifies that the IPv6 documentation prefix should be filtered in both local and public contexts. On this basis, it is suggested that both private network address blocks and documentation prefixes described in [RFC5737] are inappropriate for the purpose of RTBH configurations.

While it could be argued that there are other addresses and address prefixes which could be used for this purpose (e.g. `::/128`), or that an operator could assign an address block from their own address space for this purposes, there is currently no operational clarity on what address block would be appropriate or inappropriate to use for this purpose. By creating an assigned discard prefix for IPv6, the IETF will introduce operational clarity and good practice for implementation of IPv6 RTBH configurations.

2. A Discard Prefix for IPv6

For the purposes of implementing an IPv6 remote triggered black hole filter, a unicast address block is required. There are currently no IPv6 unicast address blocks which are specifically nominated for the purposes of implementing RTBH filters.

As [RFC5635] describes situations where more than one discard address may be used for implementing multiple remote triggered black holes, a single assigned prefix is not sufficient to cover all likely RTBH filtering situations. Consequently, an address block is required.

The prefix allocated by IANA for the purpose of implementing IPv6 remote triggered black holes is `xxx::/32`.

3. Operational Implications

This assignment MAY be carried in a dynamic routing protocol within an autonomous system. The assignment SHOULD NOT be announced to third party autonomous systems and IPv6 traffic with an destination address within this prefix SHOULD NOT be forwarded to third party autonomous systems.

On networks which implement IPv6 remote triggered black holes, some or all of this network block MAY be configured with a destination of a discard or null interface on any or all IPv6 routers within the autonomous system.

4. IANA Considerations

IANA is requested to assign a unique /32 unicast address prefix in the IPv6 address registry for the purpose of facilitating remote triggered black hole configurations.

5. Security Considerations

IPv6 addressing documents do not have any direct impact on Internet infrastructure security.

6. References

6.1. Normative References

- [RFC5635] Kumari, W. and D. McPherson, "Remote Triggered Black Hole Filtering with Unicast Reverse Path Forwarding (uRPF)", RFC 5635, August 2009.

6.2. 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.
- [RFC3849] Huston, G., Lord, A., and P. Smith, "IPv6 Address Prefix Reserved for Documentation", RFC 3849, July 2004.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5737] Arkko, J., Cotton, M., and L. Vegoda, "IPv4 Address Blocks Reserved for Documentation", RFC 5737, January 2010.

Author's Address

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

v6ops WG
Internet-Draft
Obsoletes: 3056, 3068
(if approved)
Intended status: Informational
Expires: December 26, 2011

O. Troan
Cisco
June 24, 2011

Request to move Connection of IPv6 Domains via IPv4 Clouds (6to4) to
Historic status
draft-ietf-v6ops-6to4-to-historic-05.txt

Abstract

Experience with the "Connection of IPv6 Domains via IPv4 Clouds (6to4)" IPv6 transitioning mechanism has shown that the mechanism is unsuitable for widespread deployment and use in the Internet. This document requests that RFC3056 and the companion document "An Anycast Prefix for 6to4 Relay Routers" RFC3068 are made obsolete and moved to historic status.

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 December 26, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

1. Introduction

There would appear to be no evidence of any substantial deployment of the variant of 6to4 described in [RFC3056]. Its extension specified in "An Anycast Prefix for 6to4 Relay Routers" [RFC3068] has been shown to have severe practical problems when used in the Internet. This document requests that RFC3056 and RFC3068 be moved to Historic status as defined in section 4.2.4 [RFC2026].

6to4 was designed to help transition the Internet from IPv4 to IPv6. It has been a good mechanism for experimenting with IPv6, but because of the high failure rates seen with 6to4 [HUSTON], end users may end up disabling IPv6 on hosts, and content providers are reluctant to make content available over IPv6.

[I-D.ietf-v6ops-6to4-advisory] analyses the known operational issues and describes a set of suggestions to improve 6to4 reliability, given the widespread presence of hosts and customer premises equipment that support it.

The IETF sees no evolutionary future for the mechanism and it is not recommended to include this mechanism in new implementations.

IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) [RFC5969] utilizes the same encapsulation and base mechanism as 6to4, and could be viewed as a superset of 6to4 (6to4 could be achieved by setting the 6rd prefix to 2002::/16). However, the deployment model is such that 6rd can avoid the problems described here. In this sense, 6rd can be viewed as superseding 6to4 as described in section 4.2.4 of [RFC2026]

2. Conventions

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

3. 6to4 operational problems

6to4 is a mechanism designed to allow isolated IPv6 islands to reach

each other using IPv6 over IPv4 automatic tunneling. To reach the native IPv6 Internet the mechanism uses relay routers both in the forward and reverse direction. The mechanism is supported in many IPv6 implementations. With the increased deployment of IPv6, the mechanism has been shown to have a number of fundamental shortcomings.

6to4 depends on relays both in the forward and reverse direction to enable connectivity with the native IPv6 Internet. A 6to4 node will send IPv4 encapsulated IPv6 traffic to a 6to4 relay, that is connected both to the 6to4 cloud and to native IPv6. In the reverse direction a 2002::/16 route is injected into the native IPv6 routing domain to attract traffic from native IPv6 nodes to a 6to4 relay router. It is expected that traffic will use different relays in the forward and reverse direction. RFC3068 adds an extension that allows the use of a well known IPv4 anycast address to reach the nearest 6to4 relay in the forward direction.

One model of 6to4 deployment as described in section 5.2, RFC3056, suggests that a 6to4 router should have a set of managed connections (via BGP connections) to a set of 6to4 relay routers. While this makes the forward path more controlled, it does not guarantee a functional reverse path. In any case this model has the same operational burden as manually configured tunnels and has seen no deployment in the public Internet.

List of some of the known issues with 6to4:

- o Use of relays. 6to4 depends on an unknown third- party to operate the relays between the 6to4 cloud and the native IPv6 Internet.
- o The placement of the relay can lead to increased latency, and in the case the relay is overloaded, packet loss.
- o There is generally no customer relationship between the end-user and the relay operator, or even a way for the end-user to know who the relay operator is, so no support is possible.
- o A 6to4 relay for the reverse path and an anycast 6to4 relay used for the forward path, are openly accessible, limited only by the scope of routing. 6to4 relays can be used to anonymize traffic and inject attacks into IPv6 that are very difficult to trace.
- o 6to4 may silently discard traffic in the case where protocol (41) is blocked in intermediate firewalls. Even if a firewall sent an ICMP message unreachable back, an IPv4 ICMP message rarely contains enough of the original IPv6 packet so that it can be relayed back to the IPv6 sender. That makes this problem hard to detect and react upon by the sender of the packet.
- o As 6to4 tunnels across the Internet, the IPv4 addresses used must be globally reachable. RFC3056 states that a private address [RFC1918] MUST NOT be used. 6to4 will not work in networks that

employ other addresses with limited topological span.

4. Deprecation

This document formally deprecates the 6to4 transition mechanism and the IPv6 6to4 prefix defined in [RFC3056], i.e., 2002::/16. The prefix MUST NOT be reassigned for other use except by a future IETF standards action.

Disabling 6to4 in the IPv6 Internet will take some time. The initial approach is to make 6to4 a service of "last resort" in host implementations, ensure that the 6to4 service is disabled by default in 6to4 routers, and deploy native IPv6 services. In order to limit the impact of end-users, it is recommended that operators retain their existing 6to4 relay routers and follow the recommendations found in [I-D.ietf-v6ops-6to4-advisory]. When traffic levels diminish, these routers can be decommissioned.

IPv6 nodes SHOULD treat 6to4 as a service of "last resort" as recommended in [I-D.ietf-6man-rfc3484-revise]

Implementations capable of acting as 6to4 routers SHOULD NOT enable 6to4 without explicit user configuration. In particular, enabling IPv6 forwarding on a device, SHOULD NOT automatically enable 6to4.

Existing implementations and deployments MAY continue to use 6to4.

The references to 6to4 should be removed as soon as practical from the revision of the Special-Use IPv6 Addresses [RFC5156].

The references to the 6to4 relay anycast addresses (192.88.99.0/24) should be removed as soon as practical from the revision of the Special Use IPv4 addresses [RFC5735].

Incidental references to 6to4 should be removed from other IETF documents if and when they are updated. These documents include RFC3162, RFC3178, RFC3790, RFC4191, RFC4213, RFC4389, RFC4779, RFC4852, RFC4891, RFC4903, RFC5157, RFC5245, RFC5375, RFC5971, and RFC6071.

5. IANA Considerations

IANA is requested to mark the 2002::/16 prefix as "deprecated", pointing to this document. Reassignment of the prefix for any usage requires justification via an IETF Standards Action [RFC5226].

The delegation of the 2.0.0.2.ip6.arpa domain [RFC5158] should be left in place. Redelelegation of the domain for any usage requires justification via an IETF Standards Action [RFC5226].

IANA is requested to mark the 192.88.99.0/24 prefix [RFC3068] as "deprecated", pointing to this document. Redelelegation of the domain for any usage requires justification via an IETF Standards Action [RFC5226].

6. Security Considerations

There are no new security considerations pertaining to this document. General security issues with tunnels are listed in [I-D.ietf-v6ops-tunnel-security-concerns] and more specifically to 6to4 in [RFC3964] and [I-D.ietf-v6ops-tunnel-loops].

7. Acknowledgements

The authors would like to acknowledge Tore Anderson, Dmitry Anipko, Jack Bates, Cameron Byrne, Ben Campbell, Gert Doering, Ray Hunter, Joel Jaeggli, Kurt Erik Lindqvist, Jason Livingood, Keith Moore, Tom Petch, Daniel Roesen and Mark Townsley, James Woodyatt, for their contributions and discussions on this topic.

Special thanks go to Fred Baker, Geoff Huston, Brian Carpenter, and Wes George for their significant contributions.

Many thanks to Gunter Van de Velde for documenting the harm caused by non-managed tunnels and to stimulate the creation of this document.

8. References

8.1. Normative References

- [RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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.

- [RFC5156] Blanchet, M., "Special-Use IPv6 Addresses", RFC 5156, April 2008.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5735] Cotton, M. and L. Vegoda, "Special Use IPv4 Addresses", BCP 153, RFC 5735, January 2010.

8.2. Informative References

- [HUSTON] Huston, "Flailing IPv6", December 2010,
<<http://www.potaroo.net/ispcol/2010-12/6to4fail.html>>.
- [I-D.ietf-6man-rfc3484-revise]
Matsumoto, A., Kato, J., and T. Fujisaki, "Update to RFC 3484 Default Address Selection for IPv6",
draft-ietf-6man-rfc3484-revise-03 (work in progress),
June 2011.
- [I-D.ietf-v6ops-6to4-advisory]
Carpenter, B., "Advisory Guidelines for 6to4 Deployment",
draft-ietf-v6ops-6to4-advisory-02 (work in progress),
June 2011.
- [I-D.ietf-v6ops-tunnel-loops]
Nakibly, G. and F. Templin, "Routing Loop Attack using IPv6 Automatic Tunnels: Problem Statement and Proposed Mitigations", draft-ietf-v6ops-tunnel-loops-07 (work in progress), May 2011.
- [I-D.ietf-v6ops-tunnel-security-concerns]
Krishnan, S., Thaler, D., and J. Hoagland, "Security Concerns With IP Tunneling",
draft-ietf-v6ops-tunnel-security-concerns-04 (work in progress), October 2010.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC3964] Savola, P. and C. Patel, "Security Considerations for 6to4", RFC 3964, December 2004.
- [RFC5158] Huston, G., "6to4 Reverse DNS Delegation Specification", RFC 5158, March 2008.

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

Author's Address

Ole Troan
Cisco
Oslo,
Norway

Email: ot@cisco.com

v6ops WG
Internet-Draft
Obsoletes: 3068, 6732 (if approved)
Intended status: Best Current Practice
Expires: August 1, 2015

O. Troan
Cisco
B. Carpenter, Ed.
Univ. of Auckland
January 28, 2015

Deprecating Anycast Prefix for 6to4 Relay Routers
draft-ietf-v6ops-6to4-to-historic-11.txt

Abstract

Experience with the "Connection of IPv6 Domains via IPv4 Clouds (6to4)" IPv6 transition mechanism defined in RFC 3056 has shown that when used in its anycast mode, the mechanism is unsuitable for widespread deployment and use in the Internet. This document therefore requests that RFC 3068, "An Anycast Prefix for 6to4 Relay Routers", be made obsolete and moved to historic status. It also obsoletes RFC 6732 "6to4 Provider Managed Tunnels". It recommends that future products should not support 6to4anycast and that existing deployments should be reviewed. This complements the guidelines in RFC 6343.

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 1, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

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

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Related Work	3
2. Conventions	3
3. 6to4 operational problems	3
4. Deprecation	4
5. Implementation Recommendations	5
6. Operational Recommendations	5
7. IANA Considerations	6
8. Security Considerations	6
9. Acknowledgements	6
10. References	7
10.1. Normative References	7
10.2. Informative References	8
Authors' Addresses	8

1. Introduction

The original form of the 6to4 transition mechanism [RFC3056] relies on unicast addressing. However, its extension specified in "An Anycast Prefix for 6to4 Relay Routers" [RFC3068] has been shown to have severe practical problems when used in the Internet. This document requests that RFC 3068 and RFC 6732 be moved to Historic status as defined in section 4.2.4 of [RFC2026]. It complements the deployment guidelines in [RFC6343].

6to4 was designed to help transition the Internet from IPv4 to IPv6. It has been a good mechanism for experimenting with IPv6, but because of the high failure rates seen with anycast 6to4 [HUSTON], end users may end up disabling IPv6 on hosts as a result, and in the past some content providers were reluctant to make content available over IPv6 for this reason.

[RFC6343] analyses the known operational issues in detail and describes a set of suggestions to improve 6to4 reliability, given the widespread presence of hosts and customer premises equipment that support it. The advice to disable 6to4 by default has been widely adopted in recent operating systems, and the failure modes have been widely hidden from users by many browsers adopting the "Happy Eyeballs" approach [RFC6555].

Nevertheless, a measurable amount of 6to4 traffic is still observed by IPv6 content providers. The remaining successful users of anycast 6to4 are likely to be on hosts using the obsolete policy table [RFC3484], which prefers 6to4 above IPv4, and running without Happy Eyeballs. Furthermore, they must have a route to an operational anycast relay and they must be accessing an IPv6 host that has a route to an operational return relay.

However, experience shows that operational failures caused by anycast 6to4 have continued, despite the advice in RFC 6343 being available.

1.1. Related Work

IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) [RFC5969] explicitly builds on the 6to4 mechanism, using a service provider prefix instead of 2002::/16. However, the deployment model is based on service provider support, such that 6rd avoids the problems observed with anycast 6to4.

The framework for 6to4 Provider Managed Tunnels [RFC6732] is intended to help a service provider manage 6to4 anycast tunnels. This framework only exists because of the problems observed with anycast 6to4.

2. Conventions

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

The word "deprecate" and its derivatives are used only in their generic sense of "criticize or express disapproval" and do not have any specific normative meaning. A deprecated function might exist in the Internet for many years to allow backwards compatibility.

3. 6to4 operational problems

6to4 is a mechanism designed to allow isolated IPv6 islands to reach each other using IPv6 over IPv4 automatic tunneling. To reach the native IPv6 Internet the mechanism uses relay routers both in the forward and reverse direction. The mechanism is supported in many IPv6 implementations. With the increased deployment of IPv6, the mechanism has been shown to have a number of shortcomings.

In the forward direction a 6to4 node will send IPv4 encapsulated IPv6 traffic to a 6to4 relay, that is connected both to the 6to4 cloud and to native IPv6. In the reverse direction a 2002::/16 route is

injected into the native IPv6 routing domain to attract traffic from native IPv6 nodes to a 6to4 relay router. It is expected that traffic will use different relays in the forward and reverse direction.

One model of 6to4 deployment, described in section 5.2 of RFC 3056, suggests that a 6to4 router should have a set of managed connections (via BGP connections) to a set of 6to4 relay routers. While this makes the forward path more controlled, it does not guarantee a functional reverse path. In any case this model has the same operational burden as manually configured tunnels and has seen no deployment in the public Internet.

RFC 3068 adds an extension that allows the use of a well known IPv4 anycast address to reach the nearest 6to4 relay in the forward direction. However, this anycast mechanism has a number of operational issues and problems, which are described in detail in Section 3 of [RFC6343]. This document is intended to deprecate the anycast mechanism.

Peer-to-peer usage of the 6to4 mechanism exists in the Internet, likely unknown to many operators. This usage is harmless to third parties and is not dependent on the anycast 6to4 mechanism that this document deprecates.

4. Deprecation

This document formally deprecates the anycast 6to4 transition mechanism defined in [RFC3068] and the associated anycast IPv4 address 192.88.99.1. It is no longer considered to be a useful service of last resort.

The prefix 192.88.99.0/24 MUST NOT be reassigned for other use except by a future IETF standards action.

The basic unicast 6to4 mechanism defined in [RFC3056] and the associated 6to4 IPv6 prefix 2002::/16 are not deprecated. The default address selection rules specified in [RFC6724] are not modified.

In the absence of 6to4 anycast, 6to4 Provider Managed Tunnels [RFC6732] will no longer be necessary, so they are also deprecated by this document.

Incidental references to 6to4 should be reviewed and possibly removed from other IETF documents if and when they are updated. These documents include RFC3162, RFC3178, RFC3790, RFC4191, RFC4213,

RFC4389, RFC4779, RFC4852, RFC4891, RFC4903, RFC5157, RFC5245, RFC5375, RFC5971, RFC6071 and RFC6890.

5. Implementation Recommendations

It is NOT RECOMMENDED to include the anycast 6to4 transition mechanism in new implementations. If included in any implementations, the anycast 6to4 mechanism MUST be disabled by default.

In host implementations, unicast 6to4 MUST also be disabled by default. All hosts using 6to4 MUST support the IPv6 address selection policy described in [RFC6724].

In router implementations, 6to4 MUST be disabled by default. In particular, enabling IPv6 forwarding on a device MUST NOT automatically enable 6to4.

6. Operational Recommendations

This document does not imply a recommendation for the generalized filtering of traffic or routes for 6to4 or even anycast 6to4. It simply recommends against further deployment of the anycast 6to4 mechanism, calls for current 6to4 deployments to evaluate the efficacy of continued use of the anycast 6to4 mechanism, and makes recommendations intended to prevent any use of 6to4 from hampering broader deployment and use of native IPv6 on the Internet as a whole.

Networks SHOULD NOT filter out packets whose source address is 192.88.99.1, because this is normal 6to4 traffic from a 6to4 return relay somewhere in the Internet. This includes ensuring that traffic from a local 6to4 return relay with a source address of 192.88.99.1 is allowed through anti-spoofing filters such as those described in [RFC2827] and [RFC3704] or through Unicast Reverse-Path-Forwarding (uRPF) checks [RFC5635].

The guidelines in Section 4 of [RFC6343] remain valid for those who choose to continue operating Anycast 6to4 despite its deprecation.

Current operators of an anycast 6to4 relay with the IPv4 address 192.88.99.1 SHOULD review the information in [RFC6343] and the present document, and then consider carefully whether the anycast relay can be discontinued as traffic diminishes. Internet service providers that do not operate an anycast relay but do provide their customers with a route to 192.88.99.1 SHOULD verify that it does in fact lead to an operational anycast relay, as discussed in Section 4.2.1 of [RFC6343]. Furthermore, Internet service providers and other network providers MUST NOT originate a route to

192.88.99.1, unless they actively operate and monitor an anycast 6to4 relay service as detailed in Section 4.2.1 of [RFC6343].

Operators of a 6to4 return relay responding to the IPv6 prefix 2002::/16 SHOULD review the information in [RFC6343] and the present document, and then consider carefully whether the return relay can be discontinued as traffic diminishes. To avoid confusion, note that nothing in the design of 6to4 assumes or requires that return packets are handled by the same relay as outbound packets. As discussed in Section 4.5 of RFC 6343, content providers might choose to continue operating a return relay for the benefit of their own residual 6to4 clients. Internet service providers SHOULD announce the IPv6 prefix 2002::/16 to their own customers if and only if it leads to a correctly operating return relay as described in RFC 6343. IPv6-only service providers, including those operating a NAT64 service [RFC6146], are advised that their own customers need a route to such a relay in case a residual 6to4 user served by a different service provider attempts to communicate with them.

Operators of 6to4 Provider Managed Tunnels [RFC6732] SHOULD carefully consider when this service can be discontinued as traffic diminishes.

7. IANA Considerations

The document creating the IANA IPv4 Special-Purpose Address Registry [RFC6890] included the 6to4 relay anycast prefix (192.88.99.0/24) as Table 10. Instead, IANA is requested to mark the 192.88.99.0/24 prefix originally defined by [RFC3068] as "Deprecated (6to4 Relay Anycast)", pointing to the present document. Redefinition of this prefix for any usage requires justification via an IETF Standards Action [RFC5226].

8. Security Considerations

There are no new security considerations pertaining to this document. General security issues with tunnels are listed in [RFC6169] and more specifically to 6to4 in [RFC3964] and [RFC6324].

9. Acknowledgements

The authors would like to acknowledge Tore Anderson, Mark Andrews, Dmitry Anipko, Jack Bates, Cameron Byrne, Ben Campbell, Lorenzo Colitti, Gert Doering, Nick Hilliard, Philip Homburg, Ray Hunter, Joel Jaeggli, Victor Kuarsingh, Kurt Erik Lindqvist, Jason Livingood, Jeroen Massar, Keith Moore, Tom Petch, Daniel Roesen, Mark Townsley and James Woodyatt for their contributions and discussions on this topic.

Special thanks go to Fred Baker, David Farmer, Wes George, and Geoff Huston for their significant contributions.

Many thanks to Gunter Van de Velde for documenting the harm caused by non-managed tunnels and stimulating the creation of this document.

10. References

10.1. Normative References

- [RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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.
- [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.
- [RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 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.
- [RFC6724] Thaler, D., Draves, R., Matsumoto, A., and T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, September 2012.
- [RFC6890] Cotton, M., Vegoda, L., Bonica, R., and B. Haberman, "Special-Purpose IP Address Registries", BCP 153, RFC 6890, April 2013.

10.2. Informative References

- [HUSTON] Huston, , "Flailing IPv6", December 2010, <<http://www.potaroo.net/ispcol/2010-12/6to4fail.html>>.
- [RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.
- [RFC3964] Savola, P. and C. Patel, "Security Considerations for 6to4", RFC 3964, December 2004.
- [RFC5635] Kumari, W. and D. McPherson, "Remote Triggered Black Hole Filtering with Unicast Reverse Path Forwarding (uRPF)", RFC 5635, August 2009.
- [RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.
- [RFC6169] Krishnan, S., Thaler, D., and J. Hoagland, "Security Concerns with IP Tunneling", RFC 6169, April 2011.
- [RFC6324] Nakibly, G. and F. Templin, "Routing Loop Attack Using IPv6 Automatic Tunnels: Problem Statement and Proposed Mitigations", RFC 6324, August 2011.
- [RFC6343] Carpenter, B., "Advisory Guidelines for 6to4 Deployment", RFC 6343, August 2011.
- [RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.
- [RFC6732] Kuarsingh, V., Lee, Y., and O. Vautrin, "6to4 Provider Managed Tunnels", RFC 6732, September 2012.

Authors' Addresses

Ole Troan
Cisco
Oslo
Norway

Email: ot@cisco.com

Brian Carpenter (editor)
Department of Computer Science
University of Auckland
PB 92019
Auckland 1142
New Zealand

Email: brian.e.carpenter@gmail.com

v6ops
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2012

D. Wing
A. Yourtchenko
Cisco
July 8, 2011

Happy Eyeballs: Success with Dual-Stack Hosts
draft-ietf-v6ops-happy-eyeballs-03

Abstract

When the IPv4 server and path is working but the IPv6 server or IPv6 path is down, a dual-stack client application experiences significant connection delay compared to an IPv4-only client. This is undesirable because it causes the dual-stack client to have a worse user experience. This document specifies requirements for algorithms that reduce this delay, and provides an example algorithm.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 9, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Notational Conventions	3
3. Problem Statement	3
3.1. URIs and hostnames	4
3.2. IPv6 connectivity	4
4. Algorithm Requirements	5
4.1. Adhere to Address Preference Policy	6
4.2. Behavior when Preferred Address Family has Failed	7
4.3. Reset on Network (re-)Initialization	7
4.4. Abandon Non-Winning Connections	7
5. Additional Considerations	8
5.1. Additional Network and Host Traffic	8
5.2. Determining Address Type	8
5.3. Debugging and Troubleshooting	8
5.4. Multiple Interfaces	9
5.5. Interaction with Same Origin Policy	9
5.6. Happy Eyeballs in an Operating System	9
6. Example Algorithm	9
7. Security Considerations	10
8. Acknowledgements	10
9. IANA Considerations	10
10. References	11
10.1. Normative References	11
10.2. Informational References	11
Appendix A. Changes	12
A.1. changes from -02 to -03	12
A.2. changes from -01 to -02	12
A.3. changes from -00 to -01	13
Authors' Addresses	13

1. Introduction

In order to use applications over IPv6, it is necessary that users enjoy nearly identical performance as compared to IPv4. A combination of today's applications, IPv6 tunneling, IPv6 service providers, and some of today's content providers all cause the user experience to suffer (Section 3). For IPv6, a content provider may ensure a positive user experience by using a DNS white list of IPv6 service providers who peer directly with them (e.g., [whitelist]). However, this does not scale well (to the number of DNS servers worldwide or the number of content providers worldwide), and does not react to intermittent network path outages.

Instead, applications can improve the user experience themselves, by more aggressively making connections on IPv6 and IPv4. There are a variety of algorithms that can be envisioned. This document specifies requirements for any such algorithm, with the goals that the network and servers are not inordinately harmed with a simple doubling of traffic on IPv6 and IPv4, and the host's address preference is honored (e.g., [RFC3484]).

2. Notational Conventions

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

3. Problem Statement

The basis of the IPv6/IPv4 selection problem was first described in 1994 in [RFC1671],

"The dual-stack code may get two addresses back from DNS; which does it use? During the many years of transition the Internet will contain black holes. For example, somewhere on the way from IPng host A to IPng host B there will sometimes (unpredictably) be IPv4-only routers which discard IPng packets. Also, the state of the DNS does not necessarily correspond to reality. A host for which DNS claims to know an IPng address may in fact not be running IPng at a particular moment; thus an IPng packet to that host will be discarded on delivery. Knowing that a host has both IPv4 and IPng addresses gives no information about black holes. A solution to this must be proposed and it must not depend on manually maintained information. (If this is not solved, the dual stack approach is no better than the packet translation approach.)"

As discussed in more detail in Section 3.1, it is important that the same URI and hostname be used for IPv4 and IPv6. Using separate namespaces (e.g., "ipv6.example.com") causes namespace fragmentation and reduces the ability for users to share URIs and hostnames, and complicates printed material that includes the URI or hostname.

As discussed in more detail in Section 3.2, IPv6 connectivity is broken to specific prefixes or specific hosts, or slower than native IPv4 connectivity.

3.1. URIs and hostnames

URIs are often used between users to exchange pointers to content -- such as on social networks, email, instant messaging, or other systems. Thus, production URIs and production hostnames containing references to IPv4 or IPv6 will only function if the other party is also using an application, OS, and a network that can access the URI or the hostname.

3.2. IPv6 connectivity

When IPv6 connectivity is impaired, today's IPv6-capable web browsers incur many seconds of delay before falling back to IPv4. This harms the user's experience with IPv6, which will slow the acceptance of IPv6, because IPv6 is frequently disabled in its entirety on the end systems to improve the user experience.

Reasons for such failure include no connection to the IPv6 Internet, broken 6to4 or Teredo tunnels, and broken IPv6 peering.

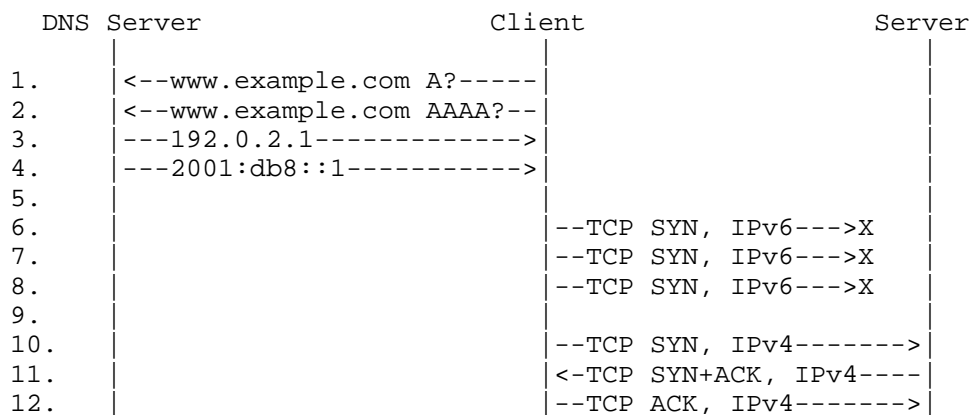


Figure 1: Existing behavior message flow

The client obtains the IPv4 and IPv6 records for the server (1-4).

The client attempts to connect using IPv6 to the server, but the IPv6 path is broken (6-8), which consumes several seconds of time. Eventually, the client attempts to connect using IPv4 (10) which succeeds.

Delays experienced by users of various browser and operating system combinations have been studied [Experiences].

4. Algorithm Requirements

A Happy Eyeballs algorithm has two primary goals:

1. Provides fast connection for users, by quickly attempting to connect using IPv6 and IPv4.
2. Avoids thrashing the network, by not always making simultaneous IPv6 and IPv4 connection attempts.

The basic idea is depicted in the following diagram:

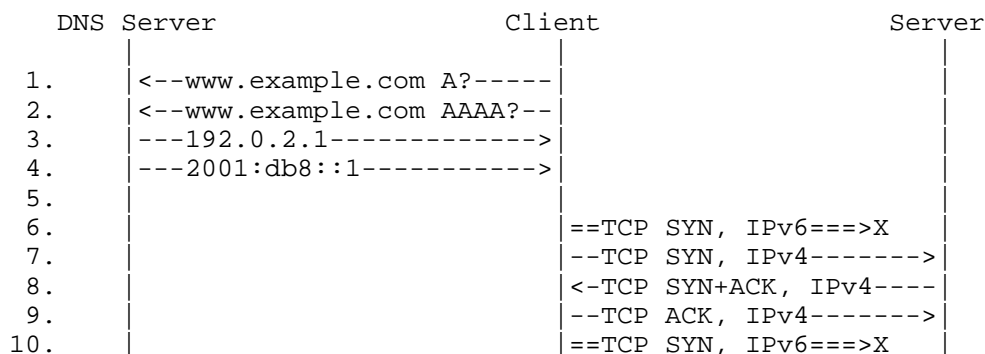


Figure 2: Happy Eyeballs flow 1, IPv6 broken

In the diagram above, the client sends two TCP SYNs at the same time over IPv6 (6) and IPv4 (7). In the diagram, the IPv6 path is broken but has little impact to the user because there is no long delay before using IPv4. The IPv6 path is retried until the application gives up (10).

After performing the above procedure, the client learns if connections to the host's IPv6 or IPv4 address were successful. The client MUST cache that information to avoid thrashing the network with excessive subsequent connection attempts. For example, in the diagram above, the client has noticed that IPv6 to that address failed, and it should provide a greater preference to using IPv4

instead.

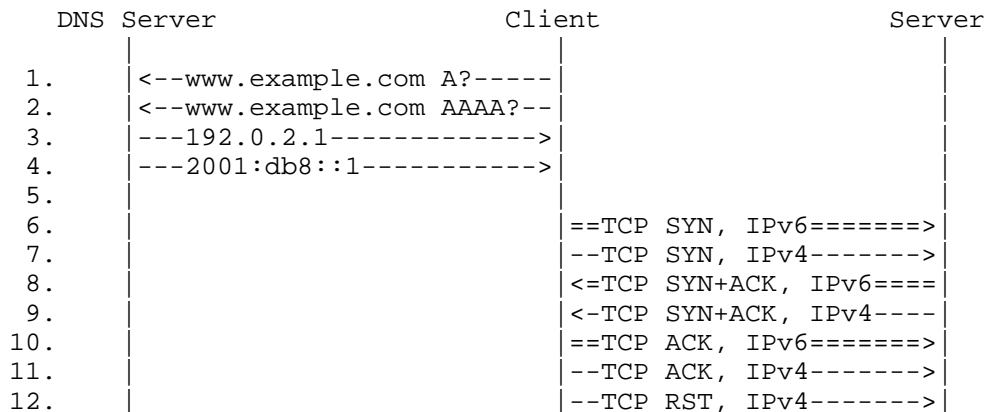


Figure 3: Happy Eyeballs flow 2, IPv6 working

The diagram above shows a case where both IPv6 and IPv4 are working, and IPv4 is abandoned (12).

Any Happy Eyeballs algorithm will persist in products for as long as the client host is dual-stacked, which will persist as long as there are IPv4-only servers on the Internet -- the so-called "long tail". Over time, as most content is available via IPv6, the amount of IPv4 traffic will decrease. This means that the IPv4 infrastructure will, over time, be sized to accomodate that decreased (and decreasing) amount of traffic. It is critical that a Happy Eyeballs algorithm not cause a surge of unnecessary traffic on that IPv4 infrastructure. To meet that goal, compliant Happy Eyeballs algorithms must adhere to the requirements in this section.

4.1. Adhere to Address Preference Policy

All hosts have an address selection policy. IPv6-capable hosts usually implement [RFC3484] and may allow the user (via configuration commands) or the network to modify that address selection policy (e.g., [I-D.ietf-6man-addr-select-opt]). In most cases, the preferred address family is IPv6.

Happy Eyeballs implementations MUST follow the host's address preference policy or, if that policy is unknown, implementations MUST prefer IPv6 over IPv4.

Justification: This reduces load on stateful IPv4 middleboxes (NAT and firewalls) and reduces IPv4 address sharing contention.

4.2. Behavior when Preferred Address Family has Failed

After making a connection attempt on a certain address family (e.g., IPv6), a Happy Eyeballs implementation will decide to initiate a second connection attempt using the other address family (e.g., IPv4).

After doing so and noticing that connections using the other address family (e.g., IPv4) are successful, a Happy Eyeballs implementation MAY make subsequent connection attempts on the successful address family (e.g., IPv4). Such an implementation MUST occasionally make connection attempts using the host's preferred address family, as it may have become functional. It is RECOMMENDED that implementations try the preferred address family at least every 10 minutes. Note: this can be achieved by connecting to both address families at the same time, which does not significantly harm the application's connection setup time for the successful address family. If connections using the preferred address family are successful, the preferred address family SHOULD be used for subsequent connections.

Justification: Once the IPv6 path becomes usable again, this reduces load on stateful IPv4 middleboxes (NAT and firewalls) and reduces IPv4 address sharing contention.

4.3. Reset on Network (re-)Initialization

Because every network has different characteristics (e.g., working or broken IPv6 or IPv4 connectivity), a Happy Eyeballs algorithm SHOULD re-initialize when the host is connected to a new network. Hosts can determine network (re-)initialization by a variety of mechanisms including DNaV4 [RFC4436], DNaV6 [RFC6059], [cx-osx], [cx-win].

Justification: This provides the best chance that IPv6 will be attempted over the new interface.

If the client application is a web browser, see also Section 5.5.

4.4. Abandon Non-Winning Connections

It is RECOMMENDED that the non-winning connections be abandoned, even though they could -- in some cases -- be put to reasonable use.

Justification: This reduces the load on the server (file descriptors, TCP control blocks), stateful middleboxes (NAT and firewalls) and, if the abandoned connection is IPv4, reduces IPv4 address sharing contention.

HTTP: The design of some sites can break because of HTTP cookies that incorporate the client's IP address and require all connections be from the same IP address. If some connections from the same client are arriving from different IP addresses (or worse, different IP address families), such applications will break. Additionally for HTTP, using the non-winning connection can interfere with the browser's Same Origin Policy (see Section 5.5).

5. Additional Considerations

This section discusses considerations and requirements that are common to new technology deployment.

5.1. Additional Network and Host Traffic

Additional network traffic and additional server load is created due to the recommendations in this document, especially when connections to the preferred address family (usually IPv6) are not completing quickly.

The procedures described in this document retain a quality user experience while transitioning from IPv4-only to dual stack, while still giving IPv6 a slight preference over IPv4 (in order to remove load from IPv4 networks, most importantly to reduce the load on IPv4 network address translators). The improvement in the user experience benefits the user to only a small detriment of the network, DNS server, and server that are serving the user.

5.2. Determining Address Type

For some transitional technologies such as a dual-stack host, it is easy for the application to recognize the native IPv6 address (learned via a AAAA query) and the native IPv4 address (learned via an A query). While IPv6/IPv4 translation makes that difficult, fortunately IPv6/IPv4 translators are not deployed on networks with dual stack clients.

5.3. Debugging and Troubleshooting

This mechanism is aimed at ensuring a reliable user experience regardless of connectivity problems affecting any single transport. However, this naturally means that applications employing these techniques are by default less useful for diagnosing issues with a particular address family. To assist in that regard, the implementations MAY also provide a mechanism to disable their Happy Eyeballs behavior via a user setting.

5.4. Multiple Interfaces

Interaction of the suggestions in this document with multiple interfaces, and interaction with the MIF working group, is for further study.

5.5. Interaction with Same Origin Policy

Web browsers implement same origin policy (SOP, [sop], [I-D.abarth-origin]), which causes subsequent connections to the same hostname to go to the same IPv4 (or IPv6) address as the previous successful connection. This is done to prevent certain types of attacks.

The same-origin policy harms user-visible responsiveness if a new connection fails (e.g., due to a transient event such as router failure or load balancer failure). While it is tempting to use Happy Eyeballs to maintain responsiveness, web browsers **MUST NOT** change their same origin policy because of Happy Eyeballs

5.6. Happy Eyeballs in an Operating System

Applications would have to change in order to use the mechanism described in this document, by either implementing the mechanism directly, or by calling APIs made available to them. To improve IPv6 connectivity experience for legacy applications (e.g., applications which simply rely on the operating system's address preference order), operating systems may consider more sophisticated approaches. These can include changing address sorting based on configuration received from the network, or observing connection failures to IPv6 and IPV4 destinations.

6. Example Algorithm

What follows is the algorithm implemented in Google Chrome and Mozilla Firefox.

1. Call `getaddrinfo()`, which returns a list of IP addresses sorted by the host's address preference policy.
2. Initiate a connection attempt with the first address in that list (e.g., IPv6).
3. If that connection does not complete within a short period of time (e.g., 200-300ms), initiate a connection attempt with the first address belonging to the other address family (e.g., IPv4)

4. The first connection that is established is used. The other connection is discarded.

Other example algorithms include [Perreault] and [Andrews].

7. Security Considerations

See Section 4.4 and Section 5.5.

8. Acknowledgements

The mechanism described in this paper was inspired by Stuart Cheshire's discussion at the IAB Plenary at IETF72, the author's understanding of Safari's operation with SRV records, Interactive Connectivity Establishment (ICE [RFC5245]), the current IPv4/IPv6 behavior of SMTP mail transfer agents, and the implementation of Happy Eyeballs in Google Chrome and Mozilla Firefox.

Thanks to Fred Baker, Jeff Kinzli, Christian Kuhtz, and Iljitsch van Beijnum for fostering the creation of this document.

Thanks to Scott Brim, Rick Jones, Stig Venaas, Erik Kline, Bjoern Zeeb, Matt Miller, Dave Thaler, and Dmitry Anipko for providing feedback on the document.

Thanks to Javier Ubillos, Simon Perreault and Mark Andrews for the active feedback and the experimental work on the independent practical implementations that they created.

Also the authors would like to thank the following individuals who participated in various email discussions on this topic: Mohacsi Janos, Pekka Savola, Ted Lemon, Carlos Martinez-Cagnazzo, Simon Perreault, Jack Bates, Jeroen Massar, Fred Baker, Javier Ubillos, Teemu Savolainen, Scott Brim, Erik Kline, Cameron Byrne, Daniel Roesen, Guillaume Leclanche, Mark Smith, Gert Doering, Martin Millnert, Tim Durack, Matthew Palmer.

9. IANA Considerations

This document has no IANA actions.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.

10.2. Informational References

- [Andrews] Andrews, M., "How to connect to a multi-homed server over TCP", January 2011, <<http://www.isc.org/community/blog/201101/how-to-connect-to-a-multi-homed-server-over-tcp>>.
- [Experiences] Savolainen, T., Miettinen, N., Veikkolainen, S., Chown, T., and J. Morse, "Experiences of host behavior in broken IPv6 networks", March 2011, <<http://www.ietf.org/proceedings/80/slides/v6ops-12.pdf>>.
- [I-D.abarth-origin] Barth, A., "The Web Origin Concept", draft-abarth-origin-09 (work in progress), November 2010.
- [I-D.ietf-6man-addr-select-opt] Matsumoto, A., Fujisaki, T., Kato, J., and T. Chown, "Distributing Address Selection Policy using DHCPv6", draft-ietf-6man-addr-select-opt-01 (work in progress), June 2011.
- [Perreault] Perreault, S., "Happy Eyeballs in Erlang", February 2011, <http://www.viagenie.ca/news/index.html#happy_eyeballs_erlang>.
- [RFC1671] Carpenter, B., "IPng White Paper on Transition and Other Considerations", RFC 1671, August 1994.
- [RFC4436] Aboba, B., Carlson, J., and S. Cheshire, "Detecting Network Attachment in IPv4 (DnAv4)", RFC 4436, March 2006.
- [RFC5245] Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", RFC 5245, April 2010.
- [RFC6059] Krishnan, S. and G. Daley, "Simple Procedures for Detecting Network Attachment in IPv6", RFC 6059,

November 2010.

- [cx-osx] Adium, "AIHostReachabilityMonitor", June 2009,
<https://bugzilla.redhat.com/show_bug.cgi?id=505105>.
- [cx-win] Microsoft, "NetworkChange.NetworkAvailabilityChanged
Event", June 2009, <[http://msdn.microsoft.com/en-us/
library/
system.net.networkinformation.networkchange.networkavailab
ilitychanged.aspx](http://msdn.microsoft.com/en-us/library/system.net.networkinformation.networkchange.networkavailabilitychanged.aspx)>.
- [sop] W3C, "Same Origin Policy", January 2010,
<http://www.w3.org/Security/wiki/Same-Origin_Policy>.
- [whitelist] Google, "Google IPv6 DNS Whitelist", January 2009,
<<http://www.google.com/intl/en/ipv6>>.

Appendix A. Changes

A.1. changes from -02 to -03

- o Re-casted this specification as a list of requirements for a compliant algorithm, rather than trying to dictate a One True algorithm.

A.2. changes from -01 to -02

- o Now honors host's address preference (RFC3484 and friends)
- o No longer requires thread-safe DNS library. It uses `getaddrinfo()`
- o No longer describes threading.
- o IPv6 is given a 200ms head start (Initial Headstart variable).
- o If the IPv6 and IPv4 connection attempts were made at nearly the same time, wait Tolerance Interval milliseconds for both to complete before deciding which one wins.
- o Renamed "global P" to "Smoothed P", and better described how it is calculated.
- o introduced the exception cache. This contains the set of networks that only work with IPv4 (or only with IPv6), so that subsequent connection attempts use that address family without them causing serious affect to Smoothed P.

- o encourages that every 10 minutes the exception cache and Smoothed P be reset. This allows IPv6 to be attempted again, so we don't get 'stuck' on IPv4.
- o If we didn't get both A and AAAA, abandon all Happy Eyeballs processing (thanks to Simon Perreault).
- o added discussion of Same Origin Policy
- o Removed discussion of NAT-PT and address learning; those are only used with IPv6-only hosts whereas this document is about dual-stack hosts contacting dual-stack servers.

A.3. changes from -00 to -01

- o added SRV section (thanks to Matt Miller)

Authors' Addresses

Dan Wing
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: dwing@cisco.com

Andrew Yourtchenko
Cisco Systems, Inc.
De Kleetlaan, 7
Diegem B-1831
Belgium

Email: ayourtch@cisco.com

v6ops
Internet-Draft
Intended status: Standards Track
Expires: June 22, 2012

D. Wing
A. Yourtchenko
Cisco
December 20, 2011

Happy Eyeballs: Success with Dual-Stack Hosts
draft-ietf-v6ops-happy-eyeballs-07

Abstract

When a server's IPv4 path and protocol is working but the server's IPv6 path and protocol are not working, a dual-stack client application experiences significant connection delay compared to an IPv4-only client. This is undesirable because it causes the dual-stack client to have a worse user experience. This document specifies requirements for algorithms that reduce this user-visible delay, and provides an algorithm.

Status of this Memo

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

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

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

This Internet-Draft will expire on June 22, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Additional Network and Host Traffic	3
2. Notational Conventions	3
3. Problem Statement	3
3.1. Hostnames	4
3.2. Delay When IPv6 is not Accessible	4
4. Algorithm Requirements	5
4.1. Delay IPv4	7
4.2. Stateful Behavior when IPv6 Fails	8
4.3. Reset on Network (re-)Initialization	9
4.4. Abandon Non-Winning Connections	9
5. Additional Considerations	10
5.1. Determining Address Type	10
5.2. Debugging and Troubleshooting	10
5.3. Three or More Interfaces	10
5.4. A and AAAA Resource Records	10
5.5. Connection time out	11
5.6. Interaction with Same Origin Policy	11
5.7. Implementation Strategies	11
6. Example Algorithm	12
7. Security Considerations	12
8. Acknowledgements	12
9. IANA Considerations	13
10. References	13
10.1. Normative References	13
10.2. Informational References	13
Appendix A. Changes	15
A.1. changes from -06 to -07	15
A.2. changes from -05 to -06	15
A.3. changes from -04 to -05	15
A.4. changes from -03 to -04	16
A.5. changes from -03 to -04	16
A.6. changes from -02 to -03	16
A.7. changes from -01 to -02	16
A.8. changes from -00 to -01	17
Authors' Addresses	17

1. Introduction

In order to use applications over IPv6, it is necessary that users enjoy nearly identical performance as compared to IPv4. A combination of today's applications, IPv6 tunneling, IPv6 service providers, and some of today's content providers all cause the user experience to suffer (Section 3). For IPv6, a content provider may ensure a positive user experience by using a DNS white list of IPv6 service providers who peer directly with them (e.g., [whitelist]). However, this does not scale well (to the number of DNS servers worldwide or the number of content providers worldwide), and does not react to intermittent network path outages.

Instead, applications reduce connection setup delays themselves, by more aggressively making connections on IPv6 and IPv4. There are a variety of algorithms that can be envisioned. This document specifies requirements for any such algorithm, with the goals that the network and servers are not inordinately harmed with a simple doubling of traffic on IPv6 and IPv4, and the host's address preference is honored (e.g., [RFC3484]).

1.1. Additional Network and Host Traffic

Additional network traffic and additional server load is created due to the recommendations in this document, especially when connections to the preferred address family (usually IPv6) are not completing quickly.

The procedures described in this document retain a quality user experience while transitioning from IPv4-only to dual stack, while still giving IPv6 a slight preference over IPv4 (in order to remove load from IPv4 networks, most importantly to reduce the load on IPv4 network address translators). The improvement in the user experience benefits the user to only a small detriment of the network, DNS server, and server that are serving the user.

2. Notational Conventions

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

3. Problem Statement

The basis of the IPv6/IPv4 selection problem was first described in 1994 in [RFC1671],

"The dual-stack code may get two addresses back from DNS; which does it use? During the many years of transition the Internet will contain black holes. For example, somewhere on the way from IPng host A to IPng host B there will sometimes (unpredictably) be IPv4-only routers which discard IPng packets. Also, the state of the DNS does not necessarily correspond to reality. A host for which DNS claims to know an IPng address may in fact not be running IPng at a particular moment; thus an IPng packet to that host will be discarded on delivery. Knowing that a host has both IPv4 and IPng addresses gives no information about black holes. A solution to this must be proposed and it must not depend on manually maintained information. (If this is not solved, the dual stack approach is no better than the packet translation approach.)"

As discussed in more detail in Section 3.1, it is important that the same hostname be used for IPv4 and IPv6.

As discussed in more detail in Section 3.2, IPv6 connectivity is broken to specific prefixes or specific hosts, or slower than native IPv4 connectivity.

The mechanism described in this document is directly applicable to connection-oriented transports (e.g., TCP, SCTP), which is the scope of this document. For connectionless transport protocols (e.g., UDP), a similar mechanism can be used if the application has request/response semantics (e.g., as done by ICE to select a working IPv6 or IPv4 media path [RFC6157]).

3.1. Hostnames

Hostnames are often used between users to exchange pointers to content -- such as on social networks, email, instant messaging, or other systems. Using separate namespaces (e.g., "ipv6.example.com") which are only accessible with certain client technology (e.g., an IPv6 client) and dependencies (e.g., a working IPv6 path) causes namespace fragmentation and reduces the ability for users to share hostnames. It also complicates printed material that includes the hostname.

The algorithm described in this document allows production hostnames to avoid these problematic references to IPv4 or IPv6.

3.2. Delay When IPv6 is not Accessible

When IPv6 connectivity is impaired, today's IPv6-capable applications (e.g., web browsers, email clients, instant messaging clients) incur many seconds of delay before falling back to IPv4. This delays

overall application operation, including harming the user's experience with IPv6, which will slow the acceptance of IPv6, because IPv6 is frequently disabled in its entirety on the end systems to improve the user experience.

Reasons for such failure include no connection to the IPv6 Internet, broken 6to4 or Teredo tunnels, and broken IPv6 peering. The following diagram shows this behavior.

The algorithm described in this document allows clients to connect to servers without significant delay, even if a path or the server is slow or down.

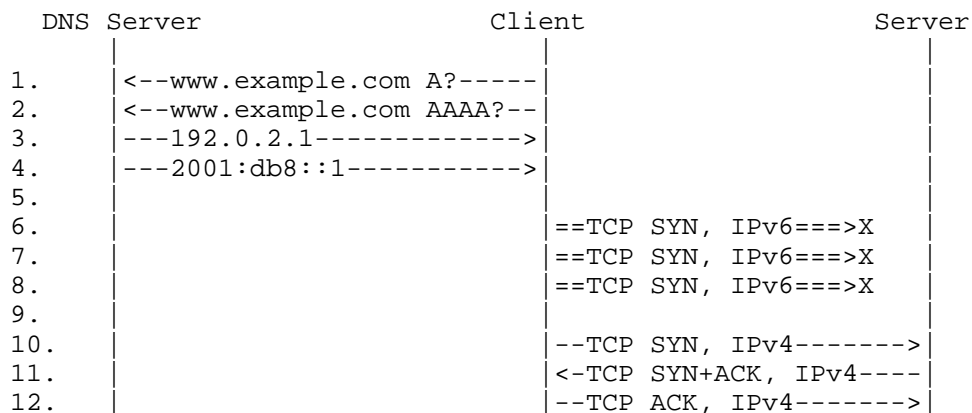


Figure 1: Existing behavior message flow

The client obtains the IPv4 and IPv6 records for the server (1-4). The client attempts to connect using IPv6 to the server, but the IPv6 path is broken (6-8), which consumes several seconds of time. Eventually, the client attempts to connect using IPv4 (10) which succeeds.

Delays experienced by users of various browser and operating system combinations have been studied [Experiences].

4. Algorithm Requirements

A Happy Eyeballs algorithm has two primary goals:

1. Provides fast connection for users, by quickly attempting to connect using IPv6 and (if that connection attempt is not quickly successful) to connect using IPv4.

2. Avoids thrashing the network, by not (always) making simultaneous connection attempts on both IPv6 and IPv4.

The basic idea is depicted in the following diagram:

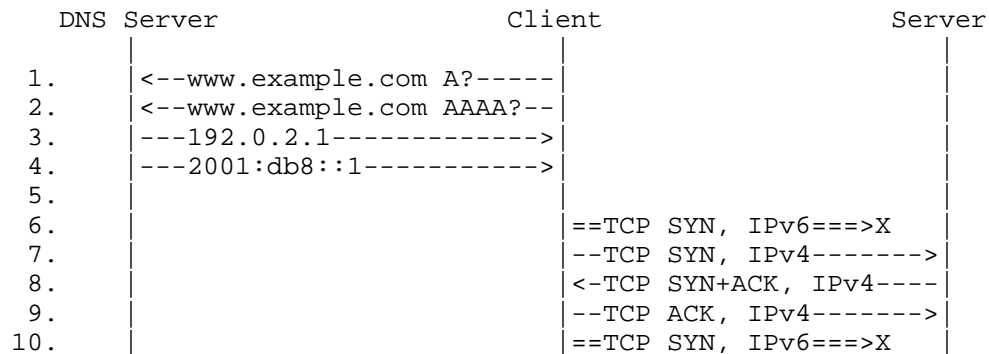


Figure 2: Happy Eyeballs flow 1, IPv6 broken

In the diagram above, the client sends two TCP SYNs at the same time over IPv6 (6) and IPv4 (7). In the diagram, the IPv6 path is broken but has little impact to the user because there is no long delay before using IPv4. The IPv6 path is retried until the application gives up (10).

After performing the above procedure, the client learns whether connections to the host's IPv6 or IPv4 address were successful. The client MUST cache information regarding the outcome of each connection attempt and uses that information to avoid thrashing the network with subsequent attempts. For example, in the example above, the cache indicates that the IPv6 connection attempt failed, and therefore the system will prefer IPv4 instead. Cache entries should be flushed when their age exceeds a system defined maximum on the order of ten minutes.

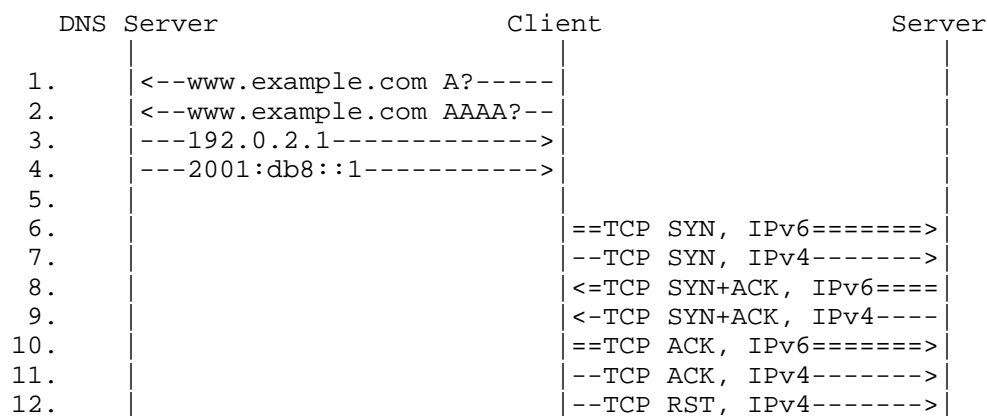


Figure 3: Happy Eyeballs flow 2, IPv6 working

The diagram above shows a case where both IPv6 and IPv4 are working, and IPv4 is abandoned (12).

Any Happy Eyeballs algorithm will persist in products for as long as the client host is dual-stacked, which will persist as long as there are IPv4-only servers on the Internet -- the so-called "long tail". Over time, as most content is available via IPv6, the amount of IPv4 traffic will decrease. This means that the IPv4 infrastructure will, over time, be sized to accommodate that decreased (and decreasing) amount of traffic. It is critical that a Happy Eyeballs algorithm not cause a surge of unnecessary traffic on that IPv4 infrastructure. To meet that goal, compliant Happy Eyeballs algorithms must adhere to the requirements in this section.

4.1. Delay IPv4

The transition to IPv6 is likely to produce a mix of different hosts within a subnetwork -- hosts that are IPv4-only, hosts that are IPv6-only (e.g., sensors), and dual-stack. This mix of hosts will exist both within an administrative domain (a single home, enterprise, hotel, or coffee shop) and between administrative domains. For example, a single home might have an IPv4-only television in one room and a dual-stack television in another room. As another example, another subscriber might have hosts that are all capable of dual-stack operation.

Due to IPv4 exhaustion, it is likely that a subscriber's hosts (both IPv4-only hosts and dual-stack hosts) will be sharing an IPv4 address with other subscribers. The dual-stack hosts have an advantage: they can utilize IPv6 or IPv4, which means it can utilize the technique described in this document. The IPv4-only hosts have a

disadvantage: they can only utilize IPv4. If all hosts (dual-stack and IPv4-only) are using IPv4, there is additional contention for the shared IPv4 address. The IPv4-only hosts cannot avoid that contention (as they can only use IPv4) while the dual-stack hosts can avoid that contention by using IPv6.

As dual-stack hosts proliferate and content becomes available over IPv6, there will be proportionally less IPv4 traffic. This is true especially for dual-stack hosts that do not implement Happy Eyeballs, because those dual-stack hosts have a very strong preference to use IPv6 (with timeouts in the tens of seconds before they will attempt to use IPv4).

When deploying IPv6, both content providers and Internet Service Providers (who supply IPv4 address sharing mechanisms such as Carrier Grade NAT (CGN)) will want to reduce their investment in IPv4 equipment -- load balancers, peering links, and address sharing devices. If a Happy Eyeballs implementation treats IPv6 and IPv4 equally by connecting to whichever address family is fastest, it will contribute to load on IPv4. This load impacts IPv4-only devices (by increasing contention of IPv4 address sharing and increasing load on IPv4 load balancers). Because of this, ISPs and content providers will find it impossible to reduce their investment in IPv4 equipment. This means that costs to migrate to IPv6 are increased, because the investment in IPv4 cannot be reduced. Furthermore, using only a metric that measures connection speed ignores the value of IPv6 over IPv4 address sharing, such as shared penalty boxes and geo-location [RFC6269].

Thus, to avoid harming IPv4-only hosts which can only utilize IPv4, implementations MUST prefer the first IP address family returned by the host's address preference policy, unless implementing a stateful algorithm described in Section 4.2. This usually means giving preference to IPv6 over IPv4, although that preference can be overridden by user configuration or by network configuration [I-D.ietf-6man-addr-select-opt]. If the host's policy is unknown or not attainable, implementations MUST prefer IPv6 over IPv4.

4.2. Stateful Behavior when IPv6 Fails

Some Happy Eyeballs algorithms are stateful -- that is, the algorithm will remember that IPv6 always fails, or that IPv6 to certain prefixes always fails, and so on. This section describes such algorithms. Stateless algorithms, which do not remember the success/failure of previous connections, are not discussed in this section.

After making a connection attempt on the preferred address family (e.g., IPv6), and failing to establish a connection within a certain

time period (see Section 5.5), a Happy Eyeballs implementation will decide to initiate a second connection attempt using the same address family or the other address family.

Such an implementation MAY make subsequent connection attempts (to the same host or to other hosts) on the successful address family (e.g., IPv4). So long as new connections are being attempted by the host, such an implementation MUST occasionally make connection attempts using the host's preferred address family, as it may have become functional again, and it SHOULD do so every 10 minutes. The 10 minute delay before re-trying a failed address family avoids the simple doubling of connection attempts on both IPv6 and IPv4. Implementation note: this can be achieved by flushing Happy Eyeballs state every every 10 minutes, which does not significantly harm the application's subsequent connection setup time. If connections using the preferred address family are again successful, the preferred address family SHOULD be used for subsequent connections. Because this implementation is stateful, it MAY track connection success (or failure) based on IPv6 or IPv4 prefix (e.g., connections to the same prefix assigned to the interface are successful whereas connections to other prefixes are failing).

4.3. Reset on Network (re-)Initialization

Because every network has different characteristics (e.g., working or broken IPv6 or IPv4 connectivity), a Happy Eyeballs algorithm SHOULD re-initialize when the interface is connected to a new network. Interfaces can determine network (re-)initialization by a variety of mechanisms (e.g., DNaV4 [RFC4436], DNaV6 [RFC6059]).

If the client application is a web browser, see also Section 5.6.

4.4. Abandon Non-Winning Connections

It is RECOMMENDED that the non-winning connections be abandoned, even though they could -- in some cases -- be put to reasonable use.

Justification: This reduces the load on the server (file descriptors, TCP control blocks), stateful middleboxes (NAT and firewalls) and, if the abandoned connection is IPv4, reduces IPv4 address sharing contention.

HTTP: The design of some sites can break because of HTTP cookies that incorporate the client's IP address and require all connections be from the same IP address. If some connections from the same client are arriving from different IP addresses (or worse, different IP address families), such applications will break. Additionally for HTTP, using the non-winning connection

can interfere with the browser's Same Origin Policy (see Section 5.6).

5. Additional Considerations

This section discusses considerations related to Happy Eyeballs.

5.1. Determining Address Type

For some transitional technologies such as a dual-stack host, it is easy for the application to recognize the native IPv6 address (learned via a AAAA query) and the native IPv4 address (learned via an A query). While IPv6/IPv4 translation makes that difficult, IPv6/IPv4 translators do not need to be deployed on networks with dual stack clients, because dual stack clients can use their native IP address family.

5.2. Debugging and Troubleshooting

This mechanism is aimed at ensuring a reliable user experience regardless of connectivity problems affecting any single transport. However, this naturally means that applications employing these techniques are by default less useful for diagnosing issues with a particular address family. To assist in that regard, the implementations MAY also provide a mechanism to disable their Happy Eyeballs behavior via a user setting, and to provide data useful for debugging (e.g., a log or way to review current preferences).

5.3. Three or More Interfaces

A dual-stack host normally has two logical interfaces: an IPv6 interface and an IPv4 interface. However, a dual-stack host might have more than two logical interfaces because of a VPN (where a third interface is the tunnel address, often assigned by the remote corporate network) or because of multiple physical interfaces such as wired and wireless Ethernet, because the host belongs to multiple VLANs, or other reasons. The interaction of Happy Eyeballs with more than two logical interfaces is for further study.

5.4. A and AAAA Resource Records

It is possible that an DNS query for an A or AAAA resource record will return more than one A or AAAA address. When this occurs, it is RECOMMENDED that a Happy Eyeballs implementation order the responses following the host's address preference policy and then try the first address. If that fails after a certain time (see Section 5.5), the next address SHOULD be the IPv4 address.

If that fails to connect after a certain time (see Section 5.5), a Happy Eyeballs implementation SHOULD try the other addresses returned; the order of these connection attempts is not important.

On the Internet today, servers commonly have multiple A records to provide load balancing across their servers. This same technique would be useful for AAAA records, as well. However, if multiple AAAA records are returned to a non-Happy Eyeballs client that has broken IPv6 connectivity, it will further increase the delay to fall back to IPv4. Thus, web site operators with native IPv6 connectivity SHOULD NOT offer multiple AAAA records. If Happy Eyeballs is widely deployed in the future, this recommendation might be revisited.

5.5. Connection time out

The primary purpose of Happy Eyeballs is to reduce the wait time for a dual stack connection to complete, especially when the IPv6 path is broken and IPv6 is preferred. Aggressive time outs (on the order of tens of milliseconds) achieve this goal, but at the cost of network traffic. This network traffic may be billable on certain networks, will create state on some middleboxes (e.g., firewalls, IDS, NAT), and will consume ports if IPv4 addresses are shared. For these reasons, it is RECOMMENDED that connection attempts be paced to give connections a chance to complete. It is RECOMMENDED that connections attempts be paced 150-250ms apart, to balance human factors against network load. Stateful algorithms are expected to be more aggressive (that is, make connection attempts closer together), as stateful algorithms maintain an estimate of the expected connection completion time.

5.6. Interaction with Same Origin Policy

Web browsers implement a Same Origin Policy [RFC6454] which causes subsequent connections to the same hostname to go to the same IPv4 (or IPv6) address as the previous successful connection. This is done to prevent certain types of attacks.

The same-origin policy harms user-visible responsiveness if a new connection fails (e.g., due to a transient event such as router failure or load balancer failure). While it is tempting to use Happy Eyeballs to maintain responsiveness, web browsers MUST NOT change their Same Origin Policy because of Happy Eyeballs, as that would create an additional security exposure.

5.7. Implementation Strategies

The simplest venue for implementation of Happy Eyeballs is within the application itself. The algorithm specified in this document is

relatively simple to implement, and would require no specific support from the operating system beyond the commonly-available APIs that provide transport service. It could also be added to applications by way of a specific Happy Eyeballs API, replacing or augmenting the transport service APIs.

To improve IPv6 connectivity experience for legacy applications (e.g., applications which simply rely on the operating system's address preference order), operating systems may consider more sophisticated approaches. These can include changing default address selection sorting ([RFC3484]) based on configuration received from the network, or observing connection failures to IPv6 and IPV4 destinations.

6. Example Algorithm

What follows is the algorithm implemented in Google Chrome and Mozilla Firefox.

1. Call `getaddinfo()`, which returns a list of IP addresses sorted by the host's address preference policy.
2. Initiate a connection attempt with the first address in that list (e.g., IPv6).
3. If that connection does not complete within a short period of time (Firefox and Chrome use 300ms), initiate a connection attempt with the first address belonging to the other address family (e.g., IPv4)
4. The first connection that is established is used. The other connection is discarded.

If an algorithm were to cache connection success/failure, the caching would occur after step 4 determined which connection was successful.

Other example algorithms include [Perreault] and [Andrews].

7. Security Considerations

See Section 4.4 and Section 5.6.

8. Acknowledgements

The mechanism described in this paper was inspired by Stuart

Cheshire's discussion at the IAB Plenary at IETF72, the author's understanding of Safari's operation with SRV records, Interactive Connectivity Establishment (ICE [RFC5245]), the current IPv4/IPv6 behavior of SMTP mail transfer agents, and the implementation of Happy Eyeballs in Google Chrome and Mozilla Firefox.

Thanks to Fred Baker, Jeff Kinzli, Christian Kuhtz, and Iljitsch van Beijnum for fostering the creation of this document.

Thanks to Scott Brim, Rick Jones, Stig Venaas, Erik Kline, Bjoern Zeeb, Matt Miller, Dave Thaler, Dmitry Anipko, Brian Carpenter, and David Crocker for their feedback.

Thanks to Javier Ubillos, Simon Perreault and Mark Andrews for the active feedback and the experimental work on the independent practical implementations that they created.

Also the authors would like to thank the following individuals who participated in various email discussions on this topic: Mohacsi Janos, Pekka Savola, Ted Lemon, Carlos Martinez-Cagnazzo, Simon Perreault, Jack Bates, Jeroen Massar, Fred Baker, Javier Ubillos, Teemu Savolainen, Scott Brim, Erik Kline, Cameron Byrne, Daniel Roesen, Guillaume Leclanche, Mark Smith, Gert Doering, Martin Millnert, Tim Durack, Matthew Palmer.

9. IANA Considerations

This document has no IANA actions.

10. References

10.1. Normative References

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

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

10.2. Informational References

[Andrews] Andrews, M., "How to connect to a multi-homed server over TCP", January 2011, <<http://www.isc.org/community/blog/201101/how-to-connect-to-a-multi-homed-server-over-tcp>>.

[Experiences]

Savolainen, T., Miettinen, N., Veikkolainen, S., Chown, T., and J. Morse, "Experiences of host behavior in broken IPv6 networks", March 2011, <<http://www.ietf.org/proceedings/80/slides/v6ops-12.pdf>>.

[I-D.ietf-6man-addr-select-opt]

Matsumoto, A., Fujisaki, T., Kato, J., and T. Chown, "Distributing Address Selection Policy using DHCPv6", draft-ietf-6man-addr-select-opt-01 (work in progress), June 2011.

[Perreault]

Perreault, S., "Happy Eyeballs in Erlang", February 2011, <http://www.viagenie.ca/news/index.html#happy_eyeballs_erlang>.

[RFC1671] Carpenter, B., "IPng White Paper on Transition and Other Considerations", RFC 1671, August 1994.

[RFC4436] Aboba, B., Carlson, J., and S. Cheshire, "Detecting Network Attachment in IPv4 (DIPv4)", RFC 4436, March 2006.

[RFC5245] Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", RFC 5245, April 2010.

[RFC6059] Krishnan, S. and G. Daley, "Simple Procedures for Detecting Network Attachment in IPv6", RFC 6059, November 2010.

[RFC6157] Camarillo, G., El Malki, K., and V. Gurbani, "IPv6 Transition in the Session Initiation Protocol (SIP)", RFC 6157, April 2011.

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

[RFC6454] Barth, A., "The Web Origin Concept", RFC 6454, December 2011.

[whitelist]

Google, "Google IPv6 DNS Whitelist", January 2009, <<http://www.google.com/intl/en/ipv6>>.

Appendix A. Changes

[RFC Editor: Please remove this section prior to publication as an RFC.]

A.1. changes from -06 to -07

- o Changed "xmpp clients" to "instant messaging clients".
- o For debugging/troubleshooting, providing a log of activity or a way to see current settings is useful.
- o tweaked abstract
- o "URIs and hostnames" -> "hostnames"
- o tweaked text on caching
- o interfaces (not hosts) notice when they are connected to a new network.
- o encourage implementations to provide log or other way to view Happy Eyeballs settings.
- o detailed that implementation can be in OS or in application.
- o 150-250ms is for human factors

A.2. changes from -05 to -06

- o Added paragraph describing current AAAA practice on the Internet (one AAAA record) due to non-Happy Eyeballs implementations, per opsdireview.
- o fixed "=" in Figure 1.
- o Removed text discussing A6. A6 is being deprecated in another document, and querying A6 is not a significant operational problem on the Internet.

A.3. changes from -04 to -05

- o Updated citations.

A.4. changes from -03 to -04

- o Make RFC3363 a non-normative reference.

A.5. changes from -03 to -04

- o Better explained why IPv6 needs to be preferred
- o Don't query A6.

A.6. changes from -02 to -03

- o Re-casted this specification as a list of requirements for a compliant algorithm, rather than trying to dictate a One True algorithm.

A.7. changes from -01 to -02

- o Now honors host's address preference (RFC3484 and friends)
- o No longer requires thread-safe DNS library. It uses getaddrinfo()
- o No longer describes threading.
- o IPv6 is given a 200ms head start (Initial Headstart variable).
- o If the IPv6 and IPv4 connection attempts were made at nearly the same time, wait Tolerance Interval milliseconds for both to complete before deciding which one wins.
- o Renamed "global P" to "Smoothed P", and better described how it is calculated.
- o introduced the exception cache. This contains the set of networks that only work with IPv4 (or only with IPv6), so that subsequent connection attempts use that address family without them causing serious affect to Smoothed P.
- o encourages that every 10 minutes the exception cache and Smoothed P be reset. This allows IPv6 to be attempted again, so we don't get 'stuck' on IPv4.
- o If we didn't get both A and AAAA, abandon all Happy Eyeballs processing (thanks to Simon Perreault).
- o added discussion of Same Origin Policy

- o Removed discussion of NAT-PT and address learning; those are only used with IPv6-only hosts whereas this document is about dual-stack hosts contacting dual-stack servers.

A.8. changes from -00 to -01

- o added SRV section (thanks to Matt Miller)

Authors' Addresses

Dan Wing
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: dwing@cisco.com

Andrew Yourtchenko
Cisco Systems, Inc.
De Kleetlaan, 7
Diegem B-1831
Belgium

Email: ayourtch@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 12, 2012

H. Singh
W. Beebe
Cisco Systems, Inc.
C. Donley
CableLabs
B. Stark
ATT
O. Troan, Ed.
Cisco Systems, Inc.
July 11, 2011

Advanced Requirements for IPv6 Customer Edge Routers
draft-ietf-v6ops-ipv6-cpe-router-bis-01

Abstract

This document continues the work undertaken by the IPv6 CE Router Phase I work in the IETF v6ops Working Group. Advanced requirements or Phase II work is covered in this document.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
2. Terminology	3
3. Conceptual Configuration Variables	4
4. Architecture	4
5. Advanced Features and Feature Requirements	6
5.1. DNS	6
5.2. Multicast Behavior	6
5.3. Routed network behavior	7
5.4. Transition Technologies Support	7
5.4.1. Dual-Stack(DS)-Lite	7
5.4.2. 6rd	9
5.4.3. Transition Technologies Coexistence	9
5.5. Quality Of Service	10
5.6. Unicast Data Forwarding	10
5.7. Additional DHCPv6 WAN Requirement	10
6. Security Considerations	10
7. Acknowledgements	10
8. Contributors	11
9. IANA Considerations	11
10. References	11
10.1. Normative References	11
10.2. Informative References	14
Authors' Addresses	14

1. Introduction

This document defines Advanced IPv6 features for a residential or small office router referred to as an IPv6 CE router. Typically these routers also support IPv4. The IPv6 End-user Network Architecture for such a router is described in [RFC6204]. This version of the document includes the requirements for Advanced features.

1.1. Requirements Language

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

2. Terminology

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 which 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 Ethernets (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. Conceptual Configuration Variables

The CE Router maintains such a list of conceptual optional configuration variables.

1. Enable an IGP on the LAN.
2. Configure 6rd configuration.
3. Configure IPv6 for 6rd to have IPv6 traffic go to the 6rd Border Relay vs. directly to peers.

4. Architecture

This document extends the architecture described in [RFC6204] to cover a strictly larger set of operational scenarios. In particular, QoS, multicast, DNS, routed network in the home, transition technologies, and conceptual configuration variables. This document also extends the model described in [RFC6204] to a two router topology where the two routers are connected back-to-back (the LAN of one router is connected to the WAN of the other router). This topology is depicted below:

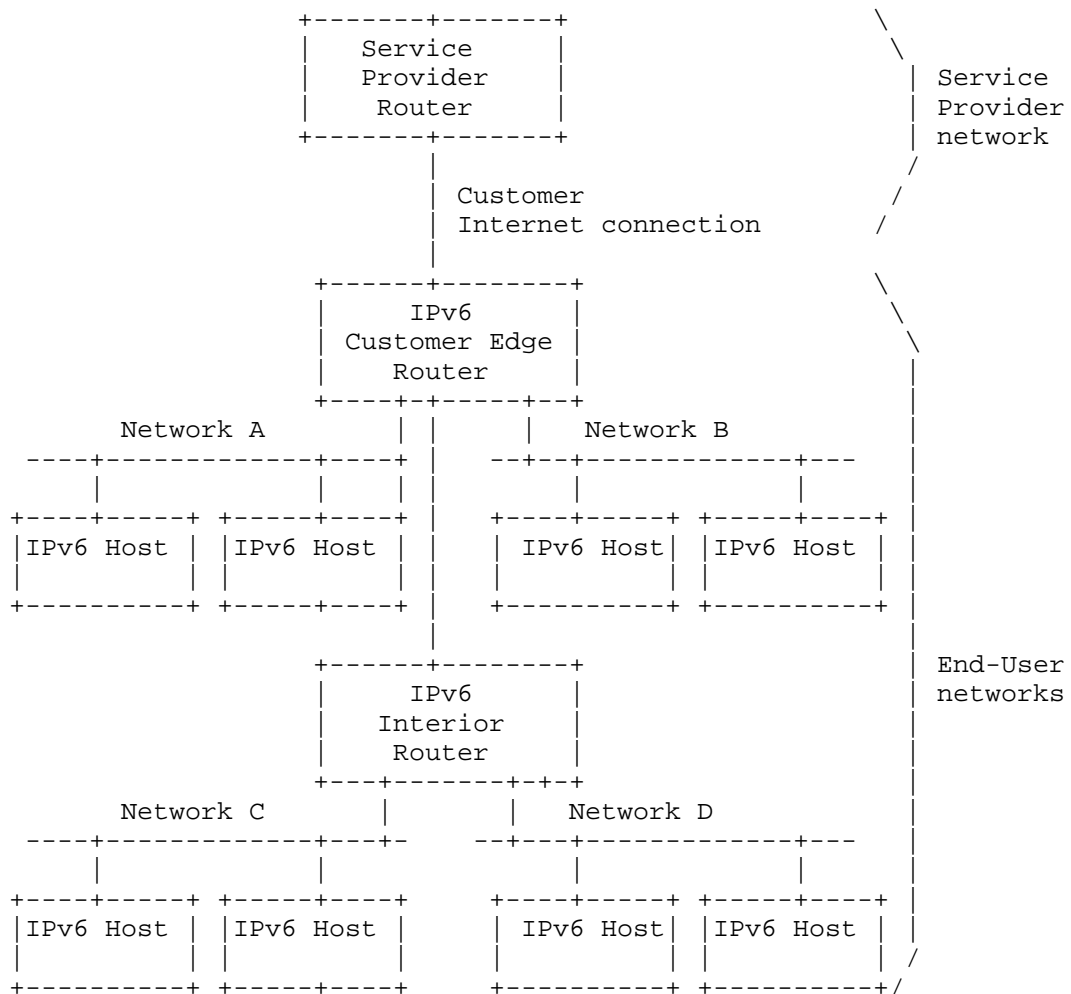


Figure 1.

For DNS, the operational expectation is that the end-user would be able to access home hosts from the home using DNS names instead of more cumbersome IPv6 addresses. Note that this is distinct from the requirement to access home hosts from outside the home.

End-users are expected to be able to receive multicast video in the home without requiring the CE router to include the cost of supporting full multicast routing protocols.

5. Advanced Features and Feature Requirements

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 supports all commonly used architectures.

5.1. DNS

D-1: The CE Router MAY include a DNS server authoritative for .local to handle local queries. If the service provider specifies one or more DNS resolvers in DHCP configuration options, the CE router SHOULD forward all non-local DNS queries unchanged to those servers. The CE Router MAY also include DNS64 functionality which is specified in [RFC6147].

5.2. Multicast Behavior

This section is only applicable to a CE Router with at least one LAN interface. A host in the home is expected to receive multicast video. Note the CE Router resides at edge of the home and the Service Provider, and the CE Router has at least one WAN connection for multiple LAN connections. In such a multiple LAN to a WAN topology at the CE Router edge, it is not necessary to run a multicast routing protocol and thus MLD Proxy as specified in [RFC4605] can be used. The CE Router discovers the hosts via a MLDv2 Router implementation on a LAN interface. A WAN interface of the CE Router interacts with the Service Provider router by sending MLD Reports and replying to MLD queries for multicast Group memberships for hosts in the home.

The CE router SHOULD implement MLD Proxy as specified in [RFC4605]. For the routed topology shown in Figure 1, each router implements a MLD Proxy. If the CE router implements MLD Proxy, the requirements on the CE Router for MLD Proxy are listed below.

WAN requirements, MLD Proxy:

WMLD-1: Consistent with [RFC4605], the CE router MUST NOT implement the router portion of MLDv2 for the WAN interface.

LAN requirements, MLD Proxy:

LMMLD-1: The CPE Router MUST follow the model described for MLD Proxy in [RFC4605] to implement multicast.

LMMLD-2: Consistent with [RFC4605], the LAN interfaces on the CPE router MUST NOT implement an MLDv2 Multicast Listener.

LAN requirements:

LM-1: If the CE Router has bridging configured between the LAN interfaces, then the LAN interfaces MUST support snooping of MLD [RFC3810] messages as per [RFC4541] .

5.3. Routed network behavior

CPE Router Behavior in a routed network:

R-1: One example of the CPE Router use in the home is shown below. The home has a broadband modem combined with a CPE Router, all in one device. The LAN interface of the device is connected to another standalone CPE Router that supports a wireless access point. To support such a network, this document recommends using prefix delegation of the prefix obtained either via IA_PD from WAN interface or a ULA from the LAN interface. The network interface of the downstream router MAY obtain an IA_PD via stateful DHCPv6. If the CPE router supports the routed network through a vendor specific automatic prefix delegation, the CPE router MUST support a DHCPv6 server or DHCPv6 relay agent. Further, if an IA_PD is used, the Service Provider or user MUST allocate an IA_PD or ULA prefix short enough to be delegated and subsequently used for SLAAC. Therefore, a prefix length shorter than /64 is needed. The CPE Router MAY support and IGP in the home network.

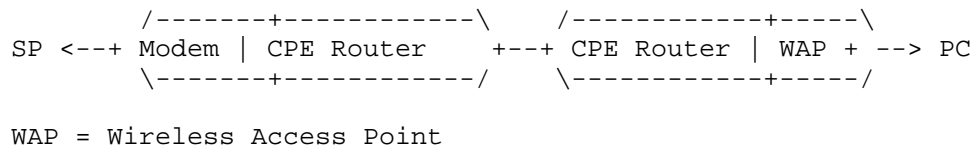


Figure 2.

5.4. Transition Technologies Support

5.4.1. Dual-Stack(DS)-Lite

Even as users migrate from IPv4 to IPv6 addressing, a significant percentage of Internet resources and content will remain accessible

only through IPv4. Also, many end-user devices will only support IPv4. As a consequence, Service Providers require mechanisms to allow customers to continue to access content and resources using IPv4 even after the last IPv4 allocations have been fully depleted. One technology that can be used for IPv4 address extension is DS-Lite.

DS-Lite enables a Service Provider to share IPv4 addresses among multiple customers by combining two well-known technologies: IP in IP (IPv4-in-IPv6) tunneling and Carrier Grade NAT. More specifically, Dual-Stack-Lite encapsulates IPv4 traffic inside an IPv6 tunnel at the IPv6 CE Router and sends it to a Service Provider Address Family Translation Router (AFTR). Configuration of the IPv6 CE Router to support IPv4 LAN traffic is outside the scope of this document.

The IPv6 CE Router SHOULD implement DS-Lite functionality as specified in [I-D.ietf-softwire-dual-stack-lite].

WAN requirements:

- DLW-1: To facilitate IPv4 extension over an IPv6 network, if the CE Router supports DS-Lite functionality, the CE Router WAN interface MUST implement a B4 Interface as specified in [I-D.ietf-softwire-dual-stack-lite].
- DLW-2: If the IPv6 CE Router implements DS-Lite functionality, the CE Router MUST support using a DS-Lite DHCPv6 option [I-D.ietf-softwire-ds-lite-tunnel-option] to configure the DS-Lite tunnel. The IPv6 CE Router MAY use other mechanisms to configure DS-Lite parameters. Such mechanisms are outside the scope of this document.
- DLW-3: IPv6 CE Router MUST NOT perform IPv4 Network Address Translation (NAT) on IPv4 traffic encapsulated using DS-Lite.
- DLW-4: If the IPv6 CE Router is configured with a public IPv4 address on its WAN interface, where public IPv4 address is defined as any address which is not in the private IP address space specified in [RFC1918] and also not in the reserved IP address space specified in [I-D.ietf-softwire-dual-stack-lite], then the IPv6 CE Router MUST disable the DS-Lite B4 element.
- DLW-5: If DS-Lite is operational on the IPv6 CE Router, multicast data MUST NOT be sent on any DS-Lite tunnel.

5.4.2. 6rd

The IPv6 CE Router can be used to offer IPv6 service to a LAN, even when the WAN access network only supports IPv4. One technology that supports IPv6 service over an IPv4 network is IPv6 Rapid Deployment (6rd). 6rd encapsulates IPv6 traffic from the end user LAN inside IPv4 at the IPv6 CE Router and sends it to a Service Provider Border Relay (BR). The IPv6 CE Router calculates a 6rd delegated IPv6 prefix during 6rd configuration, and sub-delegates the 6rd delegated prefix to devices in the LAN.

The IPv6 CE Router SHOULD implement 6rd functionality as specified in [RFC5969].

6rd requirements:

- 6RD-1: If the IPv6 CE Router implements 6rd functionality, the CE Router WAN interface MUST support at least one 6rd Virtual Interface and 6rd CE functionality as specified in [RFC5969].
- 6RD-2: If the IPv6 CE Router implements 6rd CE functionality, it MUST support user-entered configuration and using the 6rd DHCPv4 Option (212) for 6rd configuration. The IPv6 CE Router MAY use other mechanisms to configure 6rd parameters. Such mechanisms are outside the scope of this document.
- 6RD-3: If the CE router implements 6rd functionality, it MUST allow the user to specify whether all IPv6 traffic goes to the 6rd Border Relay, or whether other destinations within the same 6rd domain are routed directly to those destinations. The CE router MAY use other mechanisms to configure this. Such mechanisms are outside the scope of this document.
- 6RD-4: If 6rd is operational on the IPv6 CE Router, multicast data MUST NOT be sent on any 6rd tunnel.

5.4.3. Transition Technologies Coexistence

Run the following four in parallel to provision CPE router connectivity to the Service Provider:

1. Initiate IPv4 address acquisition.
2. Initiate IPv6 address acquisition as specified by [RFC6204].
3. If 6rd is provisioned, initiate 6rd.

4. If DS-Lite is provisioned, initiate DS-Lite.

The default route for IPv6 through the native physical interface should have preference over the 6rd tunnel interface. The default route for IPv4 through the native physical interface should have preference over the DS-Lite tunnel interface.

5.5. Quality Of Service

Q-1: The CPE router MAY support differentiated services [RFC2474].

5.6. Unicast Data Forwarding

The null route introduced by the WPD-6 requirement in [RFC6204] has lower precedence than other routes except for the default route.

5.7. Additional DHCPv6 WAN Requirement

When the WAN interface sends a DHCPV6 SOLICIT message, the CE router SHOULD request all mandatory information (IA_NA and IA_PD options) in the SOLICIT regardless of whether any partial information was received in response to previous SOLICITs.

6. Security Considerations

None.

7. Acknowledgements

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

Mikael Abrahamsson, Merete Asak, Scott Beuker, Mohamed Boucadair, Rex Bullinger, Brian Carpenter, Remi Denis-Courmont, Gert Doering, Alain Durand, Katsunori Fukuoka, Tony Hain, Thomas Herbst, Kevin Johns, Stephen Kramer, Victor Kuarsingh, Francois-Xavier Le Bail, Chad Mikkelsen, David Miles, Shin Miyakawa, Jean-Francois Mule, Michael Newbery, Carlos Pignataro, John Pomeroy, Antonio Querubin, Teemu Savolainen, Matt Schmitt, Hiroki Sato, Mark Townsley, Bernie Volz, James Woodyatt, Dan Wing and Cor Zwart

This draft 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.

8. 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.

9. IANA Considerations

This memo includes no request to IANA.

10. References

10.1. Normative References

[I-D.ietf-softwire-ds-lite-tunnel-option]

Hankins, D. and T. Mrugalski, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Option for Dual- Stack Lite", draft-ietf-softwire-ds-lite-tunnel-option-10 (work in progress), March 2011.

[I-D.ietf-softwire-dual-stack-lite]

Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual- Stack Lite Broadband Deployments Following IPv4 Exhaustion", draft-ietf-softwire-dual-stack-lite-11 (work in progress), May 2011.

[I-D.vyncke-advanced-ipv6-security]

Vyncke, E. and M. Townsley, "Advanced Security for IPv6 CPE", draft-vyncke-advanced-ipv6-security-01 (work in progress), March 2010.

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

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

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

January 1997.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2464] Crawford, M., "Transmission of IPv6 Packets over Ethernet Networks", RFC 2464, December 1998.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, 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.
- [RFC3736] Droms, R., "Stateless Dynamic Host Configuration Protocol (DHCP) Service for IPv6", RFC 3736, April 2004.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC4075] Kalusivalingam, V., "Simple Network Time Protocol (SNTP) Configuration Option for DHCPv6", RFC 4075, May 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.
- [RFC4294] Loughney, J., "IPv6 Node Requirements", RFC 4294, April 2006.

- [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.
- [RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 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.
- [RFC5571] Storer, B., Pignataro, C., Dos Santos, M., Stevant, B., Toutain, L., and J. Tremblay, "Softwire Hub and Spoke Deployment Framework with Layer Two Tunneling Protocol Version 2 (L2TPv2)", RFC 5571, June 2009.
- [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.

- [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.

10.2. Informative References

- [I-D.ietf-behave-v6v4-framework]
Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation",
draft-ietf-behave-v6v4-framework-10 (work in progress),
August 2010.
- [UPnP-IGD]
UPnP Forum, "Universal Plug and Play (UPnP) Internet Gateway Device (IGD)", November 2001,
<<http://www.upnp.org/standardizeddcps/igd.asp>>.

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
ATT
725 W Peachtree St
Atlanta, GA 30308
USA

Email: barbara.stark@att.com

Ole Troan (editor)
Cisco Systems, Inc.
Veversmauet 8
N-5017 BERGEN,
Norway

Email: ot@cisco.com

IPv6 Operations
Internet-Draft
Intended status: Informational
Expires: December 10, 2011

J. Livingood
Comcast
June 8, 2011

IPv6 AAAA DNS Whitelisting Implications
draft-ietf-v6ops-v6-aaaa-whitelisting-implications-06

Abstract

This document describes the practice and implications of whitelisting DNS recursive resolvers in order to limit AAAA resource record responses (which contain IPv6 addresses) sent by authoritative DNS servers. This is an IPv6 transition mechanism used by domains as a method for incrementally transitioning inbound traffic to a domain from IPv4 to IPv6 transport. The audience for this document is the Internet community generally, particularly IPv6 implementers.

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 December 10, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. How DNS Whitelisting Works	5
2.1. Description of the Operation of DNS Whitelisting	6
2.2. Comparison with Blacklisting	9
3. Similarities to Other DNS Operations	9
3.1. Similarities to Split DNS	9
3.2. Similarities to DNS Load Balancing	10
4. What Problems Are Implementers Trying To Solve?	10
4.1. Volume-Based Concerns	11
4.2. IPv6-Related Impairment	11
4.3. Free Versus Subscription Services	13
5. General Implementation Variations	13
5.1. Implement DNS Whitelisting Universally	13
5.2. Implement DNS Whitelisting On An Ad Hoc Basis	14
5.3. Do Not Implement DNS Whitelisting	14
5.3.1. Solve Current End User IPv6 Impairments	14
5.3.2. Gain Experience Using IPv6 Transition Names	15
5.3.3. Implement DNS Blacklisting	15
6. Concerns Regarding DNS Whitelisting	16
7. Implications of DNS Whitelisting	17
7.1. Architectural Implications	17
7.2. Public IPv6 Address Reachability Implications	18
7.3. Operational Implications	19
7.3.1. De-Whitelisting May Occur	19
7.3.2. Authoritative DNS Server Operational Implications	19
7.3.3. DNS Recursive Resolver Server Operational Implications	20
7.3.4. Monitoring Implications	21
7.3.5. Implications of Operational Momentum	22
7.3.6. Troubleshooting Implications	22
7.3.7. Additional Implications If Deployed On An Ad Hoc Basis	22
7.4. Homogeneity May Be Encouraged	23
7.5. Technology Policy Implications	23
7.6. IPv6 Adoption Implications	24
7.7. Implications with Poor IPv4 and Good IPv6 Transport	25
7.8. Implications for Users of Third-Party DNS Recursive Resolvers	25
8. Is DNS Whitelisting a Recommended Practice?	26
9. Security Considerations	26
9.1. DNSSEC Considerations	27
9.2. Authoritative DNS Response Consistency Considerations	27

10. Privacy Considerations	28
11. IANA Considerations	28
12. Contributors	28
13. Acknowledgements	29
14. References	30
14.1. Normative References	30
14.2. Informative References	31
Appendix A. Document Change Log	33
Appendix B. Open Issues	36
Author's Address	36

1. Introduction

This document describes the practice and implications of whitelisting DNS recursive resolvers in order to limit AAAA resource record (RR) responses (which contain IPv6 addresses) sent by authoritative DNS servers. This is referred to hereafter as DNS Whitelisting. This is an IPv6 transition mechanism used by domains as a method for incrementally transitioning inbound traffic to a domain from IPv4 to IPv6 transport. When implemented, a domain's authoritative DNS will return a AAAA resource record to DNS recursive resolvers [RFC1035] on the whitelist, while returning no AAAA resource records to DNS recursive resolvers which are not on the whitelist. The practice appears to have first been used by major web content sites (sometimes described hereafter as "high-traffic domains"), which have specific concerns relating to maintaining a high-quality user experience for all of their users during their transition to IPv6.

Critics of the practice of DNS Whitelisting have articulated several concerns. Among these are that:

- o DNS Whitelisting is a very different behavior from the current practice concerning the publishing of IPv4 address resource records,
- o that it may create a two-tiered Internet,
- o that policies and decision-making for whitelisting and de-whitelisting are opaque or likely to cause conflict,
- o that DNS Whitelisting reduces interest in the deployment of IPv6,
- o that new operational and management burdens are created,
- o that the practice does not scale,
- o that it violates a basic premise of cross-Internet interoperability by requiring prior arrangements,
- o and that the costs and negative implications of DNS Whitelisting outweigh the perceived benefits.

This document explores the reasons and motivations for DNS Whitelisting Section 4. It also explores the concerns regarding this practice, and whether and when the practice is recommended Section 8. Readers will hopefully better understand what DNS Whitelisting is, why some domains are implementing it, and what the implications are.

2. How DNS Whitelisting Works

Generally, using a whitelist means no traffic (or traffic of a certain type) is permitted to the destination host unless the originating host's IP address is contained in the whitelist. In contrast, using a blacklist means that all traffic is permitted to the destination host unless the originating host's IP address is contained in the blacklist.

DNS Whitelisting is implemented in authoritative DNS servers, not in DNS recursive resolvers. These authoritative DNS servers implement IP address-based restrictions on AAAA query responses. So far, DNS Whitelisting has been primarily implemented by web site operators deploying IPv6-enabled services, though this practice could affect all protocols and services within a domain. For a given operator of a website, such as `www.example.com`, the domain operator essentially applies an access control list (ACL) on the authoritative DNS servers for the domain `example.com`. The ACL is populated with the IPv4 and/or IPv6 addresses or prefix ranges of DNS recursive resolvers on the Internet, which have been authorized to receive (or access) AAAA resource record responses. These DNS recursive resolvers are operated by third parties, such as Internet Service Providers (ISPs), universities, governments, businesses, and individual end users. If a DNS recursive resolver IS NOT matched in the ACL, then AAAA resource records WILL NOT be sent in response to a query for a hostname in the `example.com` domain. However, if a DNS recursive resolver IS matched in the ACL, then AAAA resource records WILL be sent in response to a query for a given hostname in the `example.com` domain. While these are not network-layer access controls (as many ACLs are) they are nonetheless access controls that are a factor for end users and other organizations such as network operators, especially as networks and hosts transition from one network address family to another (IPv4 to IPv6). Thus, if a DNS recursive resolver is on the ACL (whitelist) then they have access to AAAA resource records for the domain.

In practice, DNS Whitelisting generally means that a very small fraction of the DNS recursive resolvers on the Internet (those in the whitelist or ACL) will receive AAAA responses. The large majority of DNS recursive resolvers on the Internet will therefore receive only A resource records containing IPv4 addresses. Thus, quite simply, the authoritative server hands out different answers depending upon who is asking; with IPv4 and IPv6 resource records for all those the authorized whitelist, and only IPv4 resource records for everyone else. See Section 2.1 and Figure 1 for more details.

DNS Whitelisting also works independently of whether an authoritative DNS server, DNS recursive resolver, or end user host uses IPv4

transport, IPv6, or both. So, for example, whitelisting may prevent sending AAAA responses even in those cases where the DNS recursive resolver has queried the authoritative server over IPv6 transport, or where the end user host's original query to the DNS recursive resolver was over IPv6 transport. One important reason for this is that even though the DNS recursive resolver may have no IPv6-related impairments, this is not a reliable predictor of whether the same is true of the end user host. This also means that a DNS whitelist can contain both IPv4 and IPv6 addresses.

Finally, DNS Whitelisting could possibly be deployed in two ways: universally on a global basis (though that would be considered harmful and is just covered to explain why this is the case), or, more realistically, on an ad hoc basis. Deployment on a universal deployment basis means that DNS Whitelisting is implemented on all authoritative DNS servers, across the entire Internet. In contrast, deployment on an ad hoc basis means that only some authoritative DNS servers, and perhaps even only a few, implement DNS Whitelisting. These two potential deployment models are described in Section 5.

Specific implementations will vary from domain to domain, based on a range of factors such as the technical capabilities of a given domain. As such, any examples listed herein should be considered general examples and are not intended to be exhaustive.

2.1. Description of the Operation of DNS Whitelisting

The system logic of DNS Whitelisting is as follows:

1. The authoritative DNS server for example.com receives DNS queries for the A (IPv4) and/or AAAA (IPv6) address resource records for the Fully Qualified Domain Name (FQDN) www.example.com, for which AAAA (IPv6) resource records exist.
2. The authoritative DNS server checks the IP address (IPv4, IPv6, or both) of the DNS recursive resolver sending the AAAA (IPv6) query against the access control list (ACL) that is the DNS Whitelist.
3. If the DNS recursive resolver's IP address IS matched in the ACL, then the response to that specific DNS recursive resolver can contain AAAA (IPv6) address resource records.
4. If the DNS recursive resolver's IP address IS NOT matched in the ACL, then the response to that specific DNS recursive resolver cannot contain AAAA (IPv6) address resource records. In this case, the server will likely return a response with the response code (RCODE) being set to 0 (No Error) with an empty answer

section for the AAAA record query.

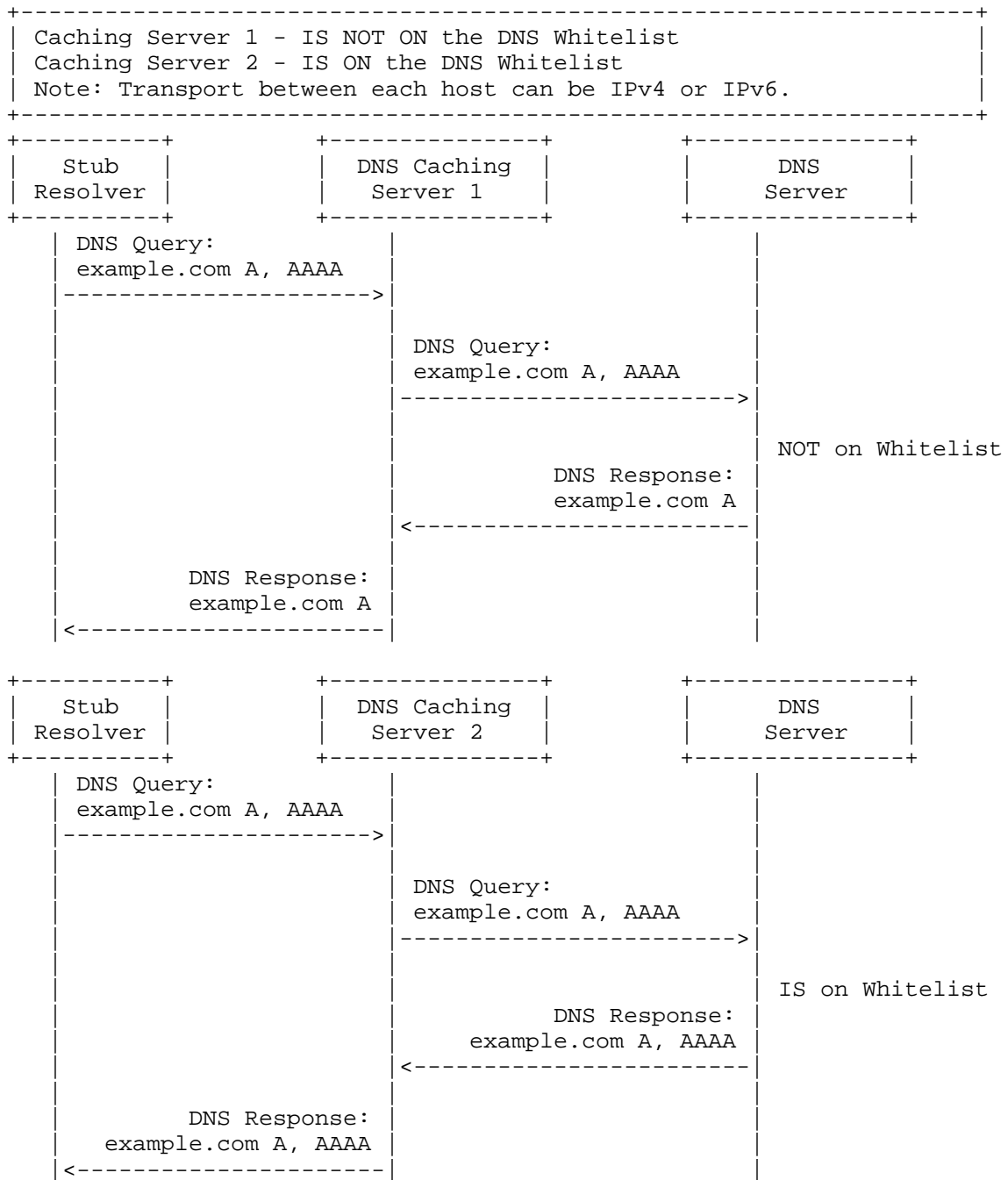


Figure 1: DNS Whitelisting Diagram

2.2. Comparison with Blacklisting

With DNS Whitelisting, DNS recursive resolvers can receive AAAA resource records only if they are on the whitelist. In contrast, blacklisting would be the opposite whereby all DNS recursive resolvers can receive AAAA resource records unless they are on the blacklist. So a whitelist contains a list of hosts allowed something, whereby a blacklist contains a list of hosts disallowed something. While the distinction between the concepts of whitelisting and blacklisting is important, this is noted specifically since some implementers of DNS Whitelisting may choose to transition to DNS Blacklisting before returning to a state without address-family-related ACLs in their authoritative DNS servers. It is unclear when and if it would be appropriate to change from whitelisting to blacklisting. Nor is it clear how implementers will judge the network conditions to have changed sufficiently to justify disabling such controls.

3. Similarities to Other DNS Operations

Some aspects of DNS Whitelisting may be considered similar to other common DNS operational techniques which are explored below.

3.1. Similarities to Split DNS

DNS Whitelisting has some similarities to so-called split DNS, briefly described in Section 3.8 of [RFC2775]. When split DNS is used, the authoritative DNS server returns different responses depending upon what host has sent the query. While [RFC2775] notes the typical use of split DNS is to provide one answer to hosts on an Intranet and a different answer to hosts on the Internet, the essence is that different answers are provided to hosts on different networks. This is basically the way that DNS Whitelisting works, whereby hosts on different networks which use different DNS recursive resolvers, receive different answers if one DNS recursive resolver is on the whitelist and the other is not.

In [RFC2956], Internet transparency and Internet fragmentation concerns regarding split DNS are detailed in Section 2.1. [RFC2956] further notes in Section 2.7, concerns regarding split DNS and that it "makes the use of Fully Qualified Domain Names (FQDNs) as endpoint identifiers more complex." Section 3.5 of [RFC2956] further recommends that maintaining a stable approach to DNS operations is key during transitions such as the one to IPv6 that is underway now, stating that "Operational stability of DNS is paramount, especially during a transition of the network layer, and both IPv6 and some network address translation techniques place a heavier burden on

DNS."

3.2. Similarities to DNS Load Balancing

DNS Whitelisting also has some similarities to DNS load balancing. There are of course many ways that DNS load balancing can be performed. In one example, multiple IP address resource records (A and/or AAAA) can be added to the DNS for a given FQDN. This approach is referred to as DNS round robin [RFC1794]. DNS round robin may also be employed where SRV resource records are used [RFC2782].

In another example, one or more of the IP address resource records in the DNS will direct traffic to a load balancer. That load balancer, in turn, may be application-aware, and pass the traffic on to one or more hosts connected to the load balancer which have different IP addresses. In cases where private IPv4 addresses are used [RFC1918], as well as when public IP addresses are used, those end hosts may not necessarily be directly reachable without passing through the load balancer first.

Additionally, a geographically-aware authoritative DNS server may be used, as is common with Content Delivery Networks (CDNs) or Global Load Balancing (GLB, also referred to as Global Server Load Balancing, or GSLB), whereby the IP address resource records returned to a resolver in response to a query will vary based on the estimated geographic location of the resolver [Wild-Resolvers]. CDNs perform this function in order to attempt to direct hosts to connect to the nearest content cache. As a result, one can see some similarities with DNS Whitelisting insofar as different IP address resource records are selectively returned to resolvers based on the IP address of each resolver (or other imputed factors related to that IP address). However, what is different is that in this case the resolvers are not deliberately blocked from receiving DNS responses containing an entire class of addresses; this load balancing function strives to perform a content location-improvement function and not an access control function.

4. What Problems Are Implementers Trying To Solve?

Implementers are attempting to protect users of their domain from having a negative experience (poor performance) when they receive DNS response containing AAAA resource records or when attempting to use IPv6 transport. There are two concerns which relate to this practice; one of which relates to IPv6-related impairment and the other which relates to the maturity or stability of IPv6 transport for high-traffic domains. Both can negatively affect the experience of end users.

Not all domains may face these challenges, though some clearly do, since the user base of each domain, traffic sources, traffic volumes, and other factors obviously varies between domains. For example, while some domains have implemented DNS Whitelisting, others have run IPv6 experiments whereby they added AAAA resource records and observed and measured errors, and then decided not to implement DNS Whitelisting [Heise]. A more widespread such experiment was World IPv6 Day [W6D], sponsored by the Internet Society, on June 8, 2011. This was a unique opportunity for hundreds of domains to add AAAA resource records to the DNS without using DNS Whitelisting, all at the same time. Domains can run their own independent experiments in the future, adding AAAA resource records for a period of time, and then analyzing any impacts or effects on traffic and the experience of end users.

4.1. Volume-Based Concerns

Some implementers are trying to gradually add IPv6 traffic to their domain since they may find that network operations, tools, processes and procedures are less mature for IPv6 as compared to IPv4. Compared to domains with small to moderate traffic volumes, whether by the count of end users or count of bytes transferred, high-traffic domains receive such a level of usage that it is prudent to undertake any network changes gradually or in a manner which minimizes any risk of disruption.

For example, one can imagine for one of the top ten sites globally that the idea of suddenly turning on a significant amount of IPv6 traffic is quite daunting. DNS Whitelisting may therefore offer such high-traffic domains one potential method for incrementally enabling IPv6. Thus, some implementers with high-traffic domains plan to use DNS Whitelisting as a necessary, though temporary, risk reduction tactic intended to ease their transition to IPv6 and minimize any perceived risk in such a transition.

4.2. IPv6-Related Impairment

Some implementers have observed that when they added AAAA resource records to their authoritative DNS servers in order to support IPv6 access to their content that a small fraction of end users had slow or otherwise impaired access to a given web site with both AAAA and A resource records. The fraction of users with such impaired access has been estimated to be as high as 0.078% of total Internet users [IETF-77-DNSOP] [NW-Article-DNSOP] [IPv6-Growth] [IPv6-Brokenness], though more recent measurements indicate this is declining [Impairment-Tracker]. In these situations, DNS recursive resolvers are added to the DNS Whitelist only when the measured level of impairment of the hosts using that resolver declines to some level

acceptable by the domain.

It is not clear if the level of IPv4-related impairment is more or less than IPv6-related impairment. As one document reviewer has pointed out, it may simply be that websites are only measuring IPv6 impairments and not IPv4 impairments, whether because IPv6 is new or whether those websites are simply unable to or are otherwise not in a position to be able to measure IPv4 impairment (since this could result in no Internet access whatsoever).

As a result of this impairment affecting end users of a given domain, a few high-traffic domains have either implemented DNS Whitelisting or are considering doing so [NW-Article-DNS-WL] [WL-Ops]. While it is outside the scope of this document to explore the various reasons why a particular user's system (host) may have impaired IPv6 access, for the users who experience this impairment it has a very real performance impact. It would affect access to all or most dual stack services to which the user attempts to connect. This negative end user experience can range from somewhat slower than usual access (as compared to native IPv4-based access), to extremely slow access, to no access to the domain whatsoever. In essence, whether the end user even has an IPv6 address or not, merely by receiving a AAAA record response the user either cannot access a FQDN or it is so slow that the user gives up and assumes the destination is unreachable.

In addition, at least one high-traffic domain has noted that they have received requests to not send DNS responses with AAAA resource records to particular DNS recursive resolvers. In this case, a DNS recursive resolver operator expressed a short-term concern that their IPv6 network infrastructure was not yet ready to handle the large traffic volume that may be associated with the hosts in their network connecting to the websites of these domains. These end user networks may also have other tools at their disposal in order to address this concern, including applying rules to network equipment such as routers and firewalls (this will necessarily vary by the type of network, as well as the technologies used and the design of a given network), as well as configuration of their DNS recursive resolvers (though modifying or suppressing AAAA resource records in a DNSSEC-signed domain on a Security-Aware Resolver will be problematic Section 9.1).

It is worth noting that the IP address of a DNS recursive resolver is not a precise indicator of the IPv6 preparedness, or lack of IPv6-related impairment, of end user hosts which query (use) a particular DNS recursive resolver. While the DNS recursive resolver may be an imperfect proxy for judging IPv6 preparedness, it is at least one of the best available methods at the current time.

4.3. Free Versus Subscription Services

It is also worth noting the differences between domains containing primarily subscription-based services compared to those containing primarily free services. In the case of free services, such as search engines, end users have no direct billing relationship with the domain and can switch sites simply by changing the address they enter into their browser (ignoring other value added services which may tie a user's preference to a given domain or otherwise create switching costs). As a result, such domains may be more sensitive to IPv6 transition issues since their users can quickly switch to another domain that is not using IPv6.

5. General Implementation Variations

In considering how DNS Whitelisting may emerge more widely, there are two deployment scenarios explored below, one of which, the ad-hoc case Section 5.2, is realistic and is happening now. The other, universal deployment Section 5.1, is only described for the sake of completeness, to highlight its difficulties, and to explain why it would be considered harmful. Other possible alternative or supplementary approaches are also outlined.

In evaluating implementing DNS Whitelisting universally and on an ad hoc basis, it is possible that reputable third parties could create and maintain DNS whitelists, in much the same way that blacklists are distributed and used for reducing email spam. In the email context, a mail operator subscribes to one or more of these lists and as such the operational processes for additions and deletions to the list are managed by a third party. A similar model could emerge for DNS Whitelisting.

In either of those scenarios a DNS recursive resolver operator will have to determine whether or not DNS Whitelisting has been implemented for a domain, since the absence of AAAA resource records may simply be indicative that the domain has not yet added IPv6 addressing for the domain, rather than that they have done so but are using DNS Whitelisting. This will be challenging at scale.

5.1. Implement DNS Whitelisting Universally

One approach is to implement DNS Whitelisting universally, which could also involve using some sort of centralized registry of DNS Whitelisting policies, contracts, processes, or other information. For this deployment scenario to occur, DNS Whitelisting functionality would need to be built into all authoritative DNS server software, and all operators of authoritative DNS servers would have to upgrade

their software in order to enable this functionality. New IETF Request for Comment (RFC) documents may need to be completed to describe how to properly configure, deploy, and maintain DNS Whitelisting across the entire Internet. As a result, it is highly unlikely that DNS Whitelisting will become universally deployed.

Such an approach is considered harmful and problematic, and almost certain not to happen.

5.2. Implement DNS Whitelisting On An Ad Hoc Basis

DNS Whitelisting is now being adopted on an ad hoc, or domain-by-domain basis. Therefore, only those domains interested in DNS Whitelisting would need to adopt the practice. Also in this scenario, ad hoc use by a particular domain is likely to be a temporary measure that has been adopted to ease the transition of the domain to IPv6. A domain, particularly a high-traffic domain, may choose to do so in order to ease their transition to IPv6 through a selective deployment so as to minimize any risks or disruptions in such a transition.

One benefit of DNS Whitelisting being deployed on an ad hoc basis is that only the domains that are interested in doing so would have to upgrade their authoritative DNS servers (or take other steps) in order to implement DNS Whitelisting. Some domains that plan to or already have implemented this and are manually updating their whitelist, while others such as CDNs have discussed the possibility of an automated method for doing so.

5.3. Do Not Implement DNS Whitelisting

As an alternative to adopting DNS Whitelisting, domains can choose not to implement DNS Whitelisting, continuing the current predominant authoritative DNS operational model on the Internet. It is then up to end users with IPv6-related impairments to discover and fix those impairments, though clearly other parties including end user host operating system developers can play a critical role. However, the concerns and risks related to traffic volume Section 4.1 should still be considered since those are not directly related to such impairments.

5.3.1. Solve Current End User IPv6 Impairments

A further extension of not implementing DNS Whitelisting, is to also endeavor fix the underlying technical problems experienced by end users during the transition to IPv6. A first step is to identify which users have such impairments and then to communicate this information to affected users. Such end user communication is likely

to be most helpful if the end user is not only alerted to a potential problem but is given careful and detailed advice on how to resolve this on their own, or where they can seek help in doing so. Section 10 may also be relevant in this case.

One challenge with this option is the potential difficulty of motivating members of the Internet community to work collectively towards this goal, sharing the labor, time, and costs related to such an effort. However, World IPv6 Day [W6D] shows that such community efforts are possible and despite any potential challenges, the Internet community continues to work to solve end user IPv6 impairments.

However, as noted above, the concerns and risks related to traffic volume Section 4.1 should still be considered since those are not directly related to such impairments.

5.3.2. Gain Experience Using IPv6 Transition Names

Another alternative is for domains to gain experience using an FQDN which has become common for domains beginning the transition to IPv6; `ipv6.example.com` and `www.ipv6.example.com`. This can be a way for a domain to gain IPv6 experience and increase IPv6 use on a relatively controlled basis, and to inform any plans for DNS Whitelisting with experience.

While this is a good first step to functionally test and prepare a domain for IPv6, the utility of the tactic is limited since users must know the transition name, the traffic volume will be low, and the traffic is unlikely to be representative of the general population of end users, among other reasons. Thus, as noted above, the concerns and risks related to traffic volume Section 4.1 should still be considered.

5.3.3. Implement DNS Blacklisting

Some domains may wish to be more permissive than if they adopted DNS Whitelisting, but still have some level of control over returning AAAA record responses. In this case an alternative may be to employ DNS Blacklisting, which would enable all DNS recursive resolvers to receive AAAA record responses except for the relatively small number that are listed in the blacklist. This could, for example, enable an implementer to only prevent such responses where there has been a relatively high level of IPv6-related impairments, until such time as these impairments can be fixed or otherwise meaningfully reduced to an acceptable level.

This approach is likely to be significantly less labor intensive for

an authoritative DNS server operator, as they would presumably focus on a smaller number of DNS recursive resolvers than if they implemented whitelisting. Thus, these authoritative DNS server operators would only need to communicate with a few DNS recursive resolver operators rather than potentially all such operators. This should result in lower labor, systems, and process requirements. This is not to say that there will be no time required to work with those parties affected by a blacklist, simply that there are likely to be fewer such interactions and that each such interaction could be shorter in duration.

The email industry has a long experience with blacklists and, very generally speaking, blacklists tend to be effective and well received when it is easy to discover if a server is on a blacklist, if there is a transparent and easily understood process for requesting removal from a blacklist, and if the decision-making criteria for placing a server on a blacklist is transparently disclosed and perceived as fair.

As noted in Section 7.3.7, it is also possible that a domain may choose to first implement DNS Whitelisting and then migrate to DNS Blacklisting.

6. Concerns Regarding DNS Whitelisting

There is concern that the practice of DNS Whitelisting for IPv6 address resource records represents a departure from the generally accepted practices regarding IPv4 address resource records in the DNS on the Internet [WL-Concerns]. Generally, once an authoritative server operator adds an A record (IPv4) to the DNS, then any DNS recursive resolver on the Internet can receive that A record in response to a query. This enables new server hosts that are connected to the Internet, and for which a FQDN such as `www.example.com` has been added to the DNS with an IPv4 address record, to be almost immediately reachable by any host on the Internet. Each end in this end-to-end model is responsible for connecting to the Internet and once they have done so they can connect to each other without additional impediments, middle networks, intervening networks, or servers either knowing about all end points or whether one is allowed to discover and contact the other. The end result is that new server hosts become more and more widely accessible as new networks and new hosts connect to the Internet over time, capitalizing on and increasing so-called "network effects" (also called network externalities).

In contrast, DNS Whitelisting may fundamentally change this model. In the altered DNS Whitelisting end-to-end model, one end (where the

end user is located) cannot readily discover the other end (where the content is located), without parts of the middle (authoritative DNS servers) making a new type of access control decision in the DNS. So in the current IPv4-based Internet when a new server host is added to the Internet it is generally widely available to all end user hosts via a FQDN. When DNS Whitelisting of IPv6 resource records is used, these new server hosts are not accessible via a FQDN by any end user hosts until such time as the operator of the authoritative DNS servers adds DNS recursive resolvers around the Internet to the DNS Whitelist.

7. Implications of DNS Whitelisting

The key DNS Whitelisting implications are detailed below.

7.1. Architectural Implications

DNS Whitelisting modifies the end-to-end model and the general notion of spontaneous interoperability of the architecture that prevails on the Internet today. This is because this approach moves additional access control information and policies into the middle of the DNS resolution path of the IPv6-addressed Internet, which generally did not exist before on the IPv4-addressed Internet, and it requires some type of prior registration with authoritative servers. This poses some risks noted in [RFC3724]. In explaining the history of the end-to-end principle, [RFC1958] states that one of the goals is to minimize the state, policies, and other functions needed in the middle of the network in order to enable end-to-end communications on the Internet. In this case, the middle network should be understood to mean anything other than the end hosts involved in communicating with one another. Some state, policies, and other functions have always been necessary to enable such end-to-end communication, but the goal of the approach has been to minimize this to the greatest extent possible.

It is also possible that DNS Whitelisting could place at risk some of the observed benefits of the end-to-end principle, as listed in Section 4.1 of [RFC3724], such as protection of innovation. [RFC3234] details issues and concerns regarding so-called middleboxes, so there may also be parallel concerns with the DNS Whitelisting approach, especially concerning modified DNS servers noted in Section 2.16 of [RFC3234], as well as more general concerns noted in Section 1.2 of [RFC3234] about the introduction of new failure modes. In particular, there may be concerns that configuration is no longer limited to two ends of a session, and that diagnosis of failures and misconfigurations becomes more complex.

Two additional sources worth considering as far as implications for the end-to-end model are concerned are [Tussle] and [Rethinking]. In [Tussle], the authors note concerns regarding the introduction of new control points, as well as "kludges" to the DNS, as risks to the goal of network transparency in the end-to-end model. Given the emerging use of DNS Whitelisting [Tussle] is an interesting and relevant document. In addition, [Rethinking] reviews similar issues that are of interest to readers of this document.

Also, it is somewhat possible that DNS Whitelisting could affect some of the architectural assumptions which underlie parts of Section 2 of [RFC4213] which outlines the dual stack approach to the IPv6 transition. DNS Whitelisting could modify the behavior of the DNS, as described in Section 2.2 of [RFC4213] and could require different sets of DNS servers to be used for hosts that are (using terms from that document) IPv6/IPv4 nodes, IPv4-only nodes, and IPv6-only nodes. As such, broad use of DNS Whitelisting may necessitate the review and/or revision (though revision is unlikely to be necessary) of standards documents which describe dual-stack and IPv6 operating modes, dual-stack architecture generally, and IPv6 transition methods, including but not limited to [RFC4213].

7.2. Public IPv6 Address Reachability Implications

It is critical to understand that the concept of reachability described here depends upon a knowledge of an address in the DNS. Thus, in order to establish reachability to an end point, a host is dependent upon looking up an IP address in the DNS when a FQDN is used. When DNS Whitelisting is used, it is quite likely that an IPv6-enabled end user host could connect to an example server host using the IPv6 address, even though the FQDN associated with that server host is restricted via a DNS whitelist. Since most Internet applications and hosts such as web servers depend upon the DNS, and as end users connect to FQDNs such as `www.example.com` and do not remember or wish to type in an IP address, the notion of reachability described here should be understood to include knowledge of how to associate a name with a network address.

The predominant experience of end user hosts and servers on the IPv4-addressed Internet today is that when a new server with a public IPv4 address is added to the DNS, that a FQDN is immediately useful for reaching it. This is a generalization and in Section 3 there are examples of common cases where this may not necessarily be the case. For the purposes of this argument, that concept of accessibility is described as "pervasive reachability". It has so far been assumed that the same expectations of pervasive reachability would exist in the IPv6-addressed Internet. However, if DNS Whitelisting is deployed, this will not be the case since only end user hosts using

DNS recursive resolvers that are included in the ACL of a given domain using DNS Whitelisting would be able to reach new servers in that given domain via IPv6 addresses. The expectation of any end user host being able to connect to any server (essentially both hosts, just at either end of the network), defined here as "pervasive reachability", will change to "restricted reachability" with IPv6.

Establishing DNS Whitelisting as an accepted practice in the early phases of mass IPv6 deployment could establish it as an integral part of how IPv6 DNS resource records are deployed globally. This risks DNS Whitelisting living on for many years as a key foundational element of domain name management on the Internet.

7.3. Operational Implications

This section explores some of the operational implications which may occur as a result of, are related to, or become necessary when engaging in the practice of DNS Whitelisting.

7.3.1. De-Whitelisting May Occur

It is possible for a DNS recursive resolver added to a whitelist to then be removed from the whitelist, also known as de-whitelisting. Since de-whitelisting can occur, through a decision by the authoritative server operator, the domain owner, or even due to a technical error, an operator of a DNS recursive resolver will have new operational and monitoring requirements and/or needs as noted in Section 7.3.3, Section 7.3.4, Section 7.3.6, and Section 7.5. One particular risk is that, especially when a high-traffic domain de-whitelists a large network, this may cause a sudden and dramatic change to networks since a large volume of traffic will then switch from IPv6 to IPv4. This can have dramatic effects on those being de-whitelisted as well as on other interconnected networks. In some cases, IPv4 network links may rapidly become congested and users of affected networks will experience network access impairments well beyond the domain which performed the de-whitelisting. Thus, once "operational stability" has been achieved between a whitelisting and whitelisted party, then de-whitelisting should generally not occur except in cases of operational emergencies, and there should be opportunities for joint troubleshooting or at least for advance warning to affected parties.

7.3.2. Authoritative DNS Server Operational Implications

DNS Whitelisting serves as a critical infrastructure service; to be useful it needs careful and extensive administration, monitoring and operation. Each new and essential mechanism creates substantial follow-on support costs.

Operators of authoritative servers (which are frequently authoritative for multiple domain names) will need to maintain an ACL on a server-wide basis affecting all domains, or on a domain-by-domain basis. As a result, operational practices and software capabilities may need to be developed in order to support such functionality. In addition, processes may need to be put in place to protect against inadvertently adding or removing IP addresses, as well as systems and/or processes to respond to such incidents if and when they occur. For example, a system may be needed to record DNS Whitelisting requests, report on their status along a workflow, add IP addresses when whitelisting has been approved, remove IP addresses when they have been de-whitelisted, log the personnel involved and timing of changes, schedule changes to occur in the future, and to roll back any inadvertent changes.

Operators may also need implement new forms of monitoring in order to apply change control, as noted briefly in Section 7.3.4.

It is important for operators of authoritative servers to recognize that the operational burden is likely to increase dramatically over time, as more and more networks transition to IPv6. As a result, the volume of new DNS Whitelisting requests will increase over time, potentially at an extraordinary growth rate, which will place an increasing burden on personnel, systems, and/or processes. Operators should also consider that any supporting systems, including the authoritative servers themselves, may experience reduced performance when a DNS whitelist becomes quite large.

7.3.3. DNS Recursive Resolver Server Operational Implications

For operators of DNS recursive resolvers, coping with DNS Whitelisting becomes expensive in time and personnel as the practice scales up. These operators include ISPs, enterprises, universities, governments; a wide range of organization types with a range of DNS-related expertise. They will need to implement new forms of monitoring, as noted briefly in Section 7.3.4. But more critically, such operators will need to add people, processes, and systems in order to manage large numbers of DNS Whitelisting applications. Since there is no common method for determining whether or not a domain is engaged in DNS Whitelisting, operators will have to apply to be whitelisted for a domain based upon one or more end user requests, which means systems, processes, and personnel for handling and responding to those requests will also be necessary.

When operators apply for DNS Whitelisting for all domains, that may mean doing so for all registered domains. Thus, some system would have to be developed to discover whether each domain has been whitelisted or not, which is touched on in Section 5 and may vary

depending upon whether DNS Whitelisting is universally deployed or is deployed on an ad hoc basis.

These operators (of DNS recursive resolvers) will need to develop processes and systems to track the status of all DNS Whitelisting applications, respond to requests for additional information related to these applications, determine when and if applications have been denied, manage appeals, and track any de-whitelisting actions.

Given the large number of domains in existence, the ease with which a new domain can be added, and the continued strong growth in the numbers of new domains, readers should not underestimate the potential significance in personnel and expense that this could represent for such operators. In addition, it is likely that systems and personnel may also be needed to handle new end user requests for domains for which to apply for DNS Whitelisting, and/or inquiries into the status of a whitelisting application, reports of de-whitelisting incidents, general inquiries related to DNS Whitelisting, and requests for DNS Whitelisting-related troubleshooting by these end users.

7.3.4. Monitoring Implications

Once a DNS recursive resolver has been whitelisted for a particular domain, then the operator of that DNS recursive resolver may need to implement monitoring in order to detect the possible loss of DNS Whitelisting in the future. This DNS recursive resolver operator could configure a monitor to check for a AAAA response in the whitelisted domain, as a check to validate continued status on the DNS whitelist. The monitor could then trigger an alert if at some point the AAAA responses were no longer received, so that operations personnel could begin troubleshooting, as outlined in Section 7.3.6.

Also, authoritative DNS server operators are likely to need to implement new forms of monitoring. In this case, they may desire to monitor for significant changes in the size of the whitelist within a certain period of time, which might be indicative of a technical error such as the entire ACL being removed. Authoritative DNS server operators may also wish to monitor their workflow process for reviewing and acting upon DNS Whitelisting applications and appeals, potentially measuring and reporting on service level commitments regarding the time an application or appeal can remain at each step of the process, regardless of whether or not such information is shared with parties other than that authoritative DNS server operator.

7.3.5. Implications of Operational Momentum

It seems plausible that once DNS Whitelisting is implemented it will be very difficult to deprecate such technical and operational practices. This assumption is based on an understanding of human nature, not to mention physics. For example, as Sir Isaac Newton noted, "Every object in a state of uniform motion tends to remain in that state of motion unless an external force is applied to it" [Motion]. Thus, once DNS Whitelisting is implemented it is quite likely that it would take considerable effort to deprecate the practice and remove it everywhere on the Internet; it may otherwise simply remain in place in perpetuity. To illustrate this point, one could consider for example that there are many email servers continuing to attempt to query anti-spam DNS blocklists which have long ago ceased to exist.

7.3.6. Troubleshooting Implications

The implications of DNS whitelisted present many challenges, as detailed throughout Section 7. These challenges may negatively affect the end users' ability to troubleshoot, as well as that of DNS recursive resolver operators, ISPs, content providers, domain owners (where they may be different from the operator of the authoritative DNS server for their domain), and other third parties. This may make the process of determining why a server is not reachable via a FQDN significantly more complex and time-consuming.

7.3.7. Additional Implications If Deployed On An Ad Hoc Basis

As more domains choose to implement DNS Whitelisting, and more networks become IPv6-capable and request to be whitelisted, scaling up operational processes, monitoring, and ACL updates will become more difficult. The increased rate of change and increased size of whitelists will increase the likelihood of configuration and other operational errors.

It is unclear when and if it would be appropriate to change from whitelisting to blacklisting. It also seems unlikely for such a change from whitelisting to blacklisting to be coordinated across the Internet, so such a change to blacklisting will likely occur on an ad-hoc basis as well (if at all).

Finally, some implementers consider DNS Whitelisting to be a temporary measure. As such, it is not clear how these implementers will judge the network conditions to have changed sufficiently to justify disabling DNS Whitelisting (or blacklisting, or other AAAA resource record access controls) and/or what the process and timing will be in order to discontinue this practice.

7.4. Homogeneity May Be Encouraged

A broad trend on the Internet is a move toward more heterogeneity. One manifestation of this is in an increasing number, variety, and customization of end user hosts, including home networks, operating systems, client software, home network devices, and personal computing devices. This trend appears to have had a positive effect on the development and growth of the Internet and has enabled end users to connect any technically compliant device or use any technically compatible software to connect to the Internet. Not only does this trend towards greater heterogeneity reduce the control which is exerted in the middle of the network, described positively in [Tussle], [Rethinking], and [RFC3724], but it also appears to help to enable greater and more rapid innovation at the edge of the network.

Some forms of so-called "network neutrality" principles around the world include the notion that any IP-capable device should be able to connect to a network, encouraging heterogeneity. These principles are often explicitly encouraged by application providers, though some of these same providers may be using DNS Whitelisting. This is ironic, as one implication of the adoption of DNS Whitelisting is that it could encourage a move back towards homogeneity resulting from greater control over devices in order to attempt to enforce technical requirements intended to reduce IPv6-related impairments. This return to an environment of more homogenous and/or controlled end user hosts could have unintended side effects on and counter-productive implications for future innovation at the edge of the network.

7.5. Technology Policy Implications

A key technology policy implication concerns the policies and processes related to reviewing and making decisions on DNS Whitelisting applications for a domain, as well as making any possible de-whitelisting decisions. Important questions may include whether these policies have been fully and transparently disclosed, are non-discriminatory, and are not anti-competitive. Key questions here may include whether appeals are allowed, what the process is, what the expected turn around time is, and whether the appeal will be handled by an independent third party.

It is also conceivable that whitelisting and de-whitelisting decisions could be quite sensitive to concerned parties beyond the operator of the domain operator and the operator of the DNS recursive resolver, including end users, application developers, content providers, advertisers, public policy groups, governments, and other entities. These concerned parties may seek to become involved in or

express opinions concerning whitelisting and/or de-whitelisting decisions.

A final concern is that decisions relating to whitelisting and de-whitelisting may occur as an expression of other commercial, governmental, and/or cultural conflicts, given the new control point which has been established with DNS Whitelisting. For example, in one imagined scenario, a domain could withhold adding a network to their DNS Whitelisting unless that network agreed to some sort of financial payment, legal agreement, agreement to sever a relationship with a competitor of the domain, etc. In another example, a music-oriented domain may be engaged in some sort of dispute with an academic network concerning copyright infringement concerns within that network, and may choose to de-whitelist that network as a negotiating technique in some sort of commercial discussion. In a final example, a major email domain may choose to de-whitelist a network due to that network sending some large volume of spam. Thus, it seems possible that whitelisting and de-whitelisting could become a vehicle for adjudicating other disputes, and that this may well have consequences for end users which are affected by such decisions and are unable to express a strong voice in such decisions.

7.6. IPv6 Adoption Implications

As noted in Section 6, the implications of DNS Whitelisting may drive end users and/or networks to delay, postpone, or cancel adoption of IPv6, or to actively seek alternatives to it. Such alternatives may include the use of multi-layer or large scale network address translation (NAT) techniques, which these parties may decide to pursue on a long-term basis to avoid the perceived costs and aggravations related to DNS Whitelisting. This could of course come at the very time that the Internet community is trying to get these very same parties interested in IPv6 and motivated to begin the transition to IPv6. As a result, parties that are likely to be concerned over the negative implications of DNS Whitelisting could logically be concerned of the negative effects that this practice could have on the adoption of IPv6 if it became widespread.

At the same time, as noted in Section 4, some high-traffic domains may find the prospect of transitioning to IPv6 daunting without having some short-term ability to incrementally control the amount and source of IPv6 traffic to their domains. Lacking such controls, some domains may choose to substantially delay their transition to IPv6.

7.7. Implications with Poor IPv4 and Good IPv6 Transport

It is possible that there could be situations where the differing quality of the IPv4 and IPv6 connectivity of an end user could cause complications in accessing content which is in a whitelisted domain, when the end user's DNS recursive resolver is not on that whitelist. While today most end users' IPv4 connectivity is typically superior to IPv6 connectivity (if such connectivity exists at all), there could be implications when the reverse is true and an end user has markedly superior IPv6 connectivity as compared to IPv4. This is admittedly theoretical but could become a factor as the transition to IPv6 continues and IPv4 address availability within networks becomes strained.

For example, in one possible scenario, a user is issued IPv6 addresses by their ISP and has a home network and devices or operating systems which fully support IPv6. As a result this theoretical user has very good IPv6 connectivity. However, this end user's ISP may have exhausted their available pool of unique IPv4 addresses, and so that ISP uses NAT in order to reuse IPv4 addresses. So for IPv4 content, the end user must send their IPv4 traffic through some additional network element (e.g. NAT, proxy, tunnel server). Use of this additional network element may cause the end user to experience sub-optimal IPv4 connectivity when certain protocols or applications are used. This user then has good IPv6 connectivity but impaired IPv4 connectivity. Furthermore, this end user's DNS recursive resolver is not whitelisted by the authoritative server for a domain that the user is trying to access, meaning the end user only gets A record responses for their impaired IPv4 transport rather than also AAAA record responses for their stable and well-performing IPv6 transport. Thus, the user's poor IPv4 connectivity situation is potentially exacerbated by not having access to a given domain's IPv6 content since they must use the address family with relatively poor performance.

7.8. Implications for Users of Third-Party DNS Recursive Resolvers

In most cases it is assumed that end users will make use of DNS recursive resolvers which are operated by their access network provider, whether that is an ISP, campus network, enterprise network, or some other type of network. However there are also cases where an end user has changed their DNS server IP addresses in their device's operating system to those of a third party which operates DNS recursive resolvers independently of end user access networks.

In these cases, an authoritative DNS server may receive a query from a DNS recursive resolver in one network, though the end user sending the original query is in an entirely different network. It may

therefore be more challenging for a DNS Whitelist implementer to determine the level of IPv6-related impairment when such third-party DNS recursive resolvers are used, given the wide variety of end user access networks which may be used and that this mix may change in unpredictable ways over time.

There may also be cases where end users' assigned DNS recursive resolvers have not been whitelisted for a particular domain, but where the end user tries to switch to a third-party DNS recursive resolver that has been whitelisted. While in most cases the end user will be able to switch to use that third-party's DNS servers, some specialized access networks, such as in hotels and conference centers, may prevent using third-party DNS servers. In these cases, end users may be frustrated at their inability to access certain content over IPv6, resulting in complaints to both a particular domain as well as the access network operator.

8. Is DNS Whitelisting a Recommended Practice?

Opinions in the Internet community concerning whether or not DNS Whitelisting is a recommended practice are understandably quite varied. However, there is clear consensus that DNS Whitelisting can be a useful tactic a domain may choose to use as they transition to IPv6. In particular, some high-traffic domains view DNS Whitelisting as one of the few practical and low-risk approaches enabling them to transition to IPv6, without which their transition may not take place for some time. However, there is also consensus is that this practice is workable only in the short-term and that it will not scale over the long-term. Thus, some domains may find DNS Whitelisting a beneficial temporary tactic in their transition to IPv6. Such temporary use during the transition to IPv6 is broadly accepted within the community, so long as it does not become a long-term practice.

9. Security Considerations

If DNS Whitelisting is adopted, then organizations which apply DNS Whitelisting policies in their authoritative servers should have procedures and systems which do not allow unauthorized parties to either remove whitelisted DNS recursive resolvers from the whitelist or add non-whitelisted DNS recursive resolvers to the whitelist, just as all configuration settings for name servers should be protected by appropriate procedures and systems. Should such unauthorized additions or removals from the whitelist can be quite damaging, and result in content providers and/or ISPs to incur substantial support costs resulting from end user and/or customer contacts. As such,

great care must be taken to control access to the whitelist for an authoritative server.

In addition, two other key security-related issues should be taken into consideration:

9.1. DNSSEC Considerations

DNS security extensions defined in [RFC4033], [RFC4034], and [RFC4035] use cryptographic digital signatures to provide origin authentication and integrity assurance for DNS data. This is done by creating signatures for DNS data on a Security-Aware Authoritative Name Server that can be used by Security-Aware Resolvers to verify the answers. Since DNS Whitelisting is implemented on an authoritative DNS server, which provides different answers depending upon which DNS resolver has sent a query, the DNSSEC chain of trust is not altered. Even though the authoritative DNS server will not always return a AAAA resource record when one exists, respective A resource records and AAAA resource records can and should both be signed. Therefore there are no DNSSEC implications per se. However, any implementer of DNS Whitelisting should be careful if they implement both DNSSEC signing of their domain and also DNS Whitelisting of that same domain. Specifically, those domains should ensure that resource records are being appropriately and reliably signed, which may present modest incremental operational and/or technical challenges.

However, as noted in fourth paragraph of Section 4.2, end user networks may also choose to implement tools at their disposal in order to address IPv6-related impairments. One of those tools could involve unspecified changes to the configuration of their DNS recursive resolvers. If those are Security-Aware Resolvers, modifying or suppressing AAAA resource records for a DNSSEC-signed domain will be problematic and could break the chain of trust established with DNSSEC.

9.2. Authoritative DNS Response Consistency Considerations

In addition to the considerations raised in Section 9.1, it is conceivable that security concerns may arise when end users or other parties notice that the responses sent from an authoritative DNS server appear to vary from one network or one DNS recursive resolver to another. This may give rise to concerns that, since the authoritative responses vary that there is some sort of security issue and/or some or none of the responses can be trusted. While this may seem a somewhat obscure concern, domains nonetheless may wish to consider this when contemplating whether or not to pursue DNS Whitelisting.

10. Privacy Considerations

As noted in Section 5.3.1, there may be methods to detect IPv6-related impairments for a particular end user. For example, this may be possible when an end user visits the website of a particular domain. In that example, there are likely no privacy considerations in automatically communicating to that end user that the domain has detected a particular impairment. However, if that domain decided to share information concerning that particular end user with their network operator or another party, then the visited domain may wish to in some manner advise the end user of this or otherwise seek to obtain the user's consent to such information sharing. This may be achieved in a wide variety of ways, from presenting a message asking the user for consent (which will of course help them solve a technical problem of which they are likely unaware) to adding this to a domain's website terms of use / service. Such information sharing and communication of such sharing to end users may well vary by geographic area and/or legal jurisdiction. Thus, a domain should consider any potential privacy issues these sorts of scenarios.

To the extent that domains or network operators decide to publish impairment statistics, they should not identify individual hosts, host identifiers, or users.

11. IANA Considerations

There are no IANA considerations in this document.

12. Contributors

The following people made significant textual contributions to this document and/or played an important role in the development and evolution of this document:

- John Brzozowski
- Chris Griffiths
- Tom Klieber
- Yiu Lee
- Rich Woundy

13. Acknowledgements

The author and contributors also wish to acknowledge the assistance of the following individuals or groups. Some of these people provided helpful and important guidance in the development of this document and/or in the development of the concepts covered in this document. Other people assisted by performing a detailed review of this document, and then providing feedback and constructive criticism for revisions to this document, or engaged in a healthy debate over the subject of the document. All of this was helpful and therefore the following individuals merit acknowledgement:

- Bernard Aboba
- Jari Arkko
- Frank Bulk
- Brian Carpenter
- Lorenzo Colitti
- Alissa Cooper
- Dave Crocker
- Ralph Droms
- Wesley Eddy
- J.D. Falk
- Adrian Farrel
- Stephen Farrell
- Tony Finch
- Karsten Fleischhauer
- Wesley George
- Philip Homburg
- Jerry Huang
- Ray Hunter

- Joel Jaeggli
- Erik Kline
- Suresh Krishnan
- Victor Kuarsingh
- John Leslie
- John Mann
- Danny McPherson
- Milo Medin
- Martin Millnert
- Russ Mundy
- Thomas Narten
- Pekka Savola
- Robert Sparks
- Barbara Stark
- Joe Touch
- Hannes Tschofenig
- Tina Tsou
- Members of the Broadband Internet Technical Advisory Group (BITAG)

14. References

14.1. Normative References

- [RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, November 1987.
- [RFC1794] Brisco, T., "DNS Support for Load Balancing", RFC 1794, April 1995.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and

- E. Lear, "Address Allocation for Private Internets",
BCP 5, RFC 1918, February 1996.
- [RFC1958] Carpenter, B., "Architectural Principles of the Internet",
RFC 1958, June 1996.
- [RFC2775] Carpenter, B., "Internet Transparency", RFC 2775,
February 2000.
- [RFC2782] Gulbrandsen, A., Vixie, P., and L. Esibov, "A DNS RR for
specifying the location of services (DNS SRV)", RFC 2782,
February 2000.
- [RFC2956] Kaat, M., "Overview of 1999 IAB Network Layer Workshop",
RFC 2956, October 2000.
- [RFC3234] Carpenter, B. and S. Brim, "Middleboxes: Taxonomy and
Issues", RFC 3234, February 2002.
- [RFC3724] Kempf, J., Austein, R., and IAB, "The Rise of the Middle
and the Future of End-to-End: Reflections on the Evolution
of the Internet Architecture", RFC 3724, March 2004.
- [RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S.
Rose, "DNS Security Introduction and Requirements",
RFC 4033, March 2005.
- [RFC4034] Arends, R., Austein, R., Larson, M., Massey, D., and S.
Rose, "Resource Records for the DNS Security Extensions",
RFC 4034, March 2005.
- [RFC4035] Arends, R., Austein, R., Larson, M., Massey, D., and S.
Rose, "Protocol Modifications for the DNS Security
Extensions", RFC 4035, March 2005.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms
for IPv6 Hosts and Routers", RFC 4213, October 2005.

14.2. Informative References

- [Heise] Heise.de, "The Big IPv6 Experiment", Heise.de
Website <http://www.h-online.com>, January 2011, <[http://www.h-online.com/features/
The-big-IPv6-experiment-1165042.html](http://www.h-online.com/features/The-big-IPv6-experiment-1165042.html)>.
- [IETF-77-DNSOP]
Gashinsky, I., "IPv6 & recursive resolvers: How do we make
the transition less painful?", IETF 77 DNS Operations

Working Group, March 2010,
<<http://www.ietf.org/proceedings/77/slides/dnsop-7.pdf>>.

[IPv6-Brokenness]

Anderson, T., "Measuring and Combating IPv6 Brokenness",
Reseaux IP Europeens (RIPE) 61st Meeting, November 2010,
<<http://ripe61.ripe.net/presentations/162-ripe61.pdf>>.

[IPv6-Growth]

Colitti, L., Gunderson, S., Kline, E., and T. Refice,
"Evaluating IPv6 adoption in the Internet", Passive and
Active Management (PAM) Conference 2010, April 2010,
<<http://www.google.com/research/pubs/archive/36240.pdf>>.

[Impairment-Tracker]

Anderson, T., "IPv6 dual-stack client loss in Norway",
Website , May 2011, <<http://www.fud.no/ipv6/>>.

[Motion]

Newton, I., "Mathematical Principles of Natural Philosophy
(Philosophiae Naturalis Principia Mathematica)",
Principia Mathematical Principles of Natural Philosophy
(Philosophiae Naturalis Principia Mathematica), July 1687,
<http://en.wikipedia.org/wiki/Newton's_laws_of_motion>.

[NW-Article-DNS-WL]

Marsan, C., "Google, Microsoft, Netflix in talks to create
shared list of IPv6 users", Network World , March 2010, <<http://www.networkworld.com/news/2010/032610-dns-ipv6-whitelist.html>>.

[NW-Article-DNSOP]

Marsan, C., "Yahoo proposes 'really ugly hack' to DNS",
Network World , March 2010, <<http://www.networkworld.com/news/2010/032610-yahoo-dns.html>>.

[Rethinking]

Blumenthal, M. and D. Clark, "Rethinking the design of the
Internet: The end to end arguments vs. the brave new
world", ACM Transactions on Internet Technology Volume 1,
Number 1, Pages 70-109, August 2001, <http://dspace.mit.edu/bitstream/handle/1721.1/1519/TPRC_Clark_Blumenthal.pdf>.

[Tussle]

Braden, R., Clark, D., Sollins, K., and J. Wroclawski,
"Tussle in Cyberspace: Defining Tomorrow's Internet",
Proceedings of ACM Sigcomm 2002, August 2002, <<http://groups.csail.mit.edu/ana/Publications/PubPDFs/Tussle2002.pdf>>.

- [W6D] The Internet Society, "World IPv6 Day - June 8, 2011",
Internet Society Website <http://www.isoc.org>,
January 2011, <<http://isoc.org/wp/worldipv6day/>>.
- [WL-Concerns] Brzozowski, J., Griffiths, C., Klieber, T., Lee, Y.,
Livingood, J., and R. Woundy, "IPv6 DNS Whitelisting -
Could It Hinder IPv6 Adoption?", ISOC Internet Society
IPv6 Deployment Workshop, April 2010, <[http://
www.comcast6.net/
IPv6_DNS_Whitelisting_Concerns_20100416.pdf](http://www.comcast6.net/IPv6_DNS_Whitelisting_Concerns_20100416.pdf)>.
- [WL-Ops] Kline, E., "IPv6 Whitelist Operations", Google Google IPv6
Implementors Conference, June 2010, <[http://
sites.google.com/site/ipv6implementors/2010/agenda/
IPv6_Whitelist_Operations.pdf](http://sites.google.com/site/ipv6implementors/2010/agenda/IPv6_Whitelist_Operations.pdf)>.
- [Wild-Resolvers] Ager, B., Smaragdakis, G., Muhlbauer, W., and S. Uhlig,
"Comparing DNS Resolvers in the Wild", ACM Sigcomm
Internet Measurement Conference 2010, November 2010,
<<http://conferences.sigcomm.org/imc/2010/papers/pl15.pdf>>.

Appendix A. Document Change Log

[RFC Editor: This section is to be removed before publication]

-06: Removed the Open Issue #8 concerning the document name, at the direction of Joel Jaeggli. Removed Open Issue #2 from J.D. Falk and removed Open Issue #3 from Ray Hunter, as confirmed on the v6ops WG mailing list. Revised the Abstract and Intro as recommended by Tony Finch. Per Dave Crocker, updated the diagram following remedial ASCII art assistance, added a reference regarding IPv4-brokenness statistics. Removed Open Issue #1, after validating proper reference placement and removing NAT444 reference. Updates per Ralph Droms' review for the IESG. Closed Open Issue #4, Per Joe Touch, moved section 8 to just after section 3 - and also moved up section 6 and merged it. Closed Open Issue #5, per Dave Crocker and John Leslie, simplifying the document more, consolidating sections, etc. Closed Open Issue #6. Closed Open Issue #7, per Jari Arkko, ensuring all motivations are accounted for, etc. Closed Open Issue #9, per Stephen Farrell, re. World IPv6 Day (retained reference but re-worded those sections). Removed the happy-eyeballs reference since this was an informative reference and the draft could be delayed due to that dependency. ALL OPEN ITEMS ARE NOW CLOSED.

-05: Additional changes requested by Stephen Farrell intended to

close his Discuss on the I-D. These changes were in Sections 6.2 and 8.3. Also shortened non-RFC references at Stephen's request.

-04: Made changed based on feedback received during IESG review. This does NOT include updated from the more general IETF last call - that will be in a -05 version of the document. Per Ralph Droms, change the title of 6.2 from "Likely Deployment Scenarios" to "General Implementation Variations", as well as changes to improve the understanding of sentences in Sections 2, 3, 4, and 8.2. Per Adrian Farrel, made a minor change to Section 3. Per Robert Sparks, to make clear in Section 2 that whitelisting is done on authoritative servers and not DNS recursive resolvers, and to improve Section 8.3 and add a reference to I-D.ietf-v6ops-happy-eyeballs. Per Wesley Eddy, updated Section 7.3.2 to address operational concerns and re-titled Section 8 from "Solutions" to "General Implementation Variations". Per Stephen Farrell, added text to Section 8.1 and Section 6.2, with a reference to 8.1 in the Introduction, to say that universal deployment is considered harmful. Added text to Section 2 per the v6ops list discussion to indicate that whitelisting is independent of the IP address family of the end user host or resolver. There was also discussion with the IESG to change the name of the draft to IPv6 DNS Resolver Whitelisting or IPv6 AAAA DNS Resolver Whitelisting (as suggested originally by John Mann) but there was not a strong consensus to do so. Added a new section 7.7, at the suggestion of Philip Homburg. Per Joe Touch, added a new Section 8.4 on blacklisting as an alternative, mentioned blacklisting in Section 2, added a new Section 7.8 on the use of 3rd party resolvers, and updated section 6.2 to change Internet Draft documents to RFCs. Minor changes from Barbara Stark. Changes to the Privacy Considerations section based on feedback from Alissa Cooper. Changed "highly-trafficked" domains to "high-traffic" domains. Per Bernard Aboba, added text noting that a whitelist may be manually or automatically updated, contrasting whitelisting with blacklisting, reorganized Section 3, added a note on multiple clearinghouses being possible. Per Pekka Savola, added a note regarding multiple clearinghouses to the Ad Hoc section, corrected grammar in Section 7.5, reworded Section 7.3.7, corrected the year in a RIPE reference citation. Also incorporated general feedback from the Broadband Internet Technical Advisory Group. Per Jari Arkko, simplified the introduction to the Implications section, played down possible impacts on RFC 4213, added caveats to Section 8.3.2 on the utility of transition names, re-wrote Section 9. Updated the Abstract and Introduction, per errors noted by Tony Finch. Updated the Security Considerations based on feedback from Russ Mundy. Per Ray Hunter, added some text to the De-Whitelisting implications section regarding effects on networks of switching from IPv6 to IPv4. Updated 7.3.1 per additional feedback from Karsten Fleischhauer. Per Dave Crocker, added a complete description of the practice to the Abstract, added a

note to the Introduction that the operational impacts are particularly acute at scale, added text to Intro to make clear this practice affects all protocols and not just HTTP, added a new query/response diagram, added text to the Abstract and Introduction noting that this is an IPv6 transition mechanism, and too many other changes to list.

-03: Several changes suggested by Joel Jaeggli at the end of WGLC. This involved swapping the order of Section 6.1 and 6.2, among other changes to make the document more readable, understandable, and tonally balanced. As suggested by Karsten Fleischhauer, added a reference to RFC 4213 in Section 7.1, as well as other suggestions to that section. As suggested by Tina Tsou, made some changes to the DNSSEC section regarding signing. As suggested by Suresh Krishnan, made several changes to improve various sections of the document, such as adding an alternative concerning the use of `ipv6.domain`, improving the system logic section, and shortening the reference titles. As suggested by Thomas Narten, added some text regarding the imperfection of making judgements as to end user host impairments based upon the DNS recursive resolver's IP and/or network. Finally, made sure that variations in the use of 'records' and 'resource records' was updated to 'resource records' for uniformity and to avoid confusion.

-02: Called for and closed out feedback on `dnsop` and `v6ops` mailing lists. Closed out open feedback items from IETF 79. Cleared I-D nits issues, added a section on whether or not this is recommended, made language less company-specific based on feedback from Martin Millnert, Wes George, and Victor Kuarsingh. Also mentioned World IPv6 Day per Wes George's suggestion. Added references to the ISOC World IPv6 Day and the Heise.de test at the suggestion of Jerry Huang, as well as an additional implication in 7.3.7. Made any speculation on IPv4 impairment noted explicitly as such, per feedback from Martin Millnert. Added a reference to DNS SRV in the load balancing section. Added various other references. Numerous changes suggested by John Brzozowski in several sections, to clean up the document. Moved up the section on why whitelisting is performed to make the document flow more logically. Added a note in the ad hoc deployment scenario explaining that a deployment may be temporary, and including more of the perceived benefits of this tactic. Added a Privacy Considerations section to address end-user detection and communication.

-01: Incorporated feedback received from Brian Carpenter (from 10/19/2010), Frank Bulk (from 11/8/2010), and Erik Kline (from 10/1/2010). Also added an informative reference at the suggestion of Wes George (from from 10/22/2010). Closed out numerous editorial notes, and made a variety of other changes.

-00: First version published as a v6ops WG draft. The preceding individual draft was draft-livingood-dns-whitelisting-implications-01. IMPORTANT TO NOTE that no changes have been made yet based on WG and list feedback. These are in queue for a -01 update.

Appendix B. Open Issues

[RFC Editor: This section is to be removed before publication]

Author's Address

Jason Livingood
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US

Email: jason_livingood@cable.comcast.com
URI: <http://www.comcast.com>

IPv6 Operations
Internet-Draft
Intended status: Informational
Expires: August 30, 2012

J. Livingood
Comcast
February 27, 2012

Considerations for Transitioning Content to IPv6
draft-ietf-v6ops-v6-aaaa-whitelisting-implications-11

Abstract

This document describes considerations for the transition of end user content on the Internet to IPv6. While this is tailored to address end user content, which is typically web-based, many aspects of this document may be more broadly applicable to the transition to IPv6 of other applications and services. This document explores the challenges involved in the transition to IPv6, potential migration tactics, possible migration phases, and other considerations. The audience for this document is the Internet community generally, particularly IPv6 implementers.

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 30, 2012.

Copyright Notice

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

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. Challenges When Transitioning Content to IPv6	4
2.1. IPv6-Related Impairment	5
2.2. Operational Maturity Concerns	5
2.3. Volume-Based Concerns	5
3. IPv6 Adoption Implications	6
4. Potential Migration Tactics	6
4.1. Solve Current End User IPv6 Impairments	7
4.2. Use IPv6-Specific Names	7
4.3. Implement DNS Resolver Whitelisting	8
4.3.1. How DNS Resolver Whitelisting Works	10
4.3.2. Similarities to Content Delivery Networks and Global Server Load Balancing	15
4.3.3. Similarities to DNS Load Balancing	15
4.3.4. Similarities to Split DNS	15
4.3.5. Related Considerations	16
4.4. Implement DNS Blacklisting	17
4.5. Transition Directly to Native Dual Stack	18
5. Potential Implementation Phases	19
5.1. No Access to IPv6 Content	19
5.2. Using IPv6-Specific Names	19
5.3. Deploying DNS Resolver Whitelisting Using Manual Processes	19
5.4. Deploying DNS Resolver Whitelisting Using Automated Processes	19
5.5. Turning Off DNS Resolver Whitelisting	19
5.6. Deploying DNS Blacklisting	20
5.7. Fully Dual-Stack Content	20
6. Other Considerations	20
6.1. Security Considerations	20
6.2. Privacy Considerations	21
6.3. Considerations with Poor IPv4 and Good IPv6 Transport	22
6.4. IANA Considerations	23
7. Contributors	23
8. Acknowledgements	23
9. References	25
9.1. Normative References	25
9.2. Informative References	26
Appendix A. Document Change Log	28
Appendix B. Open Issues	31

Author's Address	31
----------------------------	----

1. Introduction

This document describes considerations for the transition of end user content on the Internet to IPv6. While this is tailored to address end user content, which is typically web-based, many aspects of this document may be more broadly applicable to the transition to IPv6 of other applications and services. The issues explored herein will be of particular interest to major web content sites (sometimes described hereinafter as "high-service-level domains"), which have specific and unique concerns relating to maintaining a high-quality user experience for all of their users during their transition to IPv6. This document explores the challenges involved in the transition to IPv6, potential migration tactics, possible migration phases, and other considerations. Some sections of this document also include information about the potential implications of various migration tactics or phased approaches to the transition to IPv6.

2. Challenges When Transitioning Content to IPv6

The goal in transitioning content to IPv6 is to make that content natively dual-stack enabled, which provides native access to all end users via both IPv4 and IPv6. However, there are technical and operational challenges in being able to transition smoothly for all end users, which has led to the development of a variety of migration tactics. A first step in understanding various migration tactics is to first outline the challenges involved in moving content to IPv6.

Implementers of these various migration tactics are attempting to protect users of their services from having a negative experience (poor performance) when they receive DNS responses containing AAAA resource records or when attempting to use IPv6 transport. There are two main concerns which pertain to this practice; one of which is IPv6-related impairment and the other which is the maturity or stability of IPv6 transport (and associated network operations) for high-service-level domains. Both can negatively affect the experience of end users.

Not all domains may face the same challenges in transitioning content to IPv6, since the user base of each domain, traffic sources, traffic volumes, and other factors obviously will vary between domains. As a result, while some domains have used an IPv6 migration tactic, others have run brief IPv6 experiments and then decided to simply turn on IPv6 for the domain without further delay and without using any specialized IPv6 migration tactics [Heise]. Each domain should therefore consider its specific situation when formulating a plan to move to IPv6; there is not one approach that will work for every domain.

2.1. IPv6-Related Impairment

Some implementers have observed that when they added AAAA resource records to their authoritative DNS servers in order to support IPv6 access to their content that a small fraction of end users had slow or otherwise impaired access to a given web site with both AAAA and A resource records. The fraction of users with such impaired access has been estimated to be as high as 0.078% of total Internet users [IETF-77-DNSOP] [NW-Article-DNSOP] [IPv6-Growth] [IPv6-Brokenness].

While it is outside the scope of this document to explore the various reasons why a particular user's system (host) may have impaired IPv6 access, and the potential solutions [I-D.ietf-v6ops-happy-eyeballs] [RFC6343], for the users who experience this impairment it has a very real performance impact. It would impact access to all or most dual stack services to which the user attempts to connect. This negative end user experience can range from somewhat slower than usual access (as compared to native IPv4-based access), to extremely slow access, to no access to the domain's resources whatsoever. In essence, whether the end user even has an IPv6 address or not, merely by receiving a AAAA record response the user either cannot access a Fully Qualified Domain Name (FQDN, representing a service or resource sought) or it is so slow that the user gives up and assumes the destination is unreachable.

2.2. Operational Maturity Concerns

Some implementers have discovered that network operations, operations support and business support systems, and other operational processes and procedures are less mature for IPv6 as compared to IPv4, since IPv6 has not heretofore been pervasively deployed. This operational immaturity may be observed not just within the network of a given domain but also in any directly or indirectly interconnected networks. As a result, many domains consider it prudent to undertake any network changes which will cause traffic to shift to IPv6 gradually in order to provide time and experience for IPv6 operations and network practices mature.

2.3. Volume-Based Concerns

While Section 2.2 pertains to risks due to immaturity in operations, a related concern is that some technical issues may not become apparent until some moderate to high volume of traffic is sent via IPv6 to and from a domain. As above, this may be the case not just within the network of that domain but also for any directly or indirectly interconnected networks. Furthermore, compared to domains with small to moderate traffic volumes, whether by the count of end users or count of bytes transferred, high-traffic domains receive

such a level of usage that it is prudent to undertake any network changes gradually and in a manner which minimizes the risk of disruption. One can imagine that for one of the top ten sites globally, for example, the idea of suddenly turning on a significant amount of IPv6 traffic is quite daunting and would carry a relatively high risk of network and/or other disruptions.

3. IPv6 Adoption Implications

It is important that the challenges in transitioning content to IPv6 noted in Section 2 are addressed, especially for high-service-level domains. Some high-service-level domains may find the prospect of transitioning to IPv6 extremely daunting without having some ability to address these challenges and to incrementally control their transition to IPv6. Lacking such controls, some domains may choose to substantially delay their transition to IPv6. A substantial delay in content moving to IPv6 could certainly mean there are somewhat fewer motivating factors for network operators to deploy IPv6 to end user hosts (though they have many significant motivating factors that are largely independent of content). At the same time, unless network operators transition to IPv6, there are of course fewer motivations for domain owners to transition content to IPv6. Without progress in each part of the Internet ecosystem, networks and/or content sites may delay, postpone, or cease adoption of IPv6, or to actively seek alternatives to it. Such alternatives may include the use of multi-layer or large scale network address translation (NAT), which is not preferred relative to native dual stack.

Obviously, transitioning content to IPv6 is important to IPv6 adoption overall. While challenges do exist, such a transition is not an impossible task for a domain to undertake. A range of potential migration tactics, as noted below in Section 4, can help meet these challenges and enable a domain to successfully transition content and other services to IPv6.

4. Potential Migration Tactics

Domains have a wide range of potential tactics at their disposal that may be used to facilitate the migration to IPv6. This section includes many of the key tactics that could be used by a domain but it is by no means an exhaustive or exclusive list. Only a specific domain can judge whether or not a given (or any) migration tactic applies to their domain and meets their needs. A domain may also decide to pursue several of these tactics in parallel. Thus, the usefulness of each tactic and the associated pros and cons will vary from domain to domain.

4.1. Solve Current End User IPv6 Impairments

Domains can endeavor to fix the underlying technical problems experienced by their end users during the transition to IPv6, as noted in Section 2.1. One challenge with this option is that a domain may have little or no control over the network connectivity, operating system, client software (such as a web browser), and/or other capabilities of the end users of that domain. In most cases a domain is only in a position to influence and guide their end users. While this is not the same sort of direct control which may exist in an enterprise network for example, major domains are likely to be trusted by their end users and may therefore be able to influence and guide these users in solving any IPv6-related impairments.

Another challenge is that end user impairments are something that one domain on their own cannot solve. This means that domains may find it more effective to coordinate with many others in the Internet community to solve what is really a collective problem that affects the entire Internet. Of course, it can sometimes be difficult to motivate members of the Internet community to work collectively towards such a goal, sharing the labor, time, and costs related to such an effort. However, World IPv6 Day [W6D] shows that such community efforts are possible and despite any potential challenges, the Internet community continues to work together in order to solve end user IPv6 impairments.

One potential tactic may be to identify which users have such impairments and then to communicate this information to affected users. Such end user communication is likely to be most helpful if the end user is not only alerted to a potential problem but is given careful and detailed advice on how to resolve this on their own, or is guided to where they can seek help in doing so. Another potential tactic is for a domain to collect, track over time, and periodically share with the Internet community the rate of impairment observed for a domain. In any such end user IPv6-related analysis and communication, Section 6.2 is worth taking into account.

However, while these tactics can help reduce IPv6-related impairments Section 2.1, they do not address either operational maturity concerns noted in Section 2.2 or volume-based concerns noted in Section 2.3, which should be considered and addressed separately.

4.2. Use IPv6-Specific Names

Another potential migration tactic is for a domain to gain experience using a special Fully-Qualified Domain Name (FQDN). This has become typical for domains beginning the transition to IPv6, whereby an address-family-specific name such as `ipv6.example.com` or

www.ipv6.example.com is used. An end user would have to know to use this special IPv6-specific name; it is not the same name used for regular traffic.

This special IPv6-specific name directs traffic to a host or hosts which have been enabled for native IPv6 access. In some cases this name may point to hosts which are separate from those used for IPv4 traffic (via www.example.com), while in other cases it may point to the same hosts used for IPv4 traffic. A subsequent phase, if separate hosts are used to support special IPv6-specific names, is to move to the same hosts used for regular traffic in order to utilize and exercise production infrastructure more fully. Regardless of whether or not dedicated hosts are used, the use of the special name is a way to incrementally control traffic as a tool for a domain to gain IPv6 experience and increase IPv6 use on a relatively controlled basis. Any lessons learned can then inform plans for a full transition to IPv6. This also provides an opportunity for a domain to develop any necessary training for staff, to develop IPv6-related testing procedures for their production network and lab, to deploy IPv6 functionality into their production network, and to develop and deploy IPv6-related network and service monitors. It is also an opportunity to add a relatively small amount of IPv6 traffic to ensure that network gear, network interconnects, and IPv6 routing in general is working as expected.

While using a special IPv6-specific name is a good initial step to functionally test and prepare a domain for IPv6, including developing and maturing IPv6 operations, as noted in Section 2.2, the utility of the tactic is limited since users must know the IPv6-specific name, the traffic volume will be low, and the traffic is unlikely to be representative of the general population of end users (they are likely to be self-selecting early adopters and more technically advanced than average), among other reasons. As a result, any concerns and risks related to traffic volume as noted Section 2.3 should still be considered and addressed separately.

4.3. Implement DNS Resolver Whitelisting

Another potential tactic, especially when a high-service-level domain is ready to move beyond an IPv6-specific name, as described in Section 4.2, is to selectively return AAAA resource records (RRs), which contain IPv6 addresses. This selective response of DNS records is performed by an authoritative DNS servers for a domain in response to DNS queries sent by DNS recursive resolvers [RFC1035]. This is commonly referred to in the Internet community as "DNS Resolver Whitelisting", and will be referred to as such hereafter, though in essence it is simply a tactic enabling the selective return of DNS records based upon various technical factors. An end user is seeking

a resource by name, and this selective response mechanism enables what is perceived to be the most reliable and best performing IP address family to be used (IPv4 or IPv6). It shares similarities with Content Delivery Networks, Global Server Load Balancing, DNS Load Balancing, and Split DNS, as described below in Section 4.3.2, Section 4.3.3, Section 4.3.4. A few high-service-level domains have either implemented DNS Resolver Whitelisting (one of many migration tactics they have used or are using) or are considering doing so [NW-Article-DNS-WL] [WL-Ops].

This is a migration tactic used by domains as a method for incrementally transitioning inbound traffic to a domain to IPv6. If an incremental tactic like this is not used, a domain might return AAAA resource records to any relevant DNS query, meaning the domain could go quickly from no IPv6 traffic to potentially a significant amount as soon as the AAAA resource records are published. When DNS Resolver Whitelisting is implemented, a domain's authoritative DNS will selectively return a AAAA resource record to DNS recursive resolvers on a whitelist maintained by the domain, while returning no AAAA resource records to DNS recursive resolvers which are not on that whitelist. This tactic will not have a direct impact on reducing IPv6-related impairments Section 2.1, though it can help a domain address operational maturity concerns Section 2.2 and concerns and risks related to traffic volume Section 2.3. While DNS Resolver Whitelisting does not solve IPv6-related impairments, it can help a domain to avoid users that have them. As a result, the tactic removes their impact in all but the few networks that are whitelisted. DNS Resolver Whitelisting also allows a website operator to protect non-IPv6 networks (i.e. networks that do not support IPv6 and/or do not have plans to do so in the future) from IPv6-related impairments in their networks. Finally, domains using this tactic should understand that the onus is on them to ensure that the servers being whitelisted represent a network that has proven to their satisfaction that they are IPv6-ready and this will not create a poor end user experience for users of the whitelisted server.

There are of course challenges and concerns relating to DNS Resolver Whitelisting. Some of the concerns with a whitelist of DNS recursive resolvers may be held by parties other than the implementing domain, such as network operators or end users that may not have had their DNS recursive resolvers added to a whitelist. Additionally, the IP address of a DNS recursive resolver is not a precise indicator of the IPv6 preparedness, or lack of IPv6-related impairment, of end user hosts which query (use) a particular DNS recursive resolver. While the IP addresses of DNS recursive resolvers on networks known to have deployed IPv6 may be an imperfect proxy for judging IPv6 preparedness, or lack of IPv6-related impairment, it is one of the better available methods at the current time. For example,

implementers have found that it is possible to measure the level of IPv6 preparedness of the end users behind any given DNS recursive resolver by conducting ongoing measurement of the IPv6 preparedness of end users querying for one-time-use hostnames and then correlating the domain's authoritative DNS server logs with their web server logs. This can help implementers form a good picture of which DNS recursive resolvers have working IPv6 users behind them and which do not, what the latency impact of turning on IPv6 for any given DNS recursive resolver is, etc. In addition, given the current state of global IPv6 deployment, this migration tactic allows content providers to selectively expose the availability of their IPv6 services. While opinions in the Internet community concerning DNS Resolver Whitelisting are understandably quite varied, there is clear consensus that DNS Resolver Whitelisting can be a useful tactic for use during the transition of a domain to IPv6. In particular, some high-service-level domains view DNS Resolver Whitelisting as one of the few practical and low-risk approaches enabling them to transition to IPv6, without which their transition may not take place for some time. However, there is also consensus that this practice is workable on a manual basis (see below) only in the short-term and that it will not scale over the long-term. Thus, some domains may find DNS Resolver Whitelisting a beneficial temporary tactic in their transition to IPv6.

At the current time, generally speaking, a domain that implements DNS Resolver Whitelisting does so manually. This means that a domain manually maintains a list of networks that are permitted to receive IPv6 records (via their DNS resolver IP addresses) and that these networks typically submit applications, or follow some other process established by the domain, in order to be added to the DNS Whitelist. However, implementers foresee that a subsequent phase of DNS Resolver Whitelisting is likely to emerge in the future, possibly in the near future. In this new phase a domain would return IPv6 and/or IPv4 records dynamically based on automatically detected technical capabilities, location, or other factors. It would then function much like (or indeed as part of) global server load balancing, a common practice already in use today, as described in Section 4.3.2. Furthermore, in this future phase, networks would be added to and removed from a DNS Whitelist automatically, and possibly on a near-real-time basis. This means, crucially, that networks would no longer need to apply to be added to a whitelist, which may alleviate many of the key concerns that network operators may have with this tactic when it is implemented on a manual basis.

4.3.1. How DNS Resolver Whitelisting Works

Using a "whitelist" in a generic sense means that no traffic (or traffic of a certain type) is permitted to the destination host

unless the originating host's IP address is contained in the whitelist. In contrast, using a "blacklist" means that all traffic is permitted to the destination host unless the originating host's IP address is contained in the blacklist. In the case of DNS Resolver Whitelisting, the resource that an end user seeks is a name, not an IP address or IP address family. Thus, an end user is seeking a name such as `www.example.com`, without regard to the underlying IP address family (IPv4 or IPv6) which may be used to access that resource.

DNS Resolver Whitelisting is implemented in authoritative DNS servers, not in DNS recursive resolvers. These authoritative DNS servers selectively return AAAA resource records using the IP address of the DNS recursive resolver that has sent it a query. Thus, for a given operator of a website, such as `www.example.com`, the domain operator implements whitelisting on the authoritative DNS servers for the domain `example.com`. The whitelist is populated with the IPv4 and/or IPv6 addresses or prefix ranges of DNS recursive resolvers on the Internet, which have been authorized to receive AAAA resource record responses. These DNS recursive resolvers are operated by third parties, such as Internet Service Providers (ISPs), universities, governments, businesses, and individual end users. If a DNS recursive resolver is not matched in the whitelist, then AAAA resource records WILL NOT be sent in response to a query for a hostname in the `example.com` domain (and an A record would be sent). However, if a DNS recursive resolver is matched in the whitelist, then AAAA resource records WILL be sent. As a result, while Section 2.2 of [RFC4213] notes that a stub resolver can make a choice between whether to use a AAAA record or A record response, with DNS Resolver Whitelisting the authoritative DNS server can also decide whether to return a AAAA record, an A record, or both record types.

When implemented on a manual basis, DNS Resolver Whitelisting generally means that a very small fraction of the DNS recursive resolvers on the Internet (those in the whitelist) will receive AAAA responses. The large majority of DNS recursive resolvers on the Internet will therefore receive only A resource records containing IPv4 addresses. When implemented manually, domains may find the practice imposes some incremental operational burdens insofar as it can consume staff time to maintain a whitelist (such as additions and deletions to the list), respond to and review applications to be added to a whitelist, maintain good performance levels on authoritative DNS servers as the whitelist grows, create new network monitors to check the health of a whitelist function, perform new types of troubleshooting related to whitelisting, etc. In addition, manually-based whitelisting imposes some incremental burdens on operators of DNS recursive resolvers (such as network operators), since they will need to apply to be whitelisted with any implementing domains, and will subsequently need processes and systems to track

the status of whitelisting applications, respond to requests for additional information pertaining to these applications, and track any de-whitelisting actions.

When implemented on an automated basis in the future, DNS recursive resolvers listed in the whitelist could expand and contract dynamically, and possibly in near-real-time, based on a wide range of factors. As a result, it is likely that the number of DNS recursive resolvers on the whitelist will be substantially larger than when such a list is maintained manually, and it is likely the the whitelist will grow at a rapid rate. This automation can eliminate most of the significant incremental operational burdens on both implementing domains as well as operators of DNS recursive resolvers, which is clearly a factor that is motivating implementers to work to automate this function.

Section 4.3.1.1 and Figure 1 have more details on DNS Resolver Whitelisting generally. In addition, the potential deployment models of DNS Resolver Whitelisting (manual and automated) are described in Section 5. It is also important to note that DNS Resolver Whitelisting also works independently of whether an authoritative DNS server, DNS recursive resolver, or end user host uses IPv4 transport, IPv6, or both. So, for example, whitelisting may not result in the return of AAAA responses even in those cases where the DNS recursive resolver has queried the authoritative server over IPv6 transport. This may also be the case in some situations when the end user host's original query to its DNS recursive resolver was over IPv6 transport, if that DNS recursive resolver is not on a given whitelist. One important reason for this is that even though the DNS recursive resolver may have no IPv6-related impairments, this is not a reliable predictor of whether the same is true of the end user host. This also means that a DNS whitelist can contain both IPv4 and IPv6 addresses.

4.3.1.1. Description of the Operation of DNS Resolver Whitelisting

Specific implementations will vary from domain to domain, based on a range of factors such as the technical capabilities of a given domain. As such, any examples listed herein should be considered general examples and are not intended to be exhaustive.

The system logic of DNS Resolver Whitelisting is as follows:

1. The authoritative DNS server for example.com receives DNS queries for the A (IPv4) and/or AAAA (IPv6) address resource records for the Fully Qualified Domain Name (FQDN) www.example.com, for which AAAA (IPv6) resource records exist.

2. The authoritative DNS server checks the IP address (IPv4, IPv6, or both) of the DNS recursive resolver sending the AAAA (IPv6) query against the whitelist that is the DNS Whitelist.
3. If the DNS recursive resolver's IP address IS matched in the whitelist, then the response to that specific DNS recursive resolver can contain AAAA (IPv6) address resource records.
4. If the DNS recursive resolver's IP address IS NOT matched in the whitelist, then the response to that specific DNS recursive resolver cannot contain AAAA (IPv6) address resource records. In this case, the server will likely return a response with the response code (RCODE) being set to 0 (No Error) with an empty answer section for the AAAA record query.

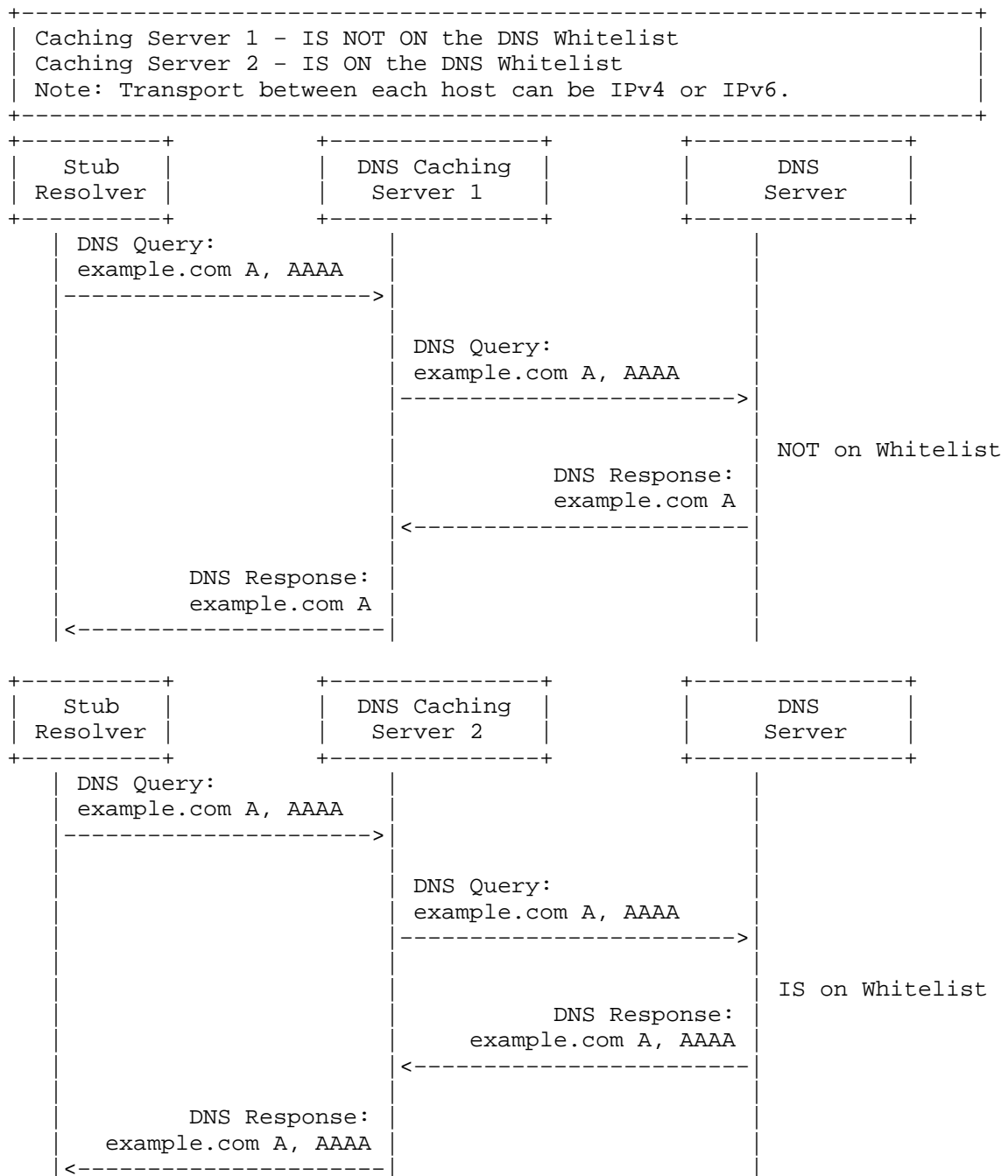


Figure 1: DNS Resolver Whitelisting Diagram

4.3.2. Similarities to Content Delivery Networks and Global Server Load Balancing

DNS Resolver Whitelisting is functionally similar to Content Delivery Networks (CDNs) and Global Server Load Balancing (GSLB). When using a CDN or GSLB, a geographically-aware authoritative DNS server function is usually part of that overall system. As a result, the use of a CDN or GSLB with an authoritative DNS server function enables the IP address resource records returned to a resolver in response to a query to vary based on the estimated geographic location of the resolver [Wild-Resolvers] or a range of other technical factors. This CDN or GSLB DNS function is performed in order to attempt to direct hosts to connect to the nearest hosts (as measured in round trip time), to the host that has the best connectivity to an end user, to route around failures, to avoid sites where maintenance work has taken down hosts, and/or to the host that will otherwise provide the best service experience for an end user at a given point in time. As a result, one can see a direct similarity to DNS Resolver Whitelisting insofar as different IP address resource records are selectively returned to resolvers based on the IP address of each resolver and/or other imputed factors related to that IP address.

4.3.3. Similarities to DNS Load Balancing

DNS Resolver Whitelisting has some similarities to DNS load balancing. There are of course many ways that DNS load balancing can be performed. In one example, multiple IP address resource records (A and/or AAAA) can be added to the DNS for a given FQDN. This approach is referred to as DNS round robin [RFC1794]. DNS round robin may also be employed where SRV resource records are used [RFC2782]. In another example, one or more of the IP address resource records in the DNS will direct traffic to a load balancer. That load balancer, in turn, may be application-aware, and pass the traffic on to one or more hosts connected to the load balancer which have different IP addresses. In cases where private IPv4 addresses are used [RFC1918], as well as when public IP addresses are used, those end hosts may not necessarily be directly reachable without passing through the load balancer first. So, similar to DNS Resolver Whitelisting, a load balancer will control what server host an end user's host communicates with when using a FQDN.

4.3.4. Similarities to Split DNS

DNS Resolver Whitelisting has some similarities to so-called split DNS, briefly described in Section 3.8 of [RFC2775]. When split DNS is used, the authoritative DNS server selectively returns different responses depending upon what host has sent the query. While

[RFC2775] notes the typical use of split DNS is to provide one answer to hosts on an Intranet (internal network) and a different answer to hosts on the Internet (external or public network), the basic idea is that different answers are provided to hosts on different networks. This is similar to the way that DNS Resolver Whitelisting works, whereby hosts on different networks which use different DNS recursive resolvers, receive different answers if one DNS recursive resolver is on the whitelist and the other is not. However, Internet transparency and Internet fragmentation concerns regarding split DNS are detailed in Section 2.1 of [RFC2956] and Section 2.7 notes concerns regarding split DNS and that it "makes the use of Fully Qualified Domain Names (FQDNs) as endpoint identifiers more complex". Section 3.5 of [RFC2956] further recommends that maintaining a stable approach to DNS operations is key during transitions, such as the one to IPv6 that is underway now, stating that "Operational stability of DNS is paramount, especially during a transition of the network layer, and both IPv6 and some network address translation techniques place a heavier burden on DNS."

4.3.5. Related Considerations

While techniques such as GLSB and DNS load balancing, which share much in common with DNS Resolver Whitelisting and are widespread, some in the community have raised a range of concerns about the practice. Some concerns are specific DNS Resolver Whitelisting [WL-Concerns]. Other concerns are not as specific and pertain to the general practice of implementing content location or other network policy controls in the "middle" of the network in a so-called "middlebox" function. Whether such DNS-related functions are really part of a middlebox is debatable. Nevertheless, implementers should at least be aware of some of the risks of middleboxes, as noted in [RFC3724]. A related document, [RFC1958] explains that the state, policies, and other functions needed in the middle of the network should be minimized as a design goal. In addition, Section 2.16 of [RFC3234] makes specific statements concerning modified DNS servers. [RFC3234] also outlines more general concerns in Section 1.2 about the introduction of new failure modes when configuration is no longer limited to two ends of a session, so that diagnosis of failures and misconfigurations could become more complex. Two additional sources worth considering are [Tussle] and [Rethinking], in which the authors note concerns regarding the introduction of new control points (such as in middleboxes), including in the DNS.

However, some state, policies, and other functions have always been necessary to enable effective, reliable, and high-quality end-to-end communications on the Internet. In addition, techniques such as Global Server Load Balancing, Content Delivery Networking, DNS Load Balancing and Split DNS are not only widely deployed but are almost

uniformly viewed as essential to the functioning of the Internet and highly beneficial to the quality of the end user experience on the Internet. These techniques have had and continue to have a beneficial effect on the experience of a wide range of Internet applications and protocols. So while there are valid concerns about implementing policy controls in the "middle" of the network, or anywhere away from edge hosts, the definition of what constitutes the middle and edge of the network is debatable in this case. This is particularly so given that GSLBs and CDNs facilitate connections from end host and the optimal content hosts, and could therefore be considered a modest and in many cases essential network policy extension of a network's edge, especially in the case of high-service-level domains.

There may be additional implications for end users that have configured their hosts to use a third party as their DNS recursive resolver, rather than the one(s) provided by their network operator. In such cases, it will be more challenging for a domain using whitelisting to determine the level of IPv6-related impairment when such third-party DNS recursive resolvers are used, given the wide variety of end user access networks which may be used and that this mix may change in unpredictable ways over time.

4.4. Implement DNS Blacklisting

With DNS Resolver Whitelisting, DNS recursive resolvers can receive AAAA resource records only if they are on the whitelist. DNS Blacklisting is by contrast the the opposite of that, whereby all DNS recursive resolvers can receive AAAA resource records unless they are on the blacklist. Some implementers of DNS Resolver Whitelisting may choose to subsequently transition to DNS Blacklisting. It is unclear when and if it may be appropriate for a domain to change from whitelisting to blacklisting. Nor is it clear how implementers will judge the network conditions to have changed sufficiently to justify disabling such controls.

When a domain uses blacklisting, they are enabling all DNS recursive resolvers to receive AAAA record responses except for what is presumed to be a relatively small number that are on the blacklist. Over time it is likely that the blacklist will become smaller as the networks associated with the blacklisted DNS recursive resolvers are able to meaningfully reduce IPv6-related impairments to some acceptable level, though it is possible that some networks may never achieve that. DNS Blacklisting is also likely less labor intensive for a domain than performing DNS Resolver Whitelisting on a manual basis. This is simply because the domain would presumably be focused on a smaller number of DNS recursive resolvers with well known IPv6-related problems.

It is also worth noting that the email industry has a long experience with blacklists and, very generally speaking, blacklists tend to be effective and well received when it is easy to discover if an IP address is on a blacklist, if there is a transparent and easily understood process for requesting removal from a blacklist, and if the decision-making criteria for placing a server on a blacklist is transparently disclosed and perceived as fair. However, in contrast to an email blacklist where a blacklisted host cannot send email to a domain at all, with DNS Resolver Whitelisting communications will still occur over IPv4 transport.

4.5. Transition Directly to Native Dual Stack

As an alternative to adopting any of the aforementioned migration tactics, domains can choose to transition to native dual stack directly by adding native IPv6 capabilities to their network and hosts and by publishing AAAA resource records in the DNS for named resources within their domain. Of course, a domain can still control this transition gradually, on a name-by-name basis, by adding native IPv6 to one name at a time, such as mail.example.com first and www.example.com later. So even a "direct" transition can be performed gradually.

It is then up to end users with IPv6-related impairments to discover and fix any applicable impairments. However, the concerns and risks related to traffic volume Section 2.3 should still be considered and managed, since those are not directly related to such impairments. Not all content providers (or other domains) may face the challenges detailed herein or face them to the same degree, since the user base of each domain, traffic sources, traffic volumes, and other factors obviously varies between domains.

For example, while some content providers have implemented DNS Resolver Whitelisting (one migration tactic), others have run IPv6 experiments whereby they added AAAA resource records and observed and measured errors, and then decided not to implement DNS Resolver Whitelisting [Heise]. A more widespread such experiment was World IPv6 Day [W6D], sponsored by the Internet Society, on June 8, 2011. This was a unique opportunity for hundreds of domains to add AAAA resource records to the DNS without using DNS Resolver Whitelisting, all at the same time. Some of the participating domains chose to leave AAAA resource records in place following the experiment based on their experiences.

Content providers can run their own independent experiments in the future, adding AAAA resource records for a brief period of time (minutes, hours, or days), and then analyzing any impacts or effects on traffic and the experience of end users. They can also simply

turn on IPv6 for their domain, which may be easier when the transition does not involve a high-service-level domain.

5. Potential Implementation Phases

The usefulness of each tactic in Section 4, and the associated pros and cons associated with each tactic, is relative to each potential implementer and will therefore vary from one implementer to another. As a result, it is not possible to say that the potential phases below make sense for every implementer. This also means that the duration of each phase will vary between implementers, and even that different implementers may skip some of these phases entirely. Finally, the tactics listed in Section 4 are by no means exclusive.

5.1. No Access to IPv6 Content

In this phase, a site is accessible only via IPv4 transport. As of the time of this document, the majority of content on the Internet is in this state and is not accessible natively over IPv6.

5.2. Using IPv6-Specific Names

One possible first step for a domain is to gain experience using a specialized new FQDN, such as `ipv6.example.com` or `www.ipv6.example.com`, as explained in Section 4.2.

5.3. Deploying DNS Resolver Whitelisting Using Manual Processes

As noted in Section 4.3, a domain could begin using DNS Resolver Whitelisting as a way to incrementally enable IPv6 access to content. This tactic may be especially interesting to high-service-level domains.

5.4. Deploying DNS Resolver Whitelisting Using Automated Processes

For a domain that decides to undertake DNS Resolver Whitelisting on a manual basis, the domain may subsequently move to perform DNS Resolver Whitelisting on an automated basis. This is explained in Section 4.3, and can significantly ease any operational burdens relating to a manually-maintained whitelist.

5.5. Turning Off DNS Resolver Whitelisting

Domains that choose to implement DNS Resolver Whitelisting generally consider it to be a temporary measure. Many implementers have announced that they plan to permanently turn off DNS Resolver Whitelisting beginning on the date of the World IPv6 Launch, on June

6, 2012 [World IPv6 Launch]. For any implementers that do not turn off DNS Resolver Whitelisting at that time, it may be unclear how each and every one will judge when the network conditions to have changed sufficiently to justify turning off DNS Resolver Whitelisting. That being said, it is clear that the extent of IPv6 deployment to end users in networks, the state of IPv6-related impairment, and the maturity of IPv6 operations are all important factors. Any such implementers may wish to take into consideration that, as a practical matter, it will be impossible to get to a point where there are no longer any IPv6-related impairments; some reasonably small number of hosts will inevitably be left behind as end users elect not to upgrade them or as some hosts are incapable of being upgraded.

5.6. Deploying DNS Blacklisting

Regardless of whether a domain has first implemented DNS Resolver Whitelisting or has never done so, DNS Blacklisting as described in Section 4.4 may become interesting. This may be at the point in time when domains wish to make their content widely available over IPv6 but still wish to protect end users of a few networks with well known IPv6 limitations from having a bad end user experience.

5.7. Fully Dual-Stack Content

A domain can arrive at this phase either following the use of a previous IPv6 migration tactic, or they may go directly to this point as noted in Section 4.5. In this phase the site's content has been made natively accessible via both IPv4 and IPv6 for all end users on the Internet, or at least without the use of any other IPv6 migration tactic.

6. Other Considerations

6.1. Security Considerations

If DNS Resolver Whitelisting is adopted, as noted in Section 4.3, then organizations which apply DNS Resolver Whitelisting policies in their authoritative servers should have procedures and systems which do not allow unauthorized parties to modify the whitelist or blacklist, just as all configuration settings for name servers should be protected by appropriate procedures and systems. Such unauthorized additions or removals from the whitelist can be damaging, causing content providers and/or network operators to incur support costs resulting from end user and/or customer contacts, as well as causing potential dramatic and disruptive swings in traffic from IPv6 to IPv4 or vice versa.

DNS security extensions defined in [RFC4033], [RFC4034], and [RFC4035] use cryptographic digital signatures to provide origin authentication and integrity assurance for DNS data. This is done by creating signatures for DNS data on a Security-Aware Authoritative Name Server that can be used by Security-Aware Resolvers to verify the answers. Since DNS Resolver Whitelisting is implemented on an authoritative DNS server, which provides different answers depending upon which DNS resolver has sent a query, the DNSSEC chain of trust is not altered. So even though an authoritative DNS server will selectively return AAAA resource records or a non-existence response, both types of response will be signed and will validate. In practical terms this means that two separate views or zones are used, each of which is signed, so that whether or not particular resource records exist, the existence or non-existence of the record can still be validated using DNSSEC. As a result, there should not be any negative impact on DNSSEC for those domains that have implemented both DNSSEC on their Security-Aware Authoritative Name Servers and also implemented DNS Resolver Whitelisting. As for any party implementing DNSSEC of course, such domains should ensure that resource records are being appropriately and reliably signed and consistent with the response being returned.

However, network operators that run DNS recursive resolvers should be careful not to modify the responses received from authoritative DNS servers. It is possible that some networks may attempt to do so in order to prevent AAAA record responses from going to end user hosts, due to some IPv6-related impairment or other lack of IPv6 readiness with that network. But when a network operates a Security-Aware Resolver, modifying or suppressing AAAA resource records for a DNSSEC-signed domain could break the chain of trust established with DNSSEC.

6.2. Privacy Considerations

As noted in Section 4.1, there is a benefit in sharing IPv6-related impairment statistics within the Internet community over time. Any statistics that are shared or disclosed publicly should be aggregate statistics, such as "the domain example.com has observed an average daily impairment rate of 0.05% in September 2011, down from 0.15% in January 2011". They should not include information that can directly or indirectly identify individuals, such as names or email addresses. Sharing only aggregate data can help protect end user privacy and any information which may be proprietary to a domain.

In addition, there are often methods to detect IPv6-related impairments for a specific end user, such as running an IPv6 test when an end user visits the website of a particular domain. Should a domain then choose to automatically communicate the facts of an

impairment to an affected user, there are likely no direct privacy considerations. However, if the domain then decided to share information concerning that particular end user with that user's network operator or another third party, then the domain may wish to consider advising the end user of this and seeking to obtain the end user's consent to share such information.

Appropriate guidelines for any information sharing likely varies by country and/or legal jurisdiction. Domains should consider any potential privacy issues when considering what information can be shared. If a domain does publish or share detailed impairment statistics, they would be well advised to avoid identifying individual hosts or users.

Finally, if a domain chooses to contact end user directly concerning their IPv6 impairments, that domain should ensure that such communication is permissible under any applicable privacy policies of the domain or its websites.

6.3. Considerations with Poor IPv4 and Good IPv6 Transport

There are situations where the differing quality of the IPv4 and IPv6 connectivity of an end user could cause complications in accessing content when a domain is using an IPv6 migration tactic. While today most end users' IPv4 connectivity is typically superior to IPv6 connectivity (if such connectivity exists at all), there could be implications when the reverse is true and an end user has markedly superior IPv6 connectivity as compared to IPv4. This is not a theoretical situation; it has been observed by at least one major content provider.

For example, in one possible scenario, a user is issued IPv6 addresses by their ISP and has a home network and devices or operating systems which fully support native IPv6. As a result this theoretical user has very good IPv6 connectivity. However, this end user's ISP has exhausted their available pool of unique IPv4 addresses, and uses NAT in order to share IPv4 addresses among end users. So for IPv4 content, the end user must send their IPv4 traffic through some additional network element (e.g. large scale NAT, proxy server, tunnel server). Use of this additional network element might cause an end user to experience sub-optimal IPv4 connectivity when certain protocols or applications are used. This user then has good IPv6 connectivity but impaired IPv4 connectivity. As a result, the user's poor IPv4 connectivity situation could potentially be exacerbated when accessing a domain which is using a migration tactic that causes this user to only be able to access content over IPv4 transport for whatever reason.

Should this sort of situation become widespread in the future, a domain may wish to take it into account when deciding how and when to transition content to IPv6.

6.4. IANA Considerations

There are no IANA considerations in this document.

7. Contributors

The following people made significant textual contributions to this document and/or played an important role in the development and evolution of this document:

- John Brzozowski
- Chris Griffiths
- Tom Klieber
- Yiu Lee
- Rich Woundy

8. Acknowledgements

The author and contributors also wish to acknowledge the assistance of the following individuals or groups. Some of these people provided helpful and important guidance in the development of this document and/or in the development of the concepts covered in this document. Other people assisted by performing a detailed review of this document, and then providing feedback and constructive criticism for revisions to this document, or engaged in a healthy debate over the subject of the document. All of this was helpful and therefore the following individuals merit acknowledgement:

- Bernard Aboba
- Mark Andrews
- Jari Arkko
- Fred Baker
- Ron Bonica

- Frank Bulk
- Brian Carpenter
- Lorenzo Colitti
- Alissa Cooper
- Dave Crocker
- Ralph Droms
- Wesley Eddy
- J.D. Falk
- Adrian Farrel
- Stephen Farrell
- Tony Finch
- Karsten Fleischhauer
- Igor Gashinsky
- Wesley George
- Philip Homburg
- Jerry Huang
- Ray Hunter
- Joel Jaeggli
- Erik Kline
- Suresh Krishnan
- Victor Kuarsingh
- Marc Lampo
- Donn Lee
- John Leslie

- John Mann
- Danny McPherson
- Milo Medin
- Martin Millnert
- Russ Mundy
- Thomas Narten
- Pekka Savola
- Robert Sparks
- Barbara Stark
- Joe Touch
- Hannes Tschofenig
- Tina Tsou
- Members of the Broadband Internet Technical Advisory Group (BITAG)

9. References

9.1. Normative References

- [RFC1035] Mockapetris, P., "Domain names - implementation and specification", STD 13, RFC 1035, November 1987.
- [RFC1794] Brisco, T., "DNS Support for Load Balancing", RFC 1794, April 1995.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC1958] Carpenter, B., "Architectural Principles of the Internet", RFC 1958, June 1996.
- [RFC2775] Carpenter, B., "Internet Transparency", RFC 2775, February 2000.
- [RFC2782] Gulbrandsen, A., Vixie, P., and L. Esibov, "A DNS RR for

specifying the location of services (DNS SRV)", RFC 2782, February 2000.

- [RFC2956] Kaat, M., "Overview of 1999 IAB Network Layer Workshop", RFC 2956, October 2000.
- [RFC3234] Carpenter, B. and S. Brim, "Middleboxes: Taxonomy and Issues", RFC 3234, February 2002.
- [RFC3724] Kempf, J., Austein, R., and IAB, "The Rise of the Middle and the Future of End-to-End: Reflections on the Evolution of the Internet Architecture", RFC 3724, March 2004.
- [RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.
- [RFC4034] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Resource Records for the DNS Security Extensions", RFC 4034, March 2005.
- [RFC4035] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Protocol Modifications for the DNS Security Extensions", RFC 4035, March 2005.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

9.2. Informative References

- [Heise] Heise.de, "The Big IPv6 Experiment", Heise.de Website <http://www.h-online.com>, January 2011, <<http://www.h-online.com/features/The-big-IPv6-experiment-1165042.html>>.
- [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.
- [IETF-77-DNSOP] Gashinsky, I., "IPv6 & recursive resolvers: How do we make the transition less painful?", IETF 77 DNS Operations Working Group, March 2010, <<http://www.ietf.org/proceedings/77/slides/dnsop-7.pdf>>.
- [IPv6-Brokenness] Anderson, T., "Measuring and Combating IPv6 Brokenness",

Reseaux IP Europeens (RIPE) 61st Meeting, November 2010,
<<http://ripe61.ripe.net/presentations/162-ripe61.pdf>>.

[IPv6-Growth]

Colitti, L., Gunderson, S., Kline, E., and T. Refice,
"Evaluating IPv6 adoption in the Internet", Passive and
Active Management (PAM) Conference 2010, April 2010,
<<http://www.google.com/research/pubs/archive/36240.pdf>>.

[NW-Article-DNS-WL]

Marsan, C., "Google, Microsoft, Netflix in talks to create
shared list of IPv6 users", Network World , March 2010, <<http://www.networkworld.com/news/2010/032610-dns-ipv6-whitelist.html>>.

[NW-Article-DNSOP]

Marsan, C., "Yahoo proposes 'really ugly hack' to DNS",
Network World , March 2010, <<http://www.networkworld.com/news/2010/032610-yahoo-dns.html>>.

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

[Rethinking]

Blumenthal, M. and D. Clark, "Rethinking the design of the
Internet: The end to end arguments vs. the brave new
world", ACM Transactions on Internet Technology Volume 1,
Number 1, Pages 70-109, August 2001, <http://dspace.mit.edu/bitstream/handle/1721.1/1519/TPRC_Clark_Blumenthal.pdf>.

[Tussle]

Braden, R., Clark, D., Sollins, K., and J. Wroclawski,
"Tussle in Cyberspace: Defining Tomorrow's Internet",
Proceedings of ACM Sigcomm 2002, August 2002, <<http://groups.csail.mit.edu/ana/Publications/PubPDFs/Tussle2002.pdf>>.

[W6D]

The Internet Society, "World IPv6 Day - June 8, 2011",
Internet Society Website <http://www.isoc.org>,
January 2011, <<http://isoc.org/wp/worldipv6day/>>.

[WL-Concerns]

Brzozowski, J., Griffiths, C., Klieber, T., Lee, Y.,
Livingood, J., and R. Woundy, "IPv6 DNS Resolver
Whitelisting - Could It Hinder IPv6 Adoption?",
ISOC Internet Society IPv6 Deployment Workshop,
April 2010, <http://www.comcast6.net/IPv6_DNS_Whitelisting_Concerns_20100416.pdf>.

[WL-Ops] Kline, E., "IPv6 Whitelist Operations", Google Google IPv6 Implementors Conference, June 2010, <http://sites.google.com/site/ipv6implementors/2010/agenda/IPv6_Whitelist_Operations.pdf>.

[Wild-Resolvers] Ager, B., Smaragdakis, G., Muhlbauer, W., and S. Uhlig, "Comparing DNS Resolvers in the Wild", ACM Sigcomm Internet Measurement Conference 2010, November 2010, <<http://conferences.sigcomm.org/imc/2010/papers/pl15.pdf>>.

[World IPv6 Launch] The Internet Society, "World IPv6 Launch Website", 2012, <<http://www.worldipv6launch.org/>>.

Appendix A. Document Change Log

[RFC Editor: This section is to be removed before publication]

-11: Minor update to one item to resolve a question from IETF Last Call (same one as -09 and -10)

-10: Minor update to one sentence to resolve a question from IETF Last Call

-09: Minor updates to resolve questions in IETF Last Call

-08: Minor updates from v6ops WGLC

-07: Significant re-write based on feedback from Jari Arkko, Joel Jaeggli, Fred Baker, Igor Gashinsky, Donn Lee, Lorenzo Colitti, and Erik Kline.

-06: Removed the Open Issue #8 concerning the document name, at the direction of Joel Jaeggli. Removed Open Issue #2 from J.D. Falk and removed Open Issue #3 from Ray Hunter, as confirmed on the v6ops WG mailing list. Revised the Abstract and Intro as recommended by Tony Finch. Per Dave Crocker, updated the diagram following remedial ASCII art assistance, added a reference regarding IPv4-brokenness statistics. Removed Open Issue #1, after validating proper reference placement and removing NAT444 reference. Updates per Ralph Droms' review for the IESG. Closed Open Issue #4, Per Joe Touch, moved section 8 to just after section 3 - and also moved up section 6 and merged it. Closed Open Issue #5, per Dave Crocker and John Leslie, simplifying the document more, consolidating sections, etc. Closed Open Issue #6. Closed Open Issue #7, per Jari Arkko, ensuring all motivations are accounted for, etc. Closed Open Issue #9, per

Stephen Farrell, re. World IPv6 Day (retained reference but reworded those sections). Removed the happy-eyeballs reference since this was an informative reference and the draft could be delayed due to that dependency. ALL OPEN ITEMS ARE NOW CLOSED.

-05: Additional changes requested by Stephen Farrell intended to close his Discuss on the I-D. These changes were in Sections 6.2 and 8.3. Also shortened non-RFC references at Stephen's request.

-04: Made changes based on feedback received during IESG review. This does NOT include updated from the more general IETF last call - that will be in a -05 version of the document. Per Ralph Droms, change the title of 6.2 from "Likely Deployment Scenarios" to "General Implementation Variations", as well as changes to improve the understanding of sentences in Sections 2, 3, 4, and 8.2. Per Adrian Farrel, made a minor change to Section 3. Per Robert Sparks, to make clear in Section 2 that whitelisting is done on authoritative servers and not DNS recursive resolvers, and to improve Section 8.3 and add a reference to I-D.ietf-v6ops-happy-eyeballs. Per Wesley Eddy, updated Section 7.3.2 to address operational concerns and re-titled Section 8 from "Solutions" to "General Implementation Variations". Per Stephen Farrell, added text to Section 8.1 and Section 6.2, with a reference to 8.1 in the Introduction, to say that universal deployment is considered harmful. Added text to Section 2 per the v6ops list discussion to indicate that whitelisting is independent of the IP address family of the end user host or resolver. There was also discussion with the IESG to change the name of the draft to IPv6 DNS Resolver Whitelisting or IPv6 AAAA DNS Resolver Whitelisting (as suggested originally by John Mann) but there was not a strong consensus to do so. Added a new section 7.7, at the suggestion of Philip Homburg. Per Joe Touch, added a new Section 8.4 on blacklisting as an alternative, mentioned blacklisting in Section 2, added a new Section 7.8 on the use of 3rd party resolvers, and updated section 6.2 to change Internet Draft documents to RFCs. Minor changes from Barbara Stark. Changes to the Privacy Considerations section based on feedback from Alissa Cooper. Changed "highly-trafficked" domains to "high-traffic" domains. Per Bernard Aboba, added text noting that a whitelist may be manually or automatically updated, contrasting whitelisting with blacklisting, reorganized Section 3, added a note on multiple clearinghouses being possible. Per Pekka Savola, added a note regarding multiple clearinghouses to the Ad Hoc section, corrected grammar in Section 7.5, reworded Section 7.3.7, corrected the year in a RIPE reference citation. Also incorporated general feedback from the Broadband Internet Technical Advisory Group. Per Jari Arkko, simplified the introduction to the Implications section, played down possible impacts on RFC 4213, added caveats to Section 8.3.2 on the utility of transition names, re-wrote Section 9. Updated the Abstract and

Introduction, per errors noted by Tony Finch. Updated the Security Considerations based on feedback from Russ Mundy. Per Ray Hunter, added some text to the De-Whitelisting implications section regarding effects on networks of switching from IPv6 to IPv4. Updated 7.3.1 per additional feedback from Karsten Fleischhauer. Per Dave Crocker, added a complete description of the practice to the Abstract, added a note to the Introduction that the operational impacts are particularly acute at scale, added text to Intro to make clear this practice affects all protocols and not just HTTP, added a new query/response diagram, added text to the Abstract and Introduction noting that this is an IPv6 transition mechanism, and too many other changes to list.

-03: Several changes suggested by Joel Jaeggli at the end of WGLC. This involved swapping the order of Section 6.1 and 6.2, among other changes to make the document more readable, understandable, and tonally balanced. As suggested by Karsten Fleischhauer, added a reference to RFC 4213 in Section 7.1, as well as other suggestions to that section. As suggested by Tina Tsou, made some changes to the DNSSEC section regarding signing. As suggested by Suresh Krishnan, made several changes to improve various sections of the document, such as adding an alternative concerning the use of `ipv6.domain`, improving the system logic section, and shortening the reference titles. As suggested by Thomas Narten, added some text regarding the imperfection of making judgements as to end user host impairments based upon the DNS recursive resolver's IP and/or network. Finally, made sure that variations in the use of 'records' and 'resource records' was updated to 'resource records' for uniformity and to avoid confusion.

-02: Called for and closed out feedback on dnsop and v6ops mailing lists. Closed out open feedback items from IETF 79. Cleared I-D nits issues, added a section on whether or not this is recommended, made language less company-specific based on feedback from Martin Millnert, Wes George, and Victor Kuarsingh. Also mentioned World IPv6 Day per Wes George's suggestion. Added references to the ISOC World IPv6 Day and the Heise.de test at the suggestion of Jerry Huang, as well as an additional implication in 7.3.7. Made any speculation on IPv4 impairment noted explicitly as such, per feedback from Martin Millnert. Added a reference to DNS SRV in the load balancing section. Added various other references. Numerous changes suggested by John Brzozowski in several sections, to clean up the document. Moved up the section on why whitelisting is performed to make the document flow more logically. Added a note in the ad hoc deployment scenario explaining that a deployment may be temporary, and including more of the perceived benefits of this tactic. Added a Privacy Considerations section to address end-user detection and communication.

-01: Incorporated feedback received from Brian Carpenter (from 10/19/2010), Frank Bulk (from 11/8/2010), and Erik Kline (from 10/1/2010). Also added an informative reference at the suggestion of Wes George (from 10/22/2010). Closed out numerous editorial notes, and made a variety of other changes.

-00: First version published as a v6ops WG draft. The preceding individual draft was draft-livingood-dns-whitelisting-implications-01. IMPORTANT TO NOTE that no changes have been made yet based on WG and list feedback. These are in queue for a -01 update.

Appendix B. Open Issues

[RFC Editor: This section is to be removed before publication]

Check references to ensure all of them are still necessary

Author's Address

Jason Livingood
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US

Email: jason_livingood@cable.comcast.com
URI: <http://www.comcast.com>

IPv6 Operations
Internet-Draft
Intended status: Informational
Expires: January 9, 2012

J. Brzozowski
C. Griffiths
Comcast
July 8, 2011

Comcast IPv6 Trial/Deployment Experiences
draft-jjmb-v6ops-comcast-ipv6-experiences-01

Abstract

This document outlines the various technologies Comcast has trialed as part of the company's ongoing IPv6 initiatives. The focus here are the technologies and experiences specific to enabling IPv6 for subscriber services like high speed data or Internet. Comcast has learned a great deal about various technologies that we feel are important to share with the community.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 9, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

Table of Contents

1. Requirements Language	3
2. Introduction	3
3. 6to4	3
4. 6RD	5
5. Native Dual Stack	6
6. Dual Stack Lite	8
7. Content and Services	8
8. Backoffice	9
9. Conclusion	9
10. IANA Considerations	10
11. Security Considerations	10
12. Acknowledgements	10
13. Normative References	10
Appendix A. Document Change Log	10
Authors' Addresses	11

1. Requirements Language

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

2. Introduction

Beginning in early 2010 Comcast announced plans to leverage the work the company has been doing related to IPv6 to conduct a number of IPv6 technology trials. These trials were specifically aimed at enabling IPv6 for subscriber services. The purpose of this document is to outline the technologies that have been trialed thus far along with experiences and observations that adopters of the same may find valuable in their own planning and deployment processes.

Further, there may be some additional feedback that the various groups within the IETF may wish to take into account as part of ongoing standards efforts.

3. 6to4

During production deployment planning the widespread use of 6to4 [RFC3068] to access content and services over IPv6 was assessed. In some scenarios 6to4 usage increased several hundred times. At the time Comcast had not deployed its own 6to4 relay infrastructure as such open relays being operated by independent third parties were by default used to facilitate 6to4-based communications. The deployment and default use of open 6to4 relays appears to be a key variable behind the sub-optimal performance associated with the use of 6to4. Operators that have not deployed IPv6 or have IPv6 incapable infrastructures should note that the use of 6to4 is likely occurring today across their infrastructure. Many operating systems and home networking devices continue to support the same and in some cases have 6to4 and other transition technologies enabled by default.

As a community there appears to be some consensus that long term the use of 6to4 is not desirable, however, in the near term it is clear that 6to4 will be used in specific scenarios. The expectation and goal is to see 6to4 usage diminish over time until use of the same is displaced by an alternate technique to access content and services over IPv6. While the debate continues over how and when to deprecate 6to4, it is clear that 6to4 should not be recommended as a primary mechanism to access content and services over IPv6.

The following documents outline the recommendations surrounding the

use and status of the 6to4 from a standards point of view:

1. [draft-ietf-v6ops-6to4-advisory]
2. [draft-ietf-v6ops-6to4-to-historic]

Comcast deployed a series of five (5) 6to4 relays in a geographically dispersed configuration across our network. The purpose of these relays was to reduce the latency typically associated with 6to4 usage. During our analysis, the use of off network, open 6to4 relays was determined to yield nearly unusable conditions depending on the geographic location of the end user relative to the open 6to4 relay. By deploying on-network 6to4 relays, latency in most cases was reduced by over 50%, which instantly yielded considerable improvements from an end user point of view. The simplistic design and deployment of these relays enabled us to rapidly put them in network, and in some cases create a better experience for some of our users who had 6to4 enabled.

Through the use of commodity i386 based servers that run a standard Linux Operating System, we reduced deployment and operating costs, while still maintaining a fault tolerant design. Each 6to4 relay was dual stacked, and with a simple kernel module, we enabled the 6to4 configuration. Some 6to4 specific configurations were required to ensure compatibility across a wide range of end points. The logic to Anycast the 6to4 records was handled by the network infrastructure providing connectivity to the 6to4 relays, and health check enabled us to automatically remove the route for any relay from the routing table in case of failure.

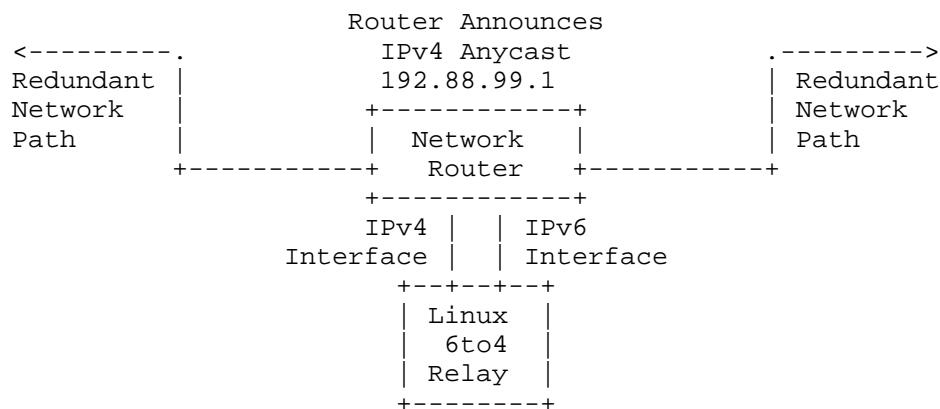


Figure 1: Comcast 6to4 Data Center View

4. 6RD

6RD [draft-townsley-ipv6-6rd] is another transition technology similar to 6to4 that Comcast has deployed as part of technology trials. While 6RD shared many similarities with 6to4 technologically there were a number of differences noted with the same that adopters of the same should consider as part of their own deployments.

As advertised 6RD frees adopters from some restrictions typically associated with 6to4 namely the use of anycast addressing (IPv4 and IPv6) and the infrastructure, like 6to4, is straightforward to deploy. However, at the time of deployment it was observed that a limited number of border relay (BR) implementations were available. This appears to be an evolving area with more implementations becoming available. Similarly it was observed that there were few if any customer edge (CE) implementations available to support a trial of the technology. As such engineering implementations were leveraged to evaluate 6RD. Further, there were no implementations available that supported the 6RD DHCPv4 options [draft-ietf-softwire-ipv6-6rd] as such every 6RD CE used for trial was manually configured with the necessary configuration required to enable 6RD. In order to support a wide scale production deployment leveraging 6RD an operator would have to ensure their DHCP infrastructure supports the required 6RD DHCPv4 options along with targeted 6RD CE devices.

Trial configurations included two (2) 6RD BRs, which were intentionally deployed in geographically dispersed configuration. An anycast design was used to enable 6RD with a well known IPv4 anycast address and FQDN for the 6RD BR. The use of the same eased manual configuration and deployment. Additionally, an IPv6 /32 was used to support the 6RD trials as such subscriber devices were able to yield a usable IPv6 /64 on the LAN side of the 6RD CE.

The quantity and location of the 6RD BRs is a key variable when planning the deployment of 6RD. Comcast specifically deployed a limited quantity of the same resulting in some end users being "closer" to the BRs than others. Proximity to the 6RD BRs is an important factor that impacts the end user experience. While 6RD yields some improvements over 6to4, 6RD is ultimately a tunneling technology as such use of the same is subject to the challenges faced by other tunneling technologies.

Placement and quantity of 6RD BRs is also a significant variable to consider when assessing impacts to performance and IPv6 geo-location.

The following provides an overview of the Comcast 6RD trial network design:

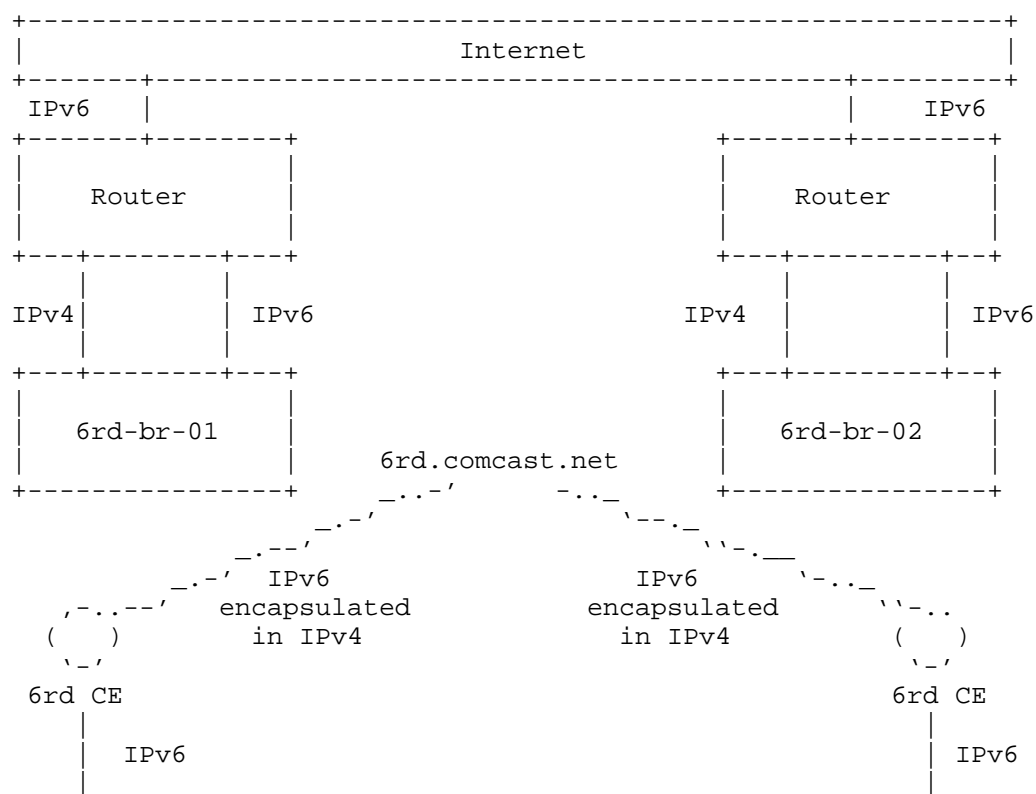


Figure 2: Comcast 6RD Overview

5. Native Dual Stack

Native dual stack is central to Comcast's IPv6 program for trial and production deployment. Native dual stack is the model where IPv4 services remain as-is with native IPv6 support introduced in parallel or simultaneously. Many of the details surrounding how this is

achieved are documented as part of the CableLabs Data Over Cable Service Interface Specification (DOCSIS) 3.0 [DOCSIS3.0]. However, relevant trial and deployment specific information that is of interest to the IETF community will be documented.

Native dual stack trials depend on the upgrade and enablement of Cable Modem Termination Systems [CMTS] to support IPv6. A CMTS is a device that end users in a cable network connect directly to using their cable modem [CM]. As with IPv4, native support for IPv6 is critical for the delivery of services to end users in a DOCSIS network. Anything less could yield an undesirable end user experience or instability in the operator network that could adversely impact larger populations of users.

Given the CMTS requirements, native dual stack trials have initially been limited to specific areas of the network. Further, where CMTS platforms have been upgraded and enabled to support IPv6 end users have been incrementally enabled with support for IPv6. Again this is to ensure a controlled introduction with a specific focus on maintaining stability. Initially, a limited combination of cable modem and IGD devices are being used to support trial activities. Overtime diversity for both cable modem and IGDs are expected to expand. To date a number of cable modems support the ability to enable native dual stack connectivity to CPEs devices behind the same. A subset of pre-DOCSIS 3.0 and all DOCSIS 3.0 devices support this capability. The population of DOCSIS devices that support these capabilities varies from operator to operator.

Trial enablement requires the stateful provisioning of an IGD using stateful DHCPv6 [RFC3315] for the IGD WAN interface and delegated prefixes [RFC3633] for LAN side connectivity. The quantity of devices supporting a native dual stack mode of operation is growing. While some devices are upgradable to support native dual stack many devices deployed today are not upgradable to support this functionality. Early implementations of devices or devices that are upgradable to support native IPv6 were found to only require and/or support the use of an IPv6 /64 for LAN side connectivity. This has been an acceptable mode of operation, however, over time IGDs will be required to support more advanced functionality including the ability to support multiple, routed IPv6 LANs. While support for a single IPv6 /64 is in place today support for shorter IPv6 prefixes is also supported. It is important for operators to ensure they design and plan support across their infrastructures for delegated prefixes that are shorter than /64.

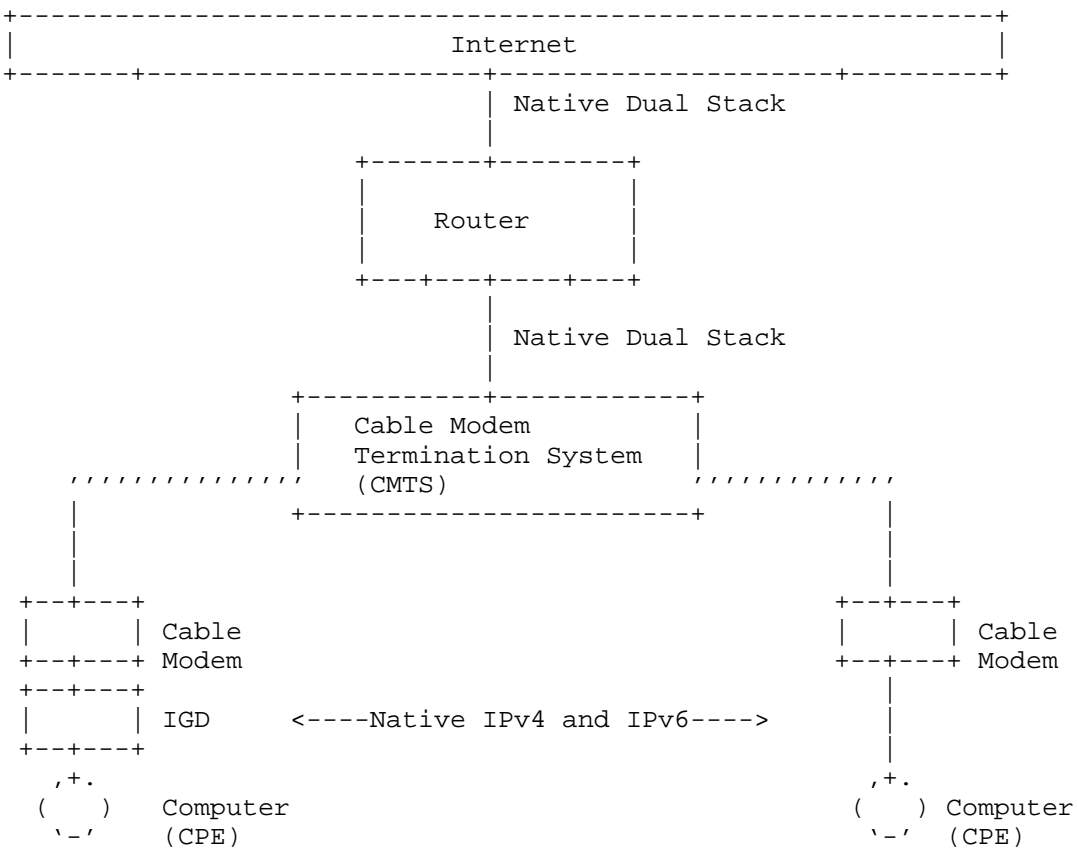


Figure 3: Comcast Native Dual Stack

6. Dual Stack Lite

Part of Comcast's trial plans includes the trialing of Dual Stack Lite. At this time trial planning for the same is underway. While Comcast plans on trialing Dual Stack Lite there are no plans at this time to deploy Dual Stack Lite beyond a limited technology trial.

7. Content and Services

During early phases of our trials Comcast leveraged reverse proxies to expedite the availability of content natively over IPv6. Open source technology running on Linux based servers was used enable the reverse proxies. To ensure that the origin content, which is IPv4 only, is available natively over IPv6 the proxy servers required

native dual stack connectivity. This model allowed us to ensure that Internet facing access to Comcast occurred natively over IPv6.

As third party CDNs introduce production quality support for IPv6 we plan to move away from the use of proxy servers and fully towards native dual stack for Comcast content and services. Native dual stack content is but the first step to ensure the same can be IPv6 only at some point in the future. As observed during World IPv6 Day it is still somewhat premature to have IPv6 only content.

Further as part of our trials Comcast has also recently enabled, in a limited fashion, Message Transfer Agents (MTA) to allow a subset of Comcast trial users to send electronic mail using SMTP. Due to the limited availability of spam mitigation for IPv6 Comcast trials does not include the receipt of electronic mail over IPv6. In order to enable the receipt of electronic mail over IPv6 spam mitigation must be in place.

8. Backoffice

We made the decision early on in our design discussions to move all systems to a dual-stack since we felt that this was the best way to transition to IPv6. We have been planning for several years to re-architect many core systems like DNS, DHCP, OSS/BSS, and Billing systems. This approach has paid off and allowed us to rapidly move towards support for dual-stack at the edge of our network, including support for our customers devices.

9. Conclusion

To date Comcast trial activities have yielded important, useful information about the various technologies that are available to facilitate the transition to IPv6. Observations and experience to date confirms that native dual stack is the preferred approach to transition to IPv6, where possible. While the various tunneling technologies are indeed straightforward to deploy there are a number of variables that must be considered when planning to deploy the same.

Support for native dual stack continues to evolve across various broadband technologies and within consumer electronics. As evidenced by World IPv6 Day many of the world's largest content providers are also making progress with their IPv6 capabilities.

10. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

11. Security Considerations

There are no security considerations at this time.

12. Acknowledgements

Thanks to the Comcast team supporting the various trial and production deployment activities:

Jonathan Boyer

Chris Griffiths

Tom Klieber

Yiu Lee

Jason Livingood

Anthony Veiga

Joel Warburton

Richard Woundy

13. Normative References

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

Appendix A. Document Change Log

[RFC Editor: This section is to be removed before publication]

-01: Added C. Griffiths as co-author. Currently working on ascii art and several new sections.

-00: First version published.

Authors' Addresses

John Jason Brzozowski
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US

Email: john_brzozowski@cable.comcast.com
URI: <http://www.comcast.com>

Chris Griffiths
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US

Email: chris_griffiths@cable.comcast.com
URI: <http://www.comcast.com>

IPv6 Operations
Internet-Draft
Intended status: Informational
Expires: April 4, 2012

J. Brzozowski
C. Griffiths
Comcast
October 2, 2011

Comcast IPv6 Trial/Deployment Experiences
draft-jjmb-v6ops-comcast-ipv6-experiences-02

Abstract

This document outlines the various technologies Comcast has trialed as part of the company's ongoing IPv6 initiatives. The focus here are the technologies and experiences specific to enabling IPv6 for subscriber services like high speed data or Internet. Comcast has learned a great deal about various technologies that we feel are important to share with the community.

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 4, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

described in the Simplified BSD License.

Table of Contents

1. Requirements Language	3
2. Introduction	3
3. 6to4	3
4. 6RD	5
5. Native Dual Stack	7
6. Dual Stack Lite	8
7. Content and Services	9
8. Backoffice	9
9. World IPv6 Day	9
10. Conclusion	10
11. IANA Considerations	10
12. Security Considerations	10
13. Acknowledgements	10
14. Normative References	11
Appendix A. Document Change Log	11
Authors' Addresses	11

1. Requirements Language

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

2. Introduction

Beginning in early 2010 Comcast announced plans to leverage the work the company has been doing related to IPv6 to conduct a number of IPv6 technology trials. These trials were specifically aimed at enabling IPv6 for subscriber services. The purpose of this document is to outline the technologies that have been trialed thus far along with experiences and observations that adopters of the same may find valuable in their own planning and deployment processes.

Further, there may be some additional feedback that the various groups within the IETF may wish to take into account as part of ongoing standards efforts.

3. 6to4

During production deployment planning the widespread use of 6to4 [RFC3068] to access content and services over IPv6 was assessed. In some scenarios 6to4 usage increased several hundred times. At the time Comcast had not deployed its own 6to4 relay infrastructure as such open relays being operated by independent third parties were by default used to facilitate 6to4-based communications. The deployment and default use of open 6to4 relays appears to be a key variable behind the sub-optimal performance associated with the use of 6to4. An important thing to note is that some home gateway vendors have turned on 6to4 by default, and in some of these implementations, they have not presented a user interface a user interface to disable it. For operators that have not deployed IPv6 or have IPv6 incapable infrastructures should note that the use of 6to4 is likely occurring today across their infrastructure. Many operating systems and home networking devices continue to support 6to4 and in some cases have 6to4 and other transition technologies enabled by default.

As a community there appears to be some consensus that long term the use of 6to4 is not desirable, however, in the near term it is clear that 6to4 will be used in specific scenarios. The expectation and goal is to see 6to4 usage diminish over time until use of the same is displaced by an alternate technique to access content and services over IPv6. While the debate continues over how and when to deprecate 6to4, it is clear that 6to4 should not be recommended as a primary

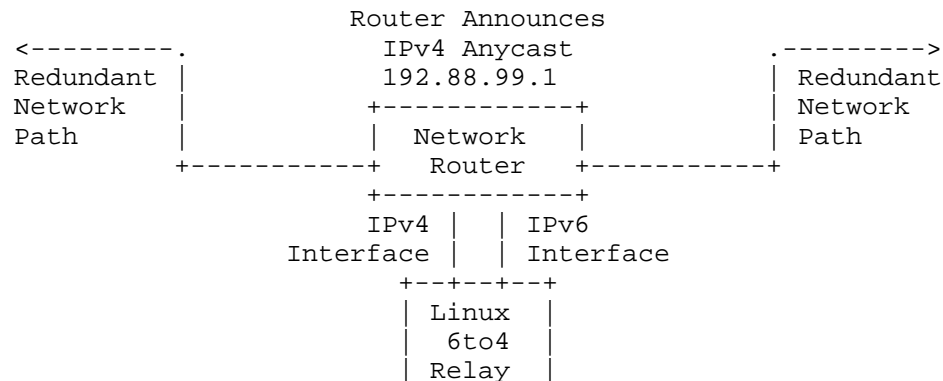
mechanism to access content and services over IPv6.

The following documents outline the recommendations surrounding the use and status of 6to4 from a standards point of view:

1. [draft-ietf-v6ops-6to4-advisory]
2. [draft-ietf-v6ops-6to4-to-historic]

Comcast deployed a series of five (5) 6to4 relays in a geographically dispersed configuration across our network. The purpose of these relays was to reduce the latency typically associated with 6to4 usage. During our analysis, the use of off network, open 6to4 relays was determined to yield nearly unusable conditions depending on the geographic location of the end user relative to the open 6to4 relay. By deploying on-network 6to4 relays, latency in most cases was reduced by over 50%, which instantly yielded considerable improvements from an end user point of view. The simplistic design and deployment of these relays enabled us to rapidly put them in network, and in some cases create a better experience for some of our users who had 6to4 enabled.

Through the use of commodity x86 based servers that run a standard Linux Operating System, we reduced deployment and operating costs, while still maintaining a fault tolerant design. Each 6to4 relay was dual stacked, and with a simple kernel module, we enabled the 6to4 configuration. Some 6to4 specific configurations were required to ensure compatibility across a wide range of end points. The logic to anycast the 6to4 records was handled by the network infrastructure providing connectivity to the 6to4 relays, and health checking enabled us to automatically remove the route for any relay from the routing table in case of failure.



+-----+

Figure 1: Comcast 6to4 Data Center View

4. 6RD

6RD [draft-townsley-ipv6-6rd] is another transition technology similar to 6to4 that Comcast has deployed as part of technology trials. While 6RD yields some improvements over 6to4, 6RD is ultimately a tunneling technology. As such, it is subject to the challenges faced by other tunneling technologies.

As advertised, 6RD frees adopters from some restrictions typically associated with 6to4. The use of anycast addressing (IPv4 and IPv6) is no longer required and the infrastructure, like 6to4, is straightforward to deploy. However, at the time of deployment it was observed that a limited number of border relay (BR) implementations were available. This appears to be an evolving area with more implementations becoming available. Similarly it was observed that there were few if any customer edge (CE) implementations available to support a trial of the technology. As such engineering implementations were leveraged to evaluate 6RD. Further, there were no implementations available that supported the 6RD DHCPv4 options [draft-ietf-softwire-ipv6-6rd]. Because of this, every 6RD CE used for trial was manually configured with the necessary information required to enable 6RD. In order to support a wide scale production deployment leveraging 6RD an operator would have to ensure their DHCP infrastructure supports the required 6RD DHCPv4 options along with targeted 6RD CE devices.

Trial configurations included two (2) 6RD BRs, which were intentionally deployed in geographically dispersed configuration. An anycast design was used to enable 6RD with a well known IPv4 anycast address and FQDN for the 6RD BR. The use of anycast eased manual configuration and deployment. Additionally, an IPv6 /32 was used to support the 6RD trials permitting subscriber devices were able to yield a usable IPv6 /64 on the LAN side of the 6RD CE.

The quantity and location of the 6RD BRs is a key variable when planning the deployment of 6RD. Comcast specifically deployed a limited quantity of BRs resulting in some end users being "closer" to the BRs than others. Proximity to the 6RD BRs is an important factor that impacts the end user experience. While 6RD yields some improvements over 6to4, 6RD is ultimately a tunneling technology as

such use of the same is subject to the challenges faced by other tunneling technologies.

Placement and quantity of 6RD BRs is also a significant variable to consider when assessing impacts to performance and IPv6 geo-location. A centralized approach to deploying 6RD BRs will yield undesirable impacts to IPv6 geo-location in that end users leveraging a particular 6RD BR that is geographically distant from their true location will not accurately represent the origin of the end user request. Conversely, deploying 6RD BRs that are near end users may require a substantial quantity of 6RD BRs depending on the operator network.

The following provides an overview of the Comcast 6RD trial network design:

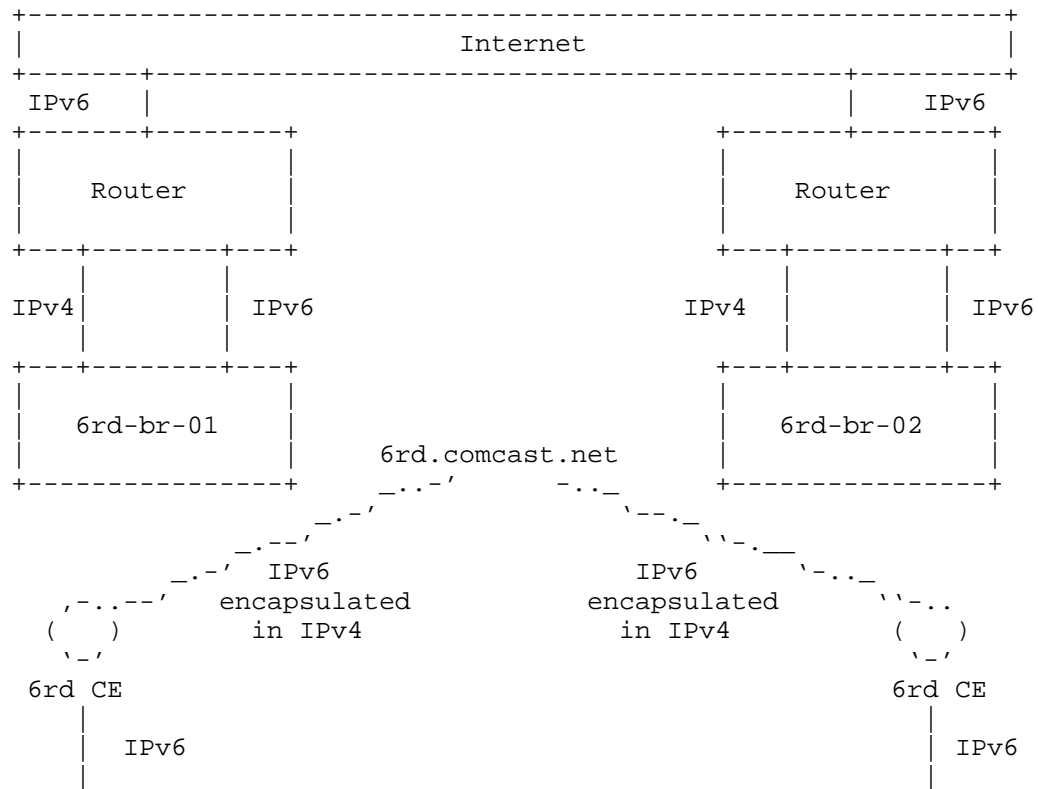


Figure 2: Comcast 6RD Overview

5. Native Dual Stack

Native dual stack is central to Comcast's IPv6 program for trial and production deployment. Native dual stack is the model where IPv4 services remain as-is with native IPv6 support introduced in parallel or simultaneously. Many of the details surrounding how this is achieved are documented as part of the CableLabs Data Over Cable Service Interface Specification (DOCSIS) 3.0 [DOCSIS3.0]. However, relevant trial and deployment specific information that is of interest to the IETF community will be documented.

Native dual stack trials depend on the upgrade and enablement of Cable Modem Termination Systems [CMTS] to support IPv6. A CMTS is a device that end users in a cable network connect directly to using their cable modem [CM]. As with IPv4, native support for IPv6 is critical for the delivery of services to end users in a DOCSIS network. Anything less could yield an undesirable end user experience or instability in the operator network that could adversely impact larger populations of users.

Given the CMTS requirements, native dual stack trials have initially been limited to specific areas of the network. Further, where CMTS platforms have been upgraded and enabled to support IPv6 end users have been incrementally enabled with support for IPv6. Again this is to ensure a controlled introduction with a specific focus on maintaining stability. Initially, a limited combination of cable modem and IGD devices are being used to support trial activities. Over time diversity for both cable modem and IGDs are expected to grow. To date a number of cable modems support the ability to enable native dual stack connectivity to CPEs devices behind them. A subset of pre-DOCSIS 3.0 and all DOCSIS 3.0 devices support this capability. The population of DOCSIS devices that support these capabilities varies from operator to operator.

Trial enablement requires the stateful provisioning of an IGD using stateful DHCPv6 [RFC3315] for the IGD WAN interface and delegated prefixes [RFC3633] for LAN side connectivity. Similarly, trial supported direct attachment of IPv6 capable CPE devices to the CM. In this configuration the CPE is provisioned with one or more IPv6 addresses via stateful DHCPv6 [RFC3315] in similar fashion to the IGD WAN interface. The quantity of devices supporting a native dual stack mode of operation is growing. While some devices are upgradable to support native dual stack many devices deployed today are not upgradable to support this functionality. Early implementations of devices or devices that are upgradable to support native IPv6 were found to only require and/or support the use of an IPv6 /64 for LAN side connectivity. This has been an acceptable mode of operation, however, over time IGDs will be required to support

more advanced functionality including the ability to support multiple, routed IPv6 LANs. While support for a single IPv6 /64 is in place today support for shorter IPv6 prefixes is also supported. It is important for operators to ensure they design and plan support across their infrastructures for delegated prefixes that are shorter than /64.

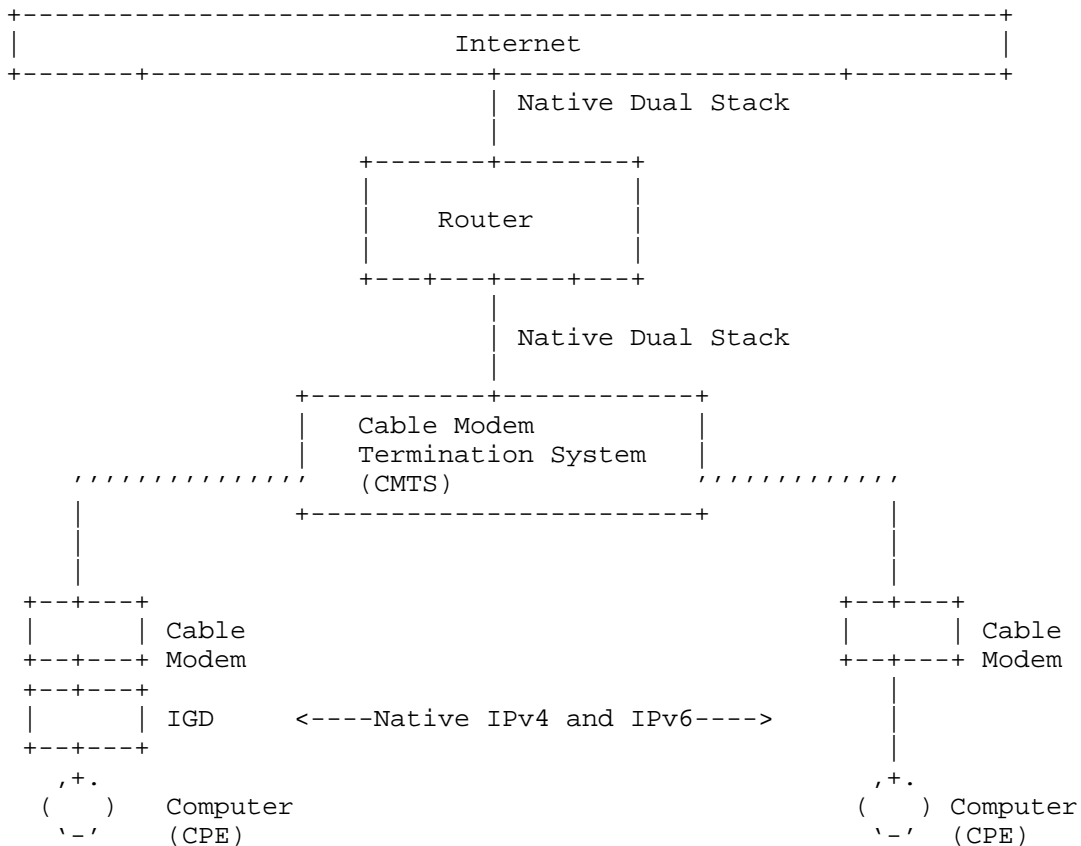


Figure 3: Comcast Native Dual Stack

6. Dual Stack Lite

Part of Comcast's trial plans includes the trialing of Dual Stack Lite. At this time trial planning for the same is underway. While Comcast plans on trialing Dual Stack Lite there are no plans at this

time to deploy Dual Stack Lite beyond a limited technology trial.

7. Content and Services

During early phases of our trials Comcast leveraged reverse proxies to expedite the availability of content natively over IPv6. Open source technology running on Linux based servers was used to enable the reverse proxies. To ensure that the origin content, which is IPv4 only, is available natively over IPv6 the proxy servers required native dual stack connectivity. This model allowed us to ensure that Internet facing access to Comcast content occurred natively over IPv6.

As third party CDNs introduce production quality support for IPv6 we plan to move away from the use of proxy servers and fully towards native dual stack for Comcast content and services. Native dual stack content is but the first step to ensure the same can be IPv6 only at some point in the future. Observations from Comcast's participation in World IPv6 day suggest it is premature to rely on IPv6-only content at this time

Further as part of our trials Comcast has also recently enabled IPv6 Message Transfer Agents (MTA), in a limited fashion, to allow a subset of Comcast trial users to send electronic mail using SMTP over IPv6.. Due to the limited availability of spam mitigation for IPv6 Comcast trials does not include the receipt of electronic mail over IPv6. In order to enable the receipt of electronic mail over IPv6 spam mitigation must be in place.

8. Backoffice

We made the decision early on in our design discussions to move all systems to a dual-stack design since we felt that this was the best way to transition to IPv6. The re-architect of many core systems like DNS, DHCP, OSS/BSS, and Billing systems took many years to plan and complete and this approach has paid off and allowed us to rapidly move towards support for dual-stack at the edge of our network, including support for our customers devices.

9. World IPv6 Day

During World IPv6 day, Comcast observed a significant increase in native IPv6 traffic once content providers enabled AAAA records for their websites. The resulting traffic has continued to increase even after World IPv6 when about 50% of the websites that participated in

World IPv6 Day left their AAAA records enabled after the day. We view this as a positive sign for continuing to drive more IPv6 traffic.

10. Conclusion

To date Comcast trial activities have yielded important, useful information about the various technologies that are available to facilitate the transition to IPv6. Observations and experience to date confirms that native dual stack is the preferred approach to transition to IPv6, where possible. While the various tunneling technologies are indeed straightforward to deploy there are a number of variables that must be considered when planning to deploy the same.

Support for native dual stack continues to evolve across various broadband technologies and within consumer electronics. As evidenced by World IPv6 Day many of the world's largest content providers are also making progress with their IPv6 capabilities.

11. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

12. Security Considerations

There are no security considerations at this time.

13. Acknowledgements

Thanks to the Comcast team supporting the various trial and production deployment activities:

Jonathan Boyer

Chris Griffiths

Tom Klieber

Yiu Lee

Jason Livingood

Anthony Veiga

Joel Warburton

Richard Woundy

14. Normative References

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

Appendix A. Document Change Log

[RFC Editor: This section is to be removed before publication]

-02: Grammatical items and re-wording of some sections. We have also added a new World IPv6 Day section.

-01: Added C. Griffiths as co-author. Currently working on ascii art and several new sections.

-00: First version published.

Authors' Addresses

John Jason Brzozowski
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US

Email: john_brzozowski@cable.comcast.com
URI: <http://www.comcast.com>

Chris Griffiths
Comcast Cable Communications
One Comcast Center
1701 John F. Kennedy Boulevard
Philadelphia, PA 19103
US

Email: chris_griffiths@comcast.com
URI: <http://www.comcast.com>

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 12, 2012

A. Keranen
J. Arkko
Ericsson
July 11, 2011

Some Measurements on World IPv6 Day from End-User Perspective
draft-keranen-ipv6day-measurements-01

Abstract

During the World IPv6 Day on June 8th, 2011, several key content providers enabled their networks to offer both IPv4 and IPv6 service. Hundreds of organizations participated in this effort, and in the months and weeks leading up to the event worked hard on preparing their networks to support this event. The event was largely unnoticed by the general public, which is a good thing as no major problems were detected. For the Internet, however, there was a major change on such a small timescale. This memo discusses measurements that the authors made from the perspective of an end-user with well-working IPv4 and IPv6 connectivity. Our measurements include the number of most popular networks providing AAAA records for their service as well as delay and connection failure statistics.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

Table of Contents

1. Introduction	3
2. Motivation and Goals	3
3. Measurement Methodology	4
4. Measurement Results	5
4.1. DNS AAAA Records	5
4.2. TCP Connection Setup	6
4.3. TCP Connection Delays	7
5. Conclusions	8
6. Security Considerations	9
7. IANA Considerations	9
8. Informative References	9
Appendix A. Acknowledgments	10
Authors' Addresses	10

1. Introduction

Many large content providers participated in World IPv6 Day on June 8, 2011. On that day, IPv6 [RFC2460] was enabled by default for 24 hours on numerous networks and sites that previously supported only IPv4. The aim was to identify any remaining issues with widespread IPv6 usage in these networks. Most of the potential problems associated with using IPv6 are, after all, of a practical nature, such as: ensuring that the necessary components have IPv6 turned on; that configurations are correct; and that any implementation bugs have been removed.

Some content providers have been reluctant to enable IPv6. The reasons for this include delays for applications attempting to connect over broken IPv6 links before falling back to IPv4, and unreliable IPv6 connectivity. Bad IPv6 routing has been behind many of the problems. Among the causes are broken 6to4 tunneling protocol connectivity, experimental IPv6 setups that are untested and unmonitored, and configuration problems with firewalls. The situation is improving as more users and operators put IPv6 to use and fix the problems that emerge.

World IPv6 Day event was largely unnoticed by the general public, which is a good thing as no major problems were detected. For the Internet, however, there was a major change on such a small timescale. This memo discusses measurements that the authors made from the perspective of an end-user with well-working IPv4 and IPv6 connectivity. Our measurements include the number of most popular networks providing AAAA records for their service as well as delay and connection failure statistics.

The rest of this memo is structured as follows. Section 2 discusses the goals of our measurements, Section 3 describes our measurement methodology, Section 4 gives our preliminary results, and Section 5 makes some conclusions.

2. Motivation and Goals

Practical IPv6 deployment plans benefit from accurate information about the extent to which IPv6 can be used for communication, and how its characteristics differ from those of IPv4. For instance, operators planning to deploy dual-stack networking may wish to understand what fraction of their traffic would move to IPv6. This information is useful for estimating sufficient capacity to deal with the IPv6 traffic and impacts to the operator's IPv4 infrastructure or carrier-grade NAT devices as their traffic is reduced. Network owners also wish to understand the extent to which they can expect

different delay characteristics or problems with IPv6 connectivity. The goals of our measurements were to help with these topics by answering the following questions:

- o What fraction of most popular Internet sites offer AAAA records? How did the World IPv6 Day change the situation?
- o How do the traffic characteristics differ between IPv4 and IPv6 on sites offering AAAA records? Are the connection failure rates similar? How are RTTs impacted?

There have been many measurements about some of these aspects from a service provider perspective, such as the Google studies on which end users have broken connectivity towards them. Our measurements start from a different angle, by assuming a well-working dual-stack connectivity on the measurement end, and then probing the rest of the Internet to understand, for instance, how likely it is to have IPv6 connectivity problems, or what are the delay differences between IPv4 and IPv6 towards the rest of the Internet. Similar studies have been performed by the Comcast IPv6 Adoption Monitor [IPv6Monitor] and RIPE NCC [RIPEv6Day].

3. Measurement Methodology

We used the top 10,000 sites of the Alexa 1 million most popular sites list [Alexa] from June 1st 2011. For each domain name in the list, we performed DNS queries with different host names. For IPv4 addresses (A records) we used host name "www" and also performed a query with just the domain name. For IPv6 addresses (AAAA records) we used also different combinations of host names that have been used for IPv6 sites, namely "www6", "ipv6", "v6", "ipv6.www", "www.ipv6", "v6.www", and "www.v6".

All DNS queries were initiated in the order listed above (first "www" and just the domain name for A-records, then "www", domain name, and different IPv6-host names for AAAA records) but the queries were done in parallel (i.e., without waiting for the previous query to finish). The first response for A and AAAA record and the corresponding host name were recorded. The queries had 3 second re-transmission timeout and if there wasn't any response for 10 seconds, all remaining queries for the site were canceled. We used a custom-made Perl script and the Net::DNS module for the DNS queries.

The measurement script used a bind9 DNS server running on the same host that was performing the measurement. The DNS cache of the server was flushed before each measurement run to be able to detect the changes in the DNS records in real-time. The host, and thus the

DNS server, was not part of DNS IPv6 whitelisting agreements.

After obtaining IP addresses for the site, if a site had both A and AAAA records, a simple C program was used to create TCP connections to the port 80 (HTTP) at the same time with IPv4 and IPv6 to the (first) IP addresses discovered from the DNS. The connection setup was repeated up to 10 times, giving up after the first failed attempt (but only after normal TCP re-transmissions). The connection setup delay was measured by recording the time right before and after the connect system call. The host used for measurements is a regular Linux PC with 2.6.32 version kernel and dual-stack Internet connection via Ethernet.

The measurements were started one week before the World IPv6 Day (on Wednesday, June 1st, 17:30 UTC) and have been running since, once every three hours. One test run takes from two to two and a half hours to finish.

The accuracy and generality of the measurement results is limited by several factors. While we run the tests in three different sites, most of the results discussed in this document present snapshots of the situation from just one measurement point, the Ericsson Research Finland premises. Also, since one measurement run takes considerably long time, the network characteristics and DNS records may have changed even during a single run. The first DNS response was used for the TCP connectivity tests and this selection may result in selecting un-optimal host; yet, slight preference is given to the "www" and only-domain-name records since their queries were started before the others. While the host performing the measurements was otherwise idle, the local network was in regular office use during the measurements. The connectivity setup delay is collected in user space, with regular, non real-time, kernel implementation, resulting in small inaccuracies in the timing information.

4. Measurement Results

4.1. DNS AAAA Records

The amount of top 10,000 sites with AAAA DNS records before, during, and after the World IPv6 Day, is shown in <http://users.piuha.net/akeranen/drafts/v6day/v6sites.pdf>. The measurements performed during the World IPv6 Day are shown on the light gray background.

When the measurements began on June 1st, there were 245 sites (2.45%) with both A and AAAA record. During the following days the number of sites was slowly increasing, reaching 306 sites at the measurement

that was started 22:30 UTC on June 7th, the evening before the World IPv6 Day. When the World IPv6 Day officially started, the following measurement (1:30 UTC) recorded already 383 sites, and the next one 472 sites. During the day number of sites with AAAA records peaked at 491 (4.91% of the measured 10,000 sites) after 19:30 UTC.

When the World IPv6 Day was over, also the number of AAAA records dropped nearly as fast as it had increased just 24 hours earlier. However, the number of sites stabilized around 310 and has not dropped below 300 since, resulting in over 3% of the top 10,000 sites having AAAA records today.

While 274 sites had IPv6 enabled in their DNS for some of the tested host names one day before the World IPv6 Day, only 116 had it for the "www" host name that is commonly used when accessing a web site. The number of "www" host names with AAAA records more than tripled during the World IPv6 Day reaching 374 sites for 3 consecutive measurement runs (i.e., at least for 6 hours). Also the number of AAAA records for the "www" host name dropped steeply after the day and has remained around 160 sites since.

Similar but more pronounced trends can be seen if only top 100 of the most popular sites are taken into considerations, as show in <http://users.piuha.net/akeranen/drafts/v6day/v6sites-top100.pdf>. Here, the number of sites with some of the tested host names having AAAA record was initially 14, jumped to 36 during the day, and eventually dropped to 13. Also, while none of the top 100 sites apparently had AAAA record for their "www" host name before and after the World IPv6 day, during the day the number peaked at 30. Thus, roughly one third of the 100 most popular sites was enabling IPv6 for the World IPv6 Day.

Two other test sites in Sweden and Canada experienced similar trends with the DNS records. However, one of the sites used an external DNS server that was part of whitelisting agreements. There the amount of sites with AAAA records before the World IPv6 Day was already higher (above 400) and hence the impact of the day was smaller when the amount of sites increased to same numbers as seen by the test site in Finland. With the whitelisted DNS server the level of sites remained above 450 also after the day.

4.2. TCP Connection Setup

To test whether the IP addresses given by the DNS actually provide connectivity to the web site, and if there is any difference in the connection setup delay and failure rates with IPv4 and IPv6, we attempted to create TCP connections for all domains that contained both A and AAAA DNS records. The fraction of sites for which the

first DNS response gave addresses that were not accessible with TCP to port 80 over IPv4 or IPv6 is shown in <http://users.piuha.net/akeranen/drafts/v6day/tcp-fails.pdf>.

There is a baseline failure rate with IPv4 around 1-3% that is fairly static throughout the test period. For hosts with AAAA records, the fraction of inaccessible sites was much higher: in the beginning up to one fourth of the tested hosts did not respond to TCP connection attempts. Much of this was likely due to the various test sites with different "IPv6 prefixes" (as discussed in Section 3); in the first run more than half of the tested sites with AAAA records used them for the first DNS response. Also, some of the hosts may not even be supposed to be accessed with HTTP but provide AAAA records for other purposes and some sites had clear configuration errors, such as localhost or link-local IPv6 addresses.

As the World IPv6 Day came closer, the number of inaccessible IPv6 sites decreased slowly and the number of sites with AAAA records increased at the same time, resulting in failure ratio dropping to roughly 20% before the day. During the day number of IPv6 sites increased rapidly but also the number of failures decreased and hence, at the end of the day, the failure ratio dropped to just above 10%. After the World IPv6 Day when many of the participating IPv6 hosts were taken off-line, the fraction of failed sites for IPv6 increased. However, since there was no increase in the absolute number of failed sites, the fraction of inaccessible sites remained at lower level, between 15 and 20 percentage, than before the day.

4.3. TCP Connection Delays

For sites that were accessible with both IPv4 and IPv6, we measured the time difference it takes to establish a TCP connection with IPv4 and IPv6. We took the median (as defined in Section 11.3 of [RFC2330]) of the time differences of all 10 connections, and then median and average (of the median) over all sites; the result is shown in <http://users.piuha.net/akeranen/drafts/v6day/mda.pdf>.

In general, the delay differences are small: median of medians stays less than 3ms off from being equal and even the mean, which is more sensitive to outliers, stays most of the time within +/- 5ms; with highest spikes reaching to -15ms (mean of median IPv6 delays being 15ms larger than for IPv4 delays). Closer inspection of the results shows that the spikes are often caused by only one or a handful of sites with bad connectivity and multiple re-transmissions of TCP SYN and ACK packets resulting in order of magnitude larger connection setup delays.

Surprisingly the median delay for IPv6 connections is in most of the

cases equal or smaller than IPv4 delay, but during the World IPv6 Day, the IPv6 delays increased slightly and became (on median) slower than IPv4 counterparts. One reason for such effect was that some of the sites that enabled IPv6 for the World IPv6 Day, had extremely low, less than ten millisecond, IPv4 delay (e.g., due to Content Delivery Network (CDN) provider hosting the IPv4 site), but "regular", over hundred millisecond, delay for the IPv6 host.

More detailed analysis of the TCP connection setup delay differences, and reasons behind them, is left for future work.

5. Conclusions

The World IPv6 Day had a very visible impact to the availability of content over IPv6, particularly when considering the top 100 content providers. It is difficult to find other examples of bigger one day swings in some characteristic of the Internet. However, real impacts to end users were small, given that when dual-stack works correctly it should not be visible at the user level and that IPv6 availability for end users themselves was small.

The key conclusions are as follows:

- o The day caused a large jump in the number of content providers providing AAAA DNS records on that day.
- o The day caused a smaller but apparently permanent increase in the number of content providers supporting AAAA.
- o Large and quick swings in the relative amount of IPv4 vs. IPv6 traffic are possible merely by supporting a dual-stack access network and having a few large content providers offer their service either globally or to this particular network over IPv6.
- o Large fraction of sites that published AAAA records for a name under their domain (be it "www" or "www6" or something else) were actually not responding to TCP SYN requests on IPv6. This fraction is far higher than what we've seen in our previous measurements, and we are still determining why that is the case. Measurement errors or problems on our side of the network cannot be ruled out at this stage. In any case, it is also clear that as new sites join, incomplete or in-progress configurations create more connectivity problems in the IPv6 Internet than we've seen before. Other measurements are needed to verify what the general level IPv6 connectivity is to addresses publicly listed in AAAA records.

- o Even if the overall level of connection failures was high, activities on and around the IPv6 day appear to have caused a significant permanent drop in the number of failures.
- o When IPv6 and IPv4 connectivity were available, the delay characteristics appear very similar. In other words, most of the providers that made IPv6 connectivity available appear to provide a production quality network. TCP connection setup delay differences due to RTT differences between IPv4 and IPv6 connections are in general low. In the remaining differences in our measurements, random packet loss plays a major role. However, some sites can experience considerable differences simply because of different content distribution mechanisms used for IPv4 and IPv6 content.

6. Security Considerations

Security issues have not been discussed in this memo.

7. IANA Considerations

This memo has no IANA implications.

8. Informative References

- [RFC2330] Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
- [IPv6Monitor] Comcast and University of Pennsylvania, "IPv6 Adoption Monitor", <<http://ipv6monitor.comcast.net>>.
- [RIPEv6Day] RIPE NCC, "World IPv6 Day Measurements", <<http://v6day.ripe.net/>>.
- [Alexa] Alexa the Web Information Company, "Alexa Top 1,000,000 Sites", <<http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>>.

Appendix A. Acknowledgments

The authors would like to thank Suresh Krishnan, Fredrik Garneij, Lorenzo Colitti, Jason Livingood, Alain Durand, Emile Aben, Jan Melen, and Tero Kauppinen for interesting discussions in this problem space.

Authors' Addresses

Ari Keranen
Ericsson
Jorvas 02420
Finland

Email: ari.keranen@ericsson.com

Jari Arkko
Ericsson
Jorvas 02420
Finland

Email: jari.arkko@piuha.net

Network Working Group
Internet-Draft
Intended status: Informational
Expires: March 15, 2013

A. Keranen
J. Arkko
Ericsson
September 11, 2012

Some Measurements on World IPv6 Day from End-User Perspective
draft-keranen-ipv6day-measurements-04

Abstract

During the World IPv6 Day on June 8th, 2011, several key content providers enabled their networks to offer both IPv4 and IPv6 services. Hundreds of organizations participated in this effort, and in the months and weeks leading up to the event worked hard on preparing their networks to support this event. The event was largely unnoticed by the general public, which is a good thing since it means that no major problems were detected. For the Internet, however, there was a major change on such a small timescale. This memo discusses measurements that the authors made from the perspective of an end-user with good IPv4 and IPv6 connectivity. Our measurements include the number of most popular networks providing AAAA records for their service as well as delay and connection failure statistics.

Status of this Memo

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

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

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

This Internet-Draft will expire on March 15, 2013.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
2. Motivation and Goals	3
3. Measurement Methodology	4
4. Measurement Results	5
4.1. DNS AAAA Records	5
4.2. TCP Connection Setup	7
4.3. TCP Connection Delays	7
5. Conclusions	8
6. Security Considerations	9
7. IANA Considerations	9
8. References	10
8.1. Normative References	10
8.2. Informative References	10
Appendix A. Acknowledgments	11
Authors' Addresses	11

1. Introduction

Many large content providers participated in World IPv6 Day on June 8, 2011. On that day, IPv6 [RFC2460] was enabled by default for 24 hours on numerous networks and sites that previously supported only IPv4. The aim was to identify any remaining issues with widespread IPv6 usage in these networks. Most of the potential problems associated with using IPv6 are, after all, of a practical nature, such as: ensuring that the necessary components have IPv6 turned on; that configurations are correct; and that any implementation bugs have been removed.

Some content providers have been reluctant to enable IPv6. The reasons for this include delays for applications attempting to connect over broken IPv6 links before falling back to IPv4 [RFC6555], and unreliable IPv6 connectivity. Bad IPv6 routing has been behind many of the problems. Among the causes are broken 6to4 tunneling protocol [RFC3056] connectivity, experimental IPv6 setups that are untested and unmonitored, and configuration problems with firewalls. The situation is improving as more users and operators put IPv6 to use and fix the problems that emerge.

World IPv6 Day event was largely unnoticed by the general public, which is a good thing since it means that no major problems were detected. Existing IPv4 connectivity was not damaged by IPv6 and also new IPv6 connectivity worked as expected in vast majority of cases. For the Internet, however, there was a major change on such a small timescale. This memo discusses measurements that the authors made from the perspective of an end-user with well-working IPv4 and IPv6 connectivity. Our measurements include the number of most popular networks providing AAAA records for their service as well as delay and connection failure statistics.

The rest of this memo is structured as follows. Section 2 discusses the goals of our measurements, Section 3 describes our measurement methodology, Section 4 gives our preliminary results, and Section 5 draws some conclusions.

2. Motivation and Goals

Practical IPv6 deployment plans benefit from accurate information about the extent to which IPv6 can be used for communication, and how its characteristics differ from those of IPv4. For instance, operators planning to deploy dual-stack networking may wish to understand what fraction of their traffic would move to IPv6. This information is useful for estimating the necessary capacity to deal with the IPv6 traffic and impacts to the operator's IPv4

infrastructure or carrier-grade NAT devices as their traffic is reduced. Network owners also wish to understand the extent to which they can expect different delay characteristics or problems with IPv6 connectivity. The goals of our measurements were to help with these topics by answering the following questions:

- o What fraction of most popular Internet sites offer AAAA records? How did the World IPv6 Day change the situation?
- o How do the traffic characteristics differ between IPv4 and IPv6 on sites offering AAAA records? Are the connection failure rates similar? How are RTTs impacted?

There have been many measurements about some of these aspects from a service provider perspective, such as the Google studies on which end users have broken connectivity towards them. Our measurements start from a different angle, by assuming good dual-stack connectivity at the measurement end, and then probing the rest of the Internet to understand, for instance, how likely there are to be IPv6 connectivity problems, or what the delay differences are between IPv4 and IPv6. Similar studies have been performed by the Comcast IPv6 Adoption Monitor [IPv6Monitor] and RIPE NCC [RIPEv6Day].

3. Measurement Methodology

We used the top 10,000 sites of the Alexa 1 million most popular sites list [Alexa] from June 1st 2011. For each domain name in the list, we performed DNS queries with different host names. For IPv4 addresses (A records) we used host name "www" and also performed a query with just the domain name. For IPv6 addresses (AAAA records) we used also different combinations of host names that have been used for IPv6 sites, namely "www6", "ipv6", "v6", "ipv6.www", "www.ipv6", "v6.www", and "www.v6".

All DNS queries were initiated in the order listed above (first "www" and just the domain name for A-records, then "www", domain name, and different IPv6-host names for AAAA records) but the queries were done in parallel (i.e., without waiting for the previous query to finish). The first response for A and AAAA records and the corresponding host names were recorded. The queries had 3 second re-transmission timeout and if there wasn't any response for 10 seconds, all remaining queries for the site were canceled. We used a custom-made Perl script and the Net::DNS [net-dns] module for the DNS queries.

The measurement script used a bind9 DNS server running on the same host as was performing the measurement. The DNS cache of the server was flushed before each measurement run in order to detect the

changes in the DNS records in real-time. The host, and thus the DNS server, was not part of DNS IPv6 whitelisting agreements.

The local network where the host performing the measurements was has native IPv6 (dual-stack) connectivity. The IPv6 connectivity to the local network was provided by an IPv6-over-IPv4 tunnel from the network's default router to the ISP's IPv6 peering point.

After obtaining IP addresses for the site, if a site had both A and AAAA records, a simple C program was used to create TCP connections to the port 80 (HTTP) simultaneously using both IPv4 and IPv6 to the (first) IP addresses discovered from the DNS. The connection setup was repeated up to 10 times, giving up after the first failed attempt (but only after normal TCP re-transmissions). The connection setup delay was measured by recording the time immediately before and after the connect system call. The host used for measurements is a regular Linux PC with 2.6.32 version kernel and dual-stack Internet connection via Ethernet.

The measurements were started one week before the World IPv6 Day (on Wednesday, June 1st, 17:30 UTC) and were running until July 11th, once every three hours. One test run takes from two to two and a half hours to complete.

The accuracy and generality of the measurement results is limited by several factors. While we ran the tests in three different sites, most of the results discussed in this document present snapshots of the situation from just one measurement point, the Ericsson Research Finland premises, near Helsinki. Also, since one measurement run takes quite a long time, the network characteristics and DNS records may change even during a single run. The first DNS response was used for the TCP connectivity tests and this selection may result in selection of a non-optimal host; yet, a slight preference is given to the "www" and only-domain-name records since their queries were started before the others. While the host performing the measurements was otherwise idle, the local network was in regular office use during the measurements. The connectivity setup delay is collected in user space, with regular, non real-time, kernel implementation, resulting in small inaccuracies in the timing information.

4. Measurement Results

4.1. DNS AAAA Records

The number of top 10,000 sites with AAAA DNS records before, during, and after the World IPv6 Day, is shown in [DNS-top10k]. The

measurements performed during the World IPv6 Day are shown on the light gray background.

When the measurements began on June 1st, there were 245 sites (2.45%) of the top 10,000 sites with both A and AAAA record. During the following days the number of such sites slowly increased, reaching 306 sites at the measurement that was started 22:30 UTC on June 7th, the evening before the World IPv6 Day. When the World IPv6 Day officially started, the following measurement (1:30 UTC) recorded 383 sites, and the next one 472 sites. During the day the number of sites with AAAA records peaked at 491 (4.91% of the measured 10,000 sites) at 19:30 UTC.

When the World IPv6 Day was over, the number of AAAA records dropped nearly as fast as it had increased just 24 hours earlier. However, the number of sites stabilized around 310 and did not drop below 300 since, resulting in over 3% of the top 10,000 sites still having AAAA records at the end of our measurements.

While 274 sites had IPv6 enabled in their DNS for some of the tested host names one day before the World IPv6 Day, only 116 had it for the "www" host name that is commonly used when accessing a web site. The number of "www" host names with AAAA records more than tripled during the World IPv6 Day reaching 374 sites for 3 consecutive measurement runs (i.e., for at least 6 hours). Also the number of AAAA records for the "www" host name dropped steeply after the day and remained at around 160 sites since.

Similar but more pronounced trends can be seen if only top 100 of the most popular sites are taken into considerations, as show in [DNS-top100]. Here, the number of sites with some of the tested host names having an AAAA record was initially 14, jumped to 36 during the day, and eventually dropped to 13. Also, while none of the top 100 sites apparently had an AAAA record for their "www" host name before and after the World IPv6 day, during the day the number peaked at 30. Thus, roughly one third of the 100 most popular sites had IPv6 enabled for the World IPv6 Day.

Two other test sites in Sweden and Canada experienced similar trends with the DNS records. However, one of the sites used an external DNS server that was part of whitelisting agreements. There the number of sites with AAAA records before the World IPv6 Day was already higher (above 400) and hence the impact of the day was smaller as the amount of sites increased to same numbers as seen by the test site in Finland. With the whitelisted DNS server the level of sites remained above 450 after the day.

4.2. TCP Connection Setup

To test whether the IP addresses given by the DNS actually provide connectivity to the web site, and if there is any difference in the connection setup delay and failure rates with IPv4 and IPv6, we attempted to create TCP connections for all domains that contained both A and AAAA DNS records. The fraction of sites for which the first DNS response gave addresses that were not accessible with TCP to port 80 over IPv4 or IPv6 is shown in [TCP-fails].

There is a baseline failure rate with IPv4 around 1-3% that is fairly static throughout the test period. For hosts with AAAA records, the fraction of inaccessible sites was much higher: in the beginning up to one fourth of the tested hosts did not respond to TCP connection attempts. Much of this was likely due to the various test sites with different "IPv6 prefixes" (as discussed in Section 3); in the first run more than half of the tested sites with AAAA records used them for the first DNS response. Also, some of the hosts may not even be supposed to be accessed with HTTP but provide AAAA records for other purposes while some sites had clear configuration errors, such as localhost or link-local IPv6 addresses.

As the World IPv6 Day came closer, the number of inaccessible IPv6 sites decreased slowly and the number of sites with AAAA records increased at the same time, resulting in the failure ratio dropping to roughly 20% before the day. During the day the number of IPv6 sites increased rapidly but also the number of failures decreased and hence, at the end of the day, the failure ratio dropped to just above 10%. After the World IPv6 Day when many of the participating IPv6 hosts were taken off-line, the fraction of failed sites for IPv6 increased. However, since there was no increase in the absolute number of failed sites, the fraction of inaccessible sites remained at a lower level, between 15% and 20%, than before the day.

4.3. TCP Connection Delays

For sites that were accessible with both IPv4 and IPv6, we measured the time difference between establishing a TCP connection with IPv4 and IPv6. We took the median (as defined in Section 11.3 of [RFC2330]) of the time differences of all 10 connections, and then median and mean (of the median) over all sites; the result is shown in [timediff].

In general, the delay differences are small: median of medians stays less than 3ms off from zero (i.e., IPv4 and IPv6 delays being equal) and even the mean, which is more sensitive to outliers, stays most of the time within +/- 5ms; with the greatest spikes reaching to roughly -15ms (i.e., mean of median IPv6 delays being 15ms larger than for

IPv4 delays). Closer inspection of the results shows that the spikes are often caused by only one or a handful of sites with bad connectivity and multiple re-transmissions of TCP SYN and ACK packets resulting in connection setup delays an order of magnitude larger.

Surprisingly the median delay for IPv6 connections is in most cases equal to or smaller than the IPv4 delay, but during the World IPv6 Day, the IPv6 delays increased slightly and became (as median) slower than their IPv4 counterparts. One reason for such an effect was that some of the sites that enabled IPv6 for the World IPv6 Day, had extremely low, less than 10ms, IPv4 delay (e.g., due to Content Delivery Network (CDN) provider hosting the IPv4 site), but "regular", over hundred millisecond, delay for the IPv6 host.

More detailed analysis of the TCP connection setup delay differences, and the reasons behind them, is left for future work.

5. Conclusions

The World IPv6 Day had a very visible impact to the availability of content over IPv6, particularly when considering the top 100 content providers. It is difficult to find other examples of bigger one day swings in some characteristic of the Internet. However, the impact on end users was small, given that when dual-stack works correctly it should not be visible at the user level and that IPv6 availability for end users themselves is small.

The key conclusions are as follows:

- o The day caused a large jump in the number of content providers providing AAAA DNS records on that day.
- o The day caused a smaller but apparently permanent increase in the number of content providers supporting AAAA.
- o Large and sudden swings in the relative amount of IPv4 vs. IPv6 traffic are possible merely by supporting a dual-stack access network and having a few large content providers offer their service either globally or to this particular network over IPv6.
- o Large fraction of sites that published AAAA records for a name under their domain (be it "www" or "www6" or something else) were actually not responding to TCP SYN requests on IPv6. This fraction is far higher than that which we've seen in our previous measurements, and we are still determining why that is the case. Measurement errors or problems on our side of the network cannot be ruled out at this stage. In any case, it is also clear that as

new sites join, incomplete or in-progress configurations create more connectivity problems in the IPv6 Internet than we've seen before. Other measurements are needed to verify what the general level IPv6 connectivity is to addresses publicly listed in AAAA records.

- o Even if the overall level of connection failures was high, activities on and around the IPv6 day appear to have caused a significant permanent drop in the number of failures.
- o When IPv6 and IPv4 connectivity were both available, the delay characteristics appear very similar. In other words, most of the providers that made IPv6 connectivity available appear to provide a production quality network. TCP connection setup delay differences due to RTT differences between IPv4 and IPv6 connections are in general low. In the remaining differences in our measurements, random packet loss plays a major role. However, some sites can experience considerable differences simply because of different content distribution mechanisms used for IPv4 and IPv6 content.

It is promising that the amount of most popular Internet content on IPv6 was surprisingly high, roughly one third of top 100 sites (during the IPv6 day or with whitelisting enabled). However, other content on the Internet forms a long tail that is harder to move to IPv6. For instance, only 3% of the 10,000 most popular web sites provided their content over IPv6 before the IPv6 day. On a positive note, the top 100 sites form a very large part of overall Internet traffic [Labovitz] and thus even the top sites moving to IPv6 could represent a significant fraction of Internet traffic on IPv6. However, this requires that users are enabled to use IPv6 in their access networks. We believe that this should be the goal of future global IPv6 efforts.

6. Security Considerations

Security issues have not been discussed in this memo.

7. IANA Considerations

This memo has no IANA implications.

8. References

8.1. Normative References

[timediff]

Keranen, A., "TCP connection setup delay differences [RFC editor: please change the references to the graphs to refer to the PDF version of the document]", June 2011, <<http://users.piuha.net/akeranen/drafts/v6day/mda.pdf>>.

[DNS-top10k]

Keranen, A., "Number of sites with AAAA DNS records in the top 10,000 most popular sites", June 2011, <<http://users.piuha.net/akeranen/drafts/v6day/v6sites.pdf>>.

[DNS-top100]

Keranen, A., "Number of sites with AAAA DNS records in the top 100 most popular sites", June 2011, <<http://users.piuha.net/akeranen/drafts/v6day/v6sites-top100.pdf>>.

[TCP-fails]

Keranen, A., "TCP connection setup failure ratio (for the first DNS response)", June 2011, <<http://users.piuha.net/akeranen/drafts/v6day/tcp-fails.pdf>>.

8.2. Informative References

[RFC2330] Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, May 1998.

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

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

[RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.

[net-dns] Fuhr, M., "Net::DNS", <<http://www.net-dns.org/>>.

[IPv6Monitor]

Comcast and University of Pennsylvania, "IPv6 Adoption Monitor", <<http://ipv6monitor.comcast.net>>.

[RIPEv6Day]

RIPE NCC, "World IPv6 Day Measurements", <<http://v6day.ripe.net/>>.

[Alexa] Alexa the Web Information Company, "Alexa Top 1,000,000 Sites",
 <<http://s3.amazonaws.com/alexa-static/top-1m.csv.zip>>.

[Labovitz] Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., and F. Jahanian, "Internet Inter-Domain Traffic",
 Proceedings of ACM SIGCOMM 2010, August 2010.

Appendix A. Acknowledgments

The authors would like to thank Suresh Krishnan, Fredrik Garneij, Lorenzo Colitti, Jason Livingood, Alain Durand, Emile Aben, Jan Melen, and Tero Kauppinen for interesting discussions in this problem space. Thanks also to Tom Petch and Bob Hinden for thorough reviews and many helpful comments.

Authors' Addresses

Ari Keranen
Ericsson
Jorvas 02420
Finland

Email: ari.keranen@ericsson.com

Jari Arkko
Ericsson
Jorvas 02420
Finland

Email: jari.arkko@piuha.net

v6ops
Internet-Draft
Intended status: Informational
Expires: January 6, 2012

V. Kuarsingh, Ed.
Rogers Communications
July 5, 2011

Wireline Incremental IPv6
draft-kuarsingh-wireline-incremental-ipv6-00

Abstract

Operators are currently challenged with enabling IPv6 within their networks while maintaining IPv4 connectivity beyond IPv4 address depletion. In the Wireline world, this will often require the replacement of access network equipment and consumer equipment along with the uplift of the core network. During the transition from the IPv4-Only service environment to the IPv6/IPv4 dual service environment, operators may often need to use multiple transition technologies and mechanisms to maintain services for their customer base. This draft is set up to show how some Wireline providers (including Cable, DSL and Fibre) may accomplish this using tunnelling, translation and native IPv6 services.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 6, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

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

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Motivation	3
3. Reasons for a Phased Approach	4
3.1. Relevance of IPv6 and IPv4	4
3.2. IPv4 Resource Challenges	4
3.3. IPv6 Introduction and Maturity	5
3.4. Impact to Operators	5
4. IPv6 Transition Technology Analysis	6
4.1. Automatic Tunnelling using 6to4 and Teredo	6
4.2. Carrier Grade NAT (NAT444)	7
4.3. 6RD	7
4.4. Native Dual Stack	8
4.5. DS-Lite	8
5. IPv6 Transition Phases	9
5.1. Phase 0 - Foundation	9
5.1.1. Phase 0 - Foundation: Training	9
5.1.2. Phase 0 - Foundation: Routing	10
5.1.3. Phase 0 - Foundation: Network Policy and Security	10
5.1.4. Phase 0 - Foundation: Transition Architecture	10
5.2. Phase 1 - Tunnelled IPv6	10
5.3. Phase 2: Native Dual Stack	11
5.4. Intermediate Phase for CGN	12
5.5. Phase 3 - Tunnelled IPv4	12
6. IANA Considerations	13
7. Security Considerations	13
8. Acknowledgements	13
9. References	13
9.1. Normative References	13
9.2. Informative References	13
Author's Address	14

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 to implement IPv6 within their networks, and many others 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 continuance 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 by identifying which technologies can be used in an incremental fashion to transition from an IPv4-only environment to an efficient IPv6/IPv4 environment. Although no single plan will work for for all operators, generically, those 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. This document also attempts to follow the principles laid out in [RFC6180] which provides guidance on using transition mechanisms. This document will show how tunnelling using 6RD and DS-Lite as well as translation via CGN can be used with native IPv6 to deliver effective dual stack 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. 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. 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 stack IPv6/IPv4 system and eventually to an IPv6 centric world.

3. Reasons for a Phased Approach

Operators may want to consider a phased approach to IPv6 service introduction for a number of reasons. These reasons include the relevance of both IPv4 and IPv6 services in the new ecosystem over the next few years. Both protocols will play a key role in providing a holistic service to customers in various ways. IPv4 resources will likely become depleted in many networks during the IPv6 transition inhibiting the general use of traditional dual stack. Additionally, IPv6 will often be a new protocol for operators and their staff further challenging their task and potentially limiting how quickly they can move to full IPv6 dependence.

3.1. 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.

Additionally, the Internet is made of of many interconnecting systems, networks and various 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.

3.2. IPv4 Resource Challenges

Since connectivity to IPv4-only endpoints and/or content will remain a reality for a 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 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. Given the reality that 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.

3.3. 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 time and was matured by experience. Although many of these experiences are transferable from IPv4 to IPv6, new experience is necessary for IPv6 nonetheless.

Engineering and Operational staff will need to become acclimatized to IPv6 which will take time as experience is gained. During this ramp up period, Operators will need to be aware that instability may occur 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/or are still under development and refinement. 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.

3.4. Impact to Operators

The lack of new IPv4 addresses related to depletion and the relative maturity state of IPv6 within the operator network will both impact what technologies can be used, when they can be used and how they can be used. Operators are welcome to evaluate the impact of these challenges on their own, but some considerations are highlighted herein.

The lack of IPv4 addresses will surely mean that any service requiring it's use as a method to deliver just IPv4, or a vehicle to deliver IPv6 may be short lived. This may also limited their usefulness to initial transition phases. Nothing precludes an operator from using technologies for longer periods of time, but the relative impacts need to be considered. Also, some technologies based on native IPv6 delivery will need to be weighed as well. This includes traditional dual stack and more importantly technologies like DS-Lite which require a native IPv6 path to the customer premise. The operator may want to wait until a certain maturity level is reached with respect to IPv6 before making IPv6 connectivity mandatory to service IPv4 flows given the potential for failure at the outset.

4. 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 how they can or may be used in a transitioning operator network.

4.1. 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 available.

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 and may be inclined to provide the best possible experience for these endpoints. Drafts such as [I-D.ietf-v6ops-6to4-advisory] 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 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.

4.2. Carrier Grade NAT (NAT444)

Carrier Grade NAT (CGN), specifically as deployed in a NAT444 scenario [I-D.ietf-behave-lsn-requirements], is also a relevant technology. CGN/NAT444 is not a IPv6 specific function, but may prove beneficial for those operators who offer dual stack services to endpoints. CGNs are known to cause certain challenges for the IPv4 service path as described in documents like [I-D.donley-nat444-impacts], but may 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/NAT444 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 NAT444. When considered in the overall IPv6 transition, CGN/NAT444 may play a vital role in delivery Internet services.

4.3. 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 IPv6. 6RD provides tunnelled connectivity to IPv6 over the existing IPv4 path. The lack of native IPv6 for a customer premise may be related to technological challenges of delivering IPv6 on a give access type or related to other operational or technical impediments that may existing in an operator's environment.

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. Over time, as the access edge is upgraded, 6RD can be replaced by native IPv6 access. 6RD can be delivered along with CGN/NAT444, but this 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 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 meet 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/NAT444 is used, there are stateful considerations to be made on the NAT444 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.

4.4. Native Dual Stack

Native Dual Stack is often referred to as the "Gold Standard" of IPv6 and IPv4 delivery. It is a method of service delivery which is already used in some deployments. Native Dual Stack does however require that Native IPv6 be delivered to the customer premise. This technology option is most desirable in many cases and can be used immediately if the access network and customer premise equipment supports IPv6, or can also be used incrementally to tunnelling options such as 6RD over time.

As time progresses, Native Dual Stack may be challenging to deliver if more IPv4 addresses are not available on the IPv4 path. For a sub-set of the IPv6 Native Dual Stack Customers, operators may include CGN/NAT444 as an assist technology. Delivering Native Dual Stack would require the operator's core and access network support IPv6 with the required assisting systems like DHCPv6, DNS, and diagnostic/management facilities to help maintain the IPv6 connection.

4.5. DS-Lite

DS-Lite, as described in [I-D.ietf-softwire-dual-stack-lite], 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 ability to continue to grow 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.

5. IPv6 Transition Phases

The Phases described below are not provided as a ridged set of stops but as a guideline which can be considered by the operator. The phases reflect the need to support IPv4 and IPv6 during transition as well as the premise that some technologies may prove beneficial at various periods during the IPv6 transition.

Operators may want to follow all these steps, skip some steps if possible or develop their own plans should they have other considerations which may be of relevance to them. The main goal however should be a set of phases that helps introduce IPv6, and allows for both protocols to work for a period of time as the Internet as whole moves to IPv6, followed by a declining need to add more IPv4 support.

Additional guidelines and information on utilizing IPv6 transition mechanisms can also be found in [RFC6180].

5.1. Phase 0 - Foundation

Before moving an organization to support IPv6 services, a foundation needs to be made to support IPv6. This foundation includes the following basic (non-exhaustive) list of items.

5.1.1. Phase 0 - Foundation: Training

Training may seem to be the most obvious step, but needs to be done effectively and widely across key technical personnel. Unlike IPv4 which had existed for a long period of time before it became "mission critical" for many organizations, IPv6 is being introduced at a time when IP services are vial for most many Internet users. This should not be taken lightly and organizations need to commit to training their staff. Staff may also have far less if any experience with

IPv6 which is not the same as with IPv4. This means the little expose they get in their training may be all they have to lean on as they seek to support IPv6 at the outset.

5.1.2. Phase 0 - Foundation: Routing

The network will all 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 the number of addresses which are made available, careful attention to a scalable and manageable architecture needs to be made. Also, given that customer environments will no longer receive a token single address as is common in IPv4, operators will need to understand the impacts of delegating large sums of addresses (Prefixes) to consumer endpoints. Delegating prefixes can be of specific importance in Cable 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.

5.1.3. Phase 0 - Foundation: Network Policy and Security

Like many principles, network policy and security need to be considered for IPv6. It is possible that many of the IPv4 policies may transfer over to the IPv6 world, others may not be applicable. 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.

5.1.4. Phase 0 - Foundation: Transition Architecture

The operator may want 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/NAT444, 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 roles. This would allow for the least disruption as the operator evolves the transition environment to meet the needs of the network.

5.2. Phase 1 - Tunnelled IPv6

During the initial phase of transition the operator may want to support IPv6 before native IPv6 services are possible on the access network. During this period of time, tunnelled access to IPv6 is likely a very viable and desirable option. 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 selection [RFC3484].

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 the retail channel products and within the OEM market making it a viable option for a wide range of operations. Retail availability of 6RD is important since not all operators control or have influence over what is deployed in the consumer site.

If the operator does not have the access network challenge of deploying Naive IPv6, they may want to skip this phase. However, the operator may still want to deploy 6to4 and/or Teredo relays to help the automatic tunnelling technology operation while Native IPv6 is deployed. This initial phase also provides the added benefit of allow the operational folks to deterministically know what the IPv6 prefix assignment is based on the IPv4 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.

5.3. Phase 2: Native Dual Stack

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 access by the customer equipment. It is also possible that the second phase be enabled in pockets of the network while the access network is undergoing upgrades.

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 work with content providers and internal groups to move content 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, incrementally route or relay customer traffic within the operator's network.

5.4. Intermediate Phase for CGN

During the first two phases, acquiring more IPv4 addresses may become challenging, therefore CGN may be required. The CGN/NAT444 infrastructure can be enabled if needed during either phase. CGN/NAT444 is less optimal in a 6RD deployment (if used with 6RD to a given endpoint) since all traffic must transverse some type of operator equipment.

In the case of Native Dual Stack, CGN/NAT444 can be used to assist in extending connectivity for the IPv4 path. During this time, for endpoints subject to the CGN/NAT444 function, the Native IPv6 path is available for higher quality connectivity. It would be the expectation that the IPv6 path becomes better utilized by the customer over time by virtue of IPv6 support in the home network and within the content provider's realm.

5.5. Phase 3 - Tunnelled IPv4

Over time, the operator will mature IPv6 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 customers. DS-Lite still requires IPv4 address sharing, but this is seen as no worse and often more advantageous than NAT444 and other address sharing options give a single NAT layer.

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 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.

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. IANA Considerations

No IANA considerations are defined at this time.

7. Security Considerations

No Additional Security Considerations are made in this document.

8. 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, and Chris Donley.

9. References

9.1. 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-01 (work in progress), February 2011.
- [RFC6180] Arkko, J. and F. Baker, "Guidelines for Using IPv6 Transition Mechanisms during IPv6 Deployment", RFC 6180, May 2011.

9.2. 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 IP address sharing schemes", draft-ietf-behave-lsn-requirements-01 (work in progress), March 2011.

- [I-D.ietf-softwire-dual-stack-lite]
Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", draft-ietf-softwire-dual-stack-lite-11 (work in progress), May 2011.
- [I-D.ietf-v6ops-6to4-advisory]
Carpenter, B., "Advisory Guidelines for 6to4 Deployment", draft-ietf-v6ops-6to4-advisory-02 (work in progress), June 2011.
- [I-D.jjmb-v6ops-comcast-ipv6-experiences]
Brzozowski, J., "Comcast IPv6 Experiences", draft-jjmb-v6ops-comcast-ipv6-experiences-00 (work in progress), March 2011.
- [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.
- [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.

Author's Address

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

Email: victor.kuarsingh@rci.rogers.com
URI: <http://www.rogers.com>

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 these 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 from 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

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	4
2. Motivation	4
3. Operator Assumptions	5
4. Reasons and Considerations for a Phased Approach	5
4.1. Relevance of IPv6 and IPv4	6
4.2. IPv4 Resource Challenges	6
4.3. IPv6 Introduction and Maturity	7
4.4. Service Management	8
4.5. Sub-Optimal Operation of Transition Technologies	8
5. IPv6 Transition Technology Analysis	9
5.1. Automatic Tunnelling using 6to4 and Teredo	9
5.2. Carrier Grade NAT (NAT444)	10
5.3. 6RD	11
5.4. Native Dual Stack	12
5.5. DS-Lite	12
5.6. NAT64	13
6. IPv6 Transition Phases	13
6.1. Phase 0 - Foundation	14
6.1.1. Phase 0 - Foundation: Training	14
6.1.2. Phase 0 - Foundation: Routing	15
6.1.3. Phase 0 - Foundation: Network Policy and Security	15
6.1.4. Phase 0 - Foundation: Transition Architecture	15
6.1.5. Phase 0- Foundation: Tools and Management	16
6.2. Phase 1 - Tunnelled IPv6	16
6.2.1. 6RD Deployment Considerations	17
6.3. Phase 2: Native Dual Stack	20
6.3.1. Native Dual Stack Deployment Considerations	20
6.4. Intermediate Phase for CGN	21
6.4.1. CGN Deployment Considerations	22
6.5. Phase 3 - Tunnelled IPv4	23
6.5.1. DS-Lite Deployment Considerations	24
7. IANA Considerations	25
8. Security Considerations	25
9. Acknowledgements	25
10. References	25
10.1. Normative References	25
10.2. Informative References	26
Author's Address	27

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 Stack 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 approach 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 be 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 as 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 than 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 wish to constrain such customers to IPv4 delivery until Native IPv6 is available. Also, the operator may wish 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 that 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 assumed 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 be 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 use 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 (CGN), 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 exist 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 many 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 large 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 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 interworking 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.

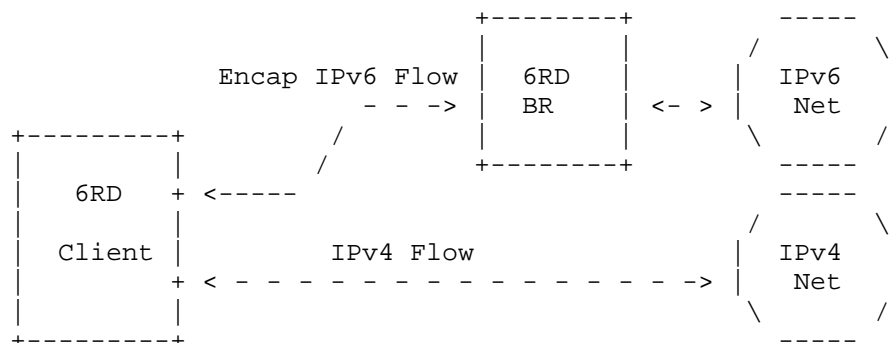


Figure 1: 6RD Basic Model

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 network 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.

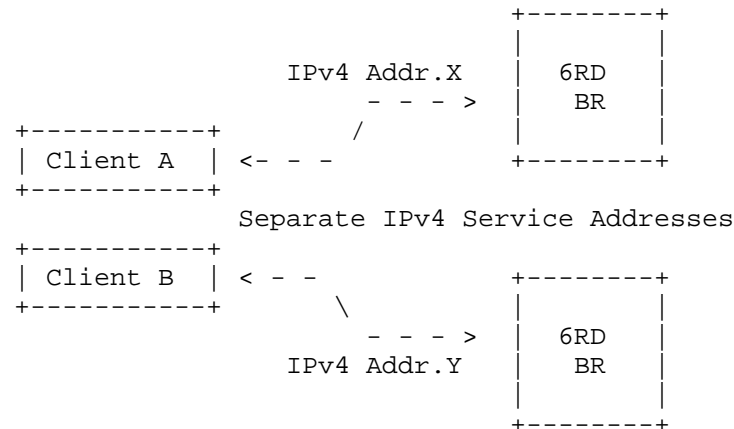


Figure 2: 6RD Multiple IPv4 Service Address Model

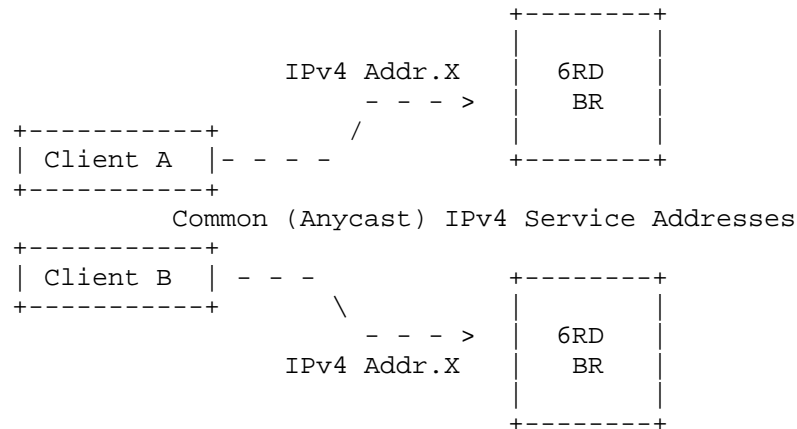


Figure 3: 6RD Anycast IPv4 Service Address Model

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.

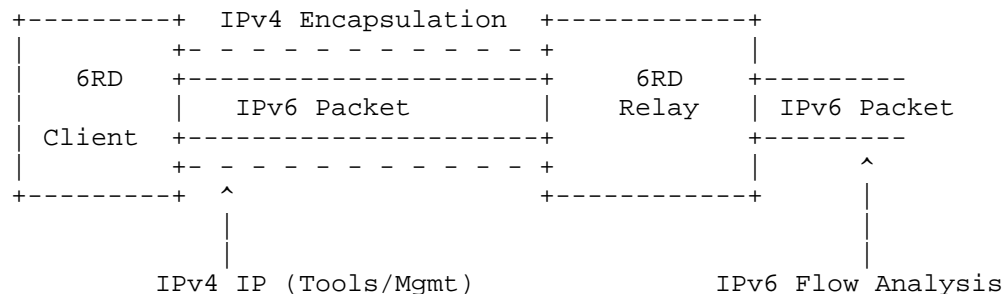


Figure 4: 6RD Tools and Flow Management

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

straight 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 need to build new tools that help manage 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 planned 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).

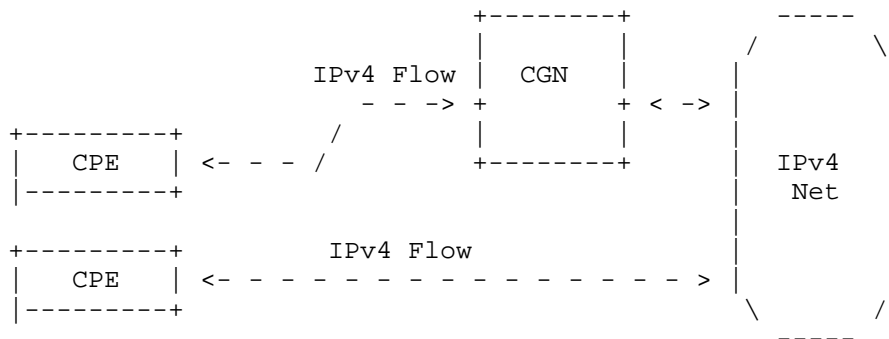


Figure 5: Overlay CGN Deployment

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 than optimal performance.

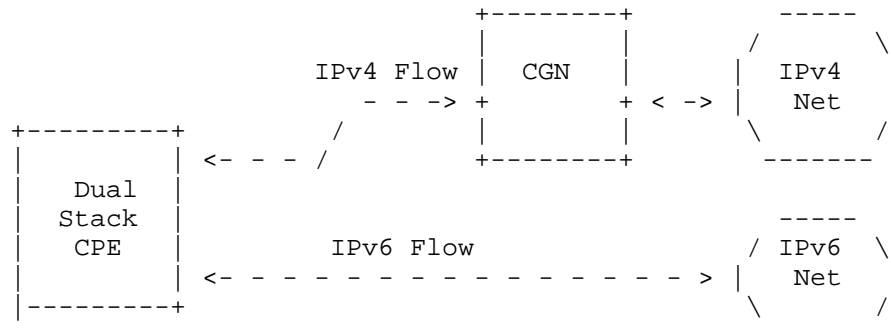


Figure 6: Dual Stack with CGN

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.

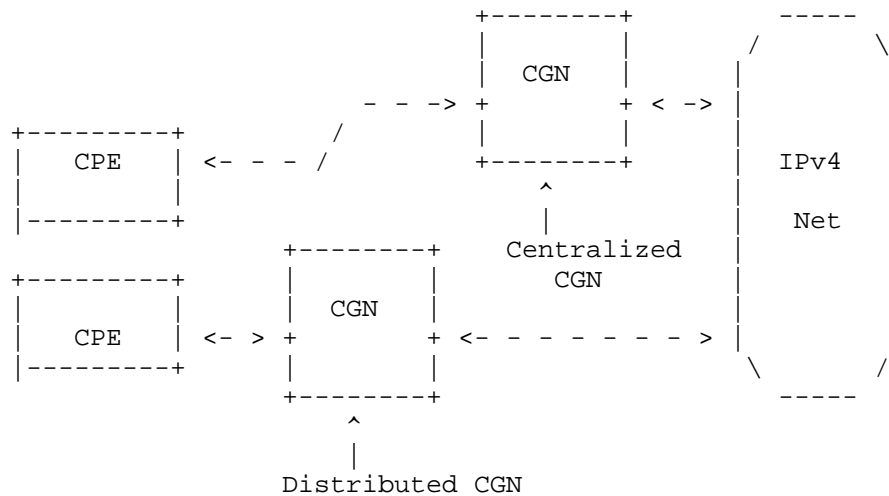


Figure 7: CGN Deployment: Centralized vs. Distributed

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 than 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.

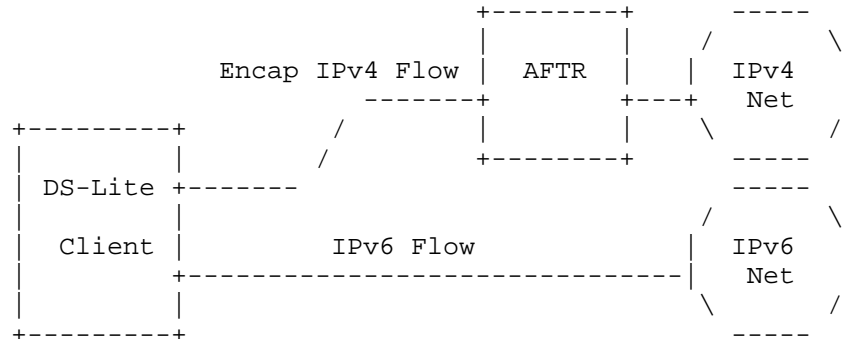


Figure 8: DS-Lite Basic Model

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.

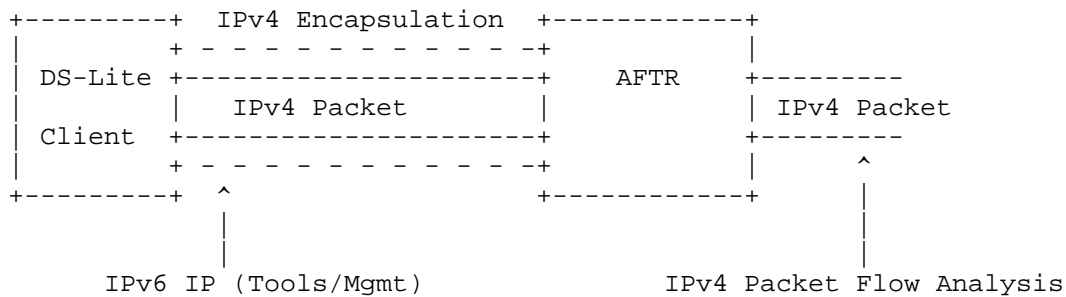


Figure 9: DS-Lite Tools and Flow Analysis

DS-Lite also requires client support. If the operator has chosen 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: Informational
Expires: December 18, 2011

B. Sarikaya
F. Xia
Huawei USA
T. Lemon
Nominum
June 16, 2011

DHCPv6 Prefix Delegation as IPv6 Migration Tool in Mobile Networks
draft-sarikaya-v6ops-prefix-delegation-07.txt

Abstract

As interest on IPv6 deployment is increasing in cellular networks several migration issues are being raised and IPv6 prefix management is the one addressed in this document. Based on the idea that DHCPv6 servers can manage prefixes, we address prefix management issues such as the access router offloading delegation and release tasks of the prefixes to a DHCPv6 server using DHCPv6 Prefix Delegation. The access router first requests a prefix for an incoming mobile node from the DHCPv6 server. The access router may next do stateless or stateful address allocation to the mobile node, e.g. with a Router Advertisement or using DHCP. We also describe prefix management using Authentication Authorization and Accounting servers.

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 December 18, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
2. Terminology	4
3. Prefix Delegation Using DHCPv6	4
3.1. Prefix Request Procedure for Stateless Address Configuration	4
3.2. Prefix Request Procedure for Stateful Address Configuration	6
3.3. MN as Requesting Router in Prefix Delegation	7
3.4. Prefix Release Procedure	8
3.5. Miscellaneous Considerations	8
3.5.1. How to Generate IAID	8
3.5.2. Policy to Delegate Prefixes	9
4. Prefix Delegation Using RADIUS and Diameter	9
5. Security Considerations	10
6. IANA Considerations	10
7. Acknowledgements	11
8. References	11
8.1. Normative References	11
8.2. Informative References	12
Authors' Addresses	12

1. Introduction

Figure 1 illustrates the key elements of a typical cellular access network. In a Long Term Evolution (LTE) network, access router is the packet data network (PDN) gateway [ThreeGPP23401].

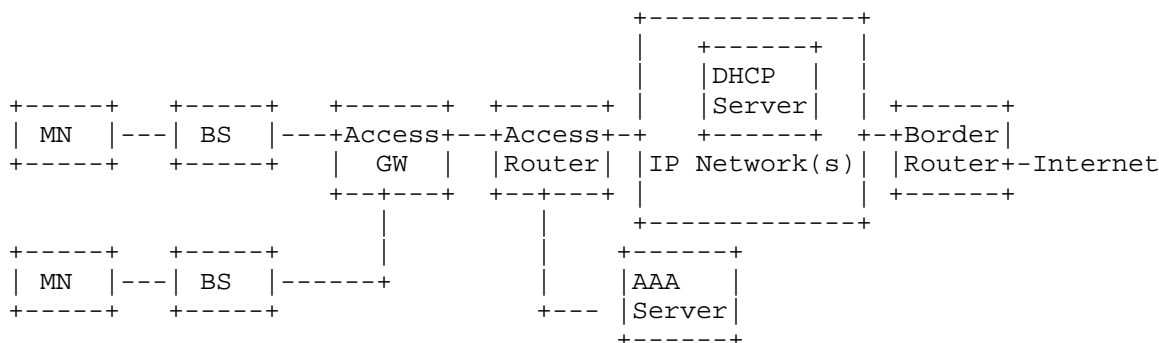


Figure 1: Key elements of a typical cellular network

Mobile node (MN) attaches to a base station (BS) through LTE air interface. A BS manages connectivity of UEs and extends connections to an Access Gateway (GW), e.g. the serving gateway (S-GW) in an LTE network. The access gateway and the Access Router (AR) are connected with an IP network. The access router is the first hop router of MNs and it is in charge of address/prefix management.

Access router is connected to an IP network which is owned by the operator which is connected to the public Internet via a Border Router. The network contains servers for subscriber management including Quality of Service, billing and accounting as well as Dynamic Host Configuration Protocol (DHCP) server [I-D.ietf-v6ops-v6-in-mobile-networks].

As to IPv6 addressing, because mobile network links are point-to-point (p2p) Per-MN interface prefix model is used [RFC3314], [RFC3316]. In Per-MN interface prefix model, prefix management is an issue.

When an MN attaches an AR, the AR requests one or more prefixes for the MN. When the MN detaches the AR, the prefixes should be released. When the MN becomes idle, the AR should hold the prefixes allocated.

This document describes how to use DHCPv6 Prefix Delegation (PD) in mobile networks such as networks based on standards developed by the 3rd Generation Partnership Project (3GPP) and it could easily be

adopted to Worldwide Interoperability for Microwave Access (WiMAX) Forum as well. In view of migration to IPv6, the number of mobile nodes connected to the network at a given time may become very high. Traditional techniques such as prefix pools are not scalable. In such cases DHCPv6 PD becomes the viable approach to take.

2. Terminology

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].

This document uses the terminology defined in [RFC3315], [RFC3314], [RFC3316] and [RFC3633].

3. Prefix Delegation Using DHCPv6

Access router refers to the cellular network entity that has DHCP Client. According to [ThreeGPP23401] DHCP Client is located in PDN Gateway. So AR is the PDN Gateway in LTE architecture.

3.1. Prefix Request Procedure for Stateless Address Configuration

There are two function modules in the AR, DHCP Client and DHCP Relay. DHCP messages should be relayed if the AR and a DHCP server are not connected directly, otherwise DHCP relay function in the AR is not necessary. Figure 2 illustrates the scenario that the AR and the DHCP Server aren't connected directly:

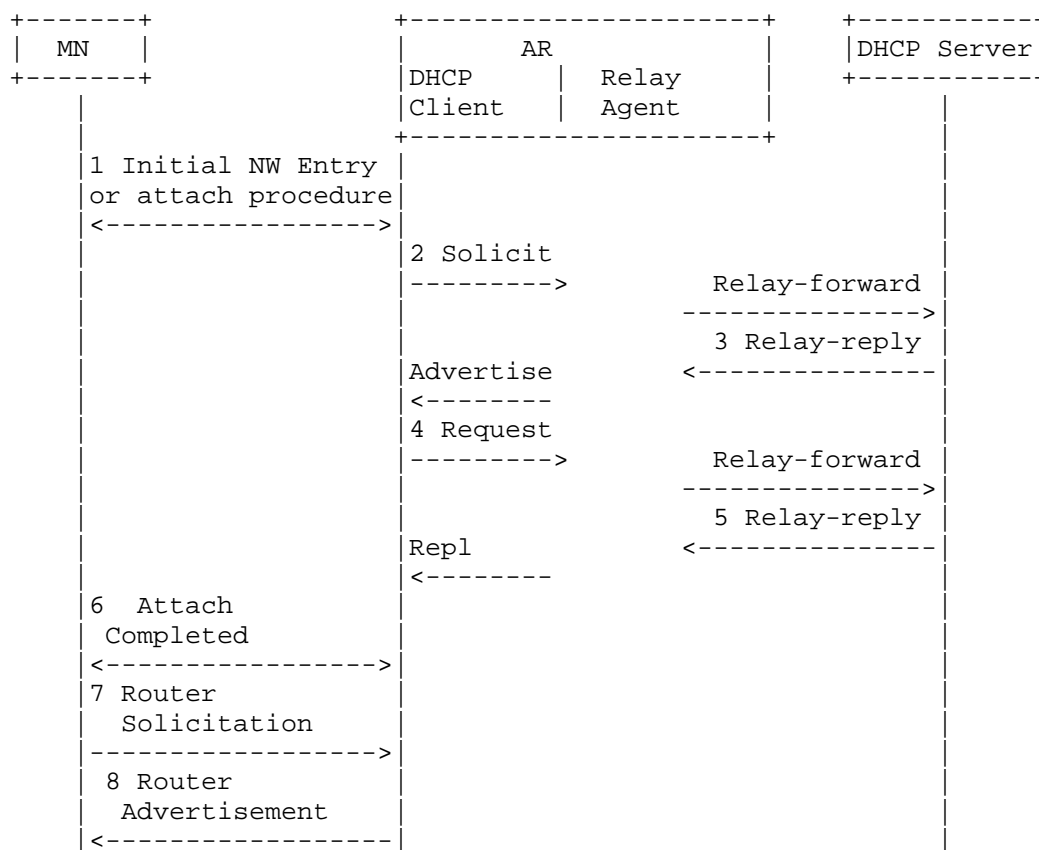


Figure 2: Prefix request

1. An MN (UE=User Equipment in 3GPP) performs initial network entry and authentication procedures, a.k.a. attach procedure.
2. On successful completion of Step 1, the AR initiates DHCP Solicit procedure to request prefixes for the MN. The DHCP Client in AR creates and transmits a Solicit message as described in sections 17.1.1, "Creation of Solicit Messages" and 17.1.2, "Transmission of Solicit Messages" of [RFC3315]. The DHCP Client in AR that supports DHCPv6 Prefix Delegation [RFC3633] creates an Identity Association for Prefix Delegation (IA_PD) and assigns it an Identity Association Identifier (IAID). The client MUST include the IA_PD option in the Solicit message. DHCP Client as Requesting Router MUST set prefix-length field to 64 to request a /64 prefix. Next, the Relay Agent in AR sends Relay-Forward message to the DHCP Server encapsulating Solicit message.

3. The DHCP server sends an Advertise message to the AR in the same way as described in section 17.2.2, "Creation and transmission of Advertise messages" of [RFC3315]. Advertise message with IA_PD shows that the DHCP server is capable of delegating prefixes. This message is received encapsulated in Relay-Reply message by the Relay Agent in AR and sent as Advertise message to the DHCP Client in AR.
4. The AR (DHCP Client and Relay Agent) uses the same message exchanges as described in section 18, "DHCP Client-Initiated Configuration Exchange" of [RFC3315] and [RFC3633] to obtain or update prefixes from the DHCP server. The AR (DHCP Client and Relay Agent) and the DHCP server use the IA_PD Prefix option to exchange information about prefixes in much the same way as IA Address options are used for assigned addresses. This is accomplished by the AR sending a DHCP Request message and the DHCP server sending a DHCP Reply message.
5. AR stores the prefix information it received in the Reply message.
6. A connection between MN and AR is established and the link becomes active. This step completes the PDP Context Activation Procedure in UMTS and PDN connection establishment in LTE networks.
7. The MN MAY send a Router Solicitation message to solicit the AR to send a Router Advertisement message.
8. The AR advertises the prefixes received in IA_PD option to MN with router advertisement (RA) once the PDP Context/PDN connection is established or in response to Router Solicitation message sent from the MN.

4-way exchange between AR as requesting router (RR) and DHCP server as delegating router (DR) in Figure 2 MAY be reduced into a two message exchange using the Rapid Commit option [RFC3315]. DHCP Client in AR acting as RR includes a Rapid Commit option in the Solicit message. DR then sends a Reply message containing one or more prefixes.

3.2. Prefix Request Procedure for Stateful Address Configuration

Stateful address configuration requires a different architecture than shown in Figure 2. There are two function modules in the AR, DHCP Server and DHCP Client.

After the initial attach is completed, a connection to the AR is established for the MN. DHCP Client function at the AR as requesting router and DHCP server as delegating router follow Steps 2 through 5 of the procedure shown in Figure 2 to get the new prefix for this interface of MN from IA_PD Option exchange defined in [RFC3633].

DHCPv6 client at the MN sends DHCP Request to AR. DHCP Server function at the AR MUST use the IA_PD option received in DHCP PD exchange to assign an address to MN. IA_PD option MUST contain the prefix. AR sends DHCP Reply message to MN containing IA address option (IAADDR). Figure 3 shows the message sequence.

MN configures its interface with the address assigned by DHCP server in DHCP Reply message.

In Figure 3 AR may be the home gateway of a fixed network to which MN gets connected during MN's handover.

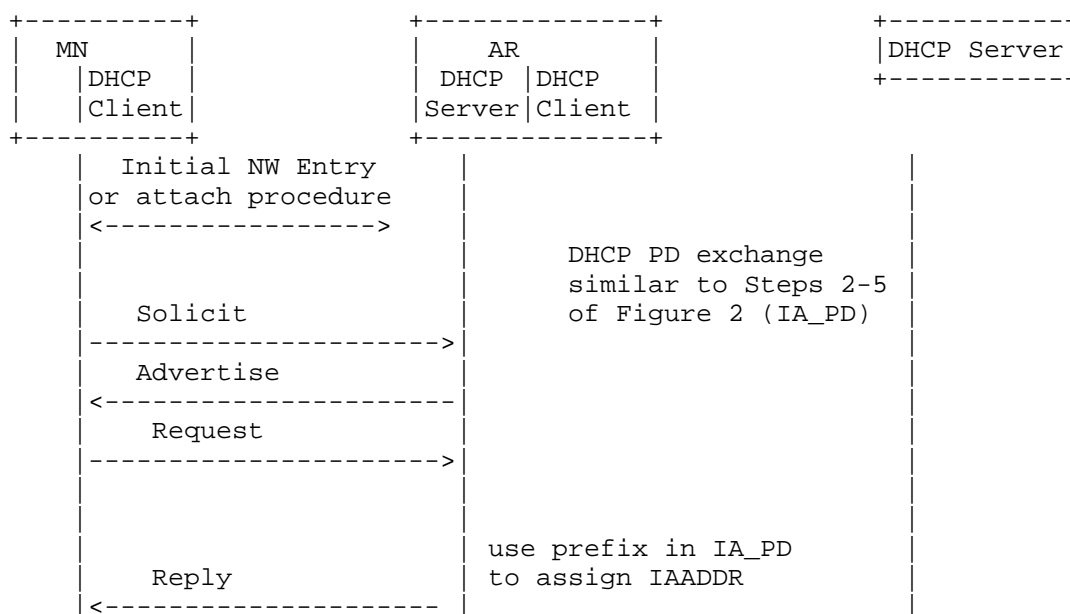


Figure 3: Stateful Address Configuration Following PD

3.3. MN as Requesting Router in Prefix Delegation

AR may use DHCPv6 prefix delegation exchange to get a delegated prefix shorter than /64 by setting prefix-length field to a value less than 64, e.g. 56 to get a /56 prefix. Each newly attaching MN first goes through the steps in Figure 2 in which AR requests a shorter prefix to establish a default connection with the AR.

MN may next request additional /64 prefixes from the AR using DHCPv6 prefix delegation where MN is the requesting router and AR is the delegating router [I-D.ietf-v6ops-3gpp-eps]. In this case the call

flow is similar to Figure 3. Solicit message must include the IA_PD option with prefix-length field set to 64. MN may request more than one /64 prefixes. AR as delegating router must delegate these prefixes excluding the prefix assigned to the default connection.

3.4. Prefix Release Procedure

Prefixes can be released in two ways, prefix aging or DHCP release procedure. In the former way, a prefix SHOULD NOT be used by an MN when the prefix ages, and the DHCP Server can delegate it to another MN. A prefix lifetime is delivered from the DHCPv6 server to the MN through DHCP IA_PD Prefix option [RFC3633] and RA Prefix Information option [RFC4861]. Figure 4 illustrates how the AR releases prefixes to an DHCP Server which isn't connected directly:

1. An MN detachment signaling, such as switch-off or handover, triggers prefix release procedure.
2. The AR initiates a Release message to give back the prefixes to the DHCP server.
3. The server responds with a Reply message, and then the prefixes can be reused by other MNs.

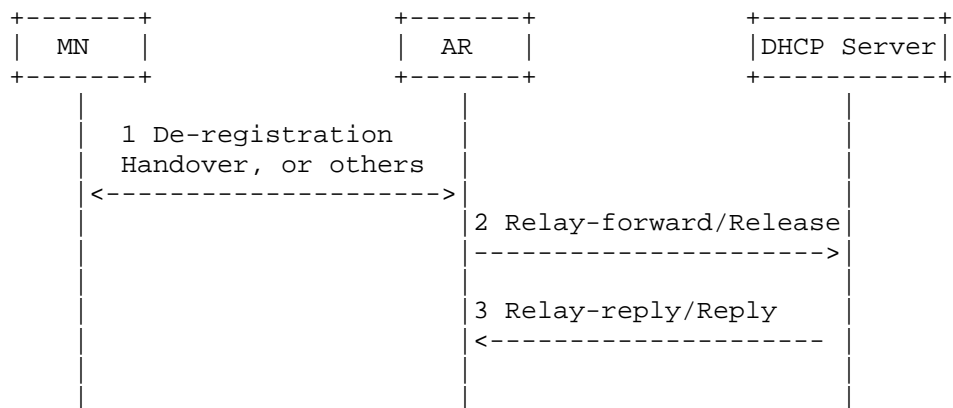


Figure 4: Prefix Release

3.5. Miscellaneous Considerations

3.5.1. How to Generate IAID

IAID is 4 bytes in length and should be unique in an AR scope. Prefix table SHOULD be maintained. Prefix table contains IAID, MAC address and the prefix(es) assigned to MN. In LTE networks, International Mobile station Equipment Identity (IMEI) uniquely identifies MN's interface and thus corresponds to the MAC address.

MAC address of the interface SHOULD be stored in the prefix table and this field is used as the key for searching the table.

IAID SHOULD be set to Start_IAID, an integer of 4 octets. The following IAID generation algorithm is used:

1. Set this IAID value in IA_PD Prefix Option. Request prefix for this MN as in Section 3.1 or Section 3.2.
2. Store IAID, MAC address and the prefix(es) received in the next entry of the prefix table.
3. Increment IAID.

Prefix table entry for an MN that handover to another AR MUST be removed. IAID value is released to be reused.

3.5.2. Policy to Delegate Prefixes

AR should broadcast the prefix(es) dynamically upstream as the route information of all the MNs connected to this AR. In point-to-point links, this causes high routing protocol traffic (IGP, OSPF, etc.) due to Per-MN interface prefixes. To solve the problem, route aggregation SHOULD be used. For example, each AR can be assigned a /48 or /32 prefix (aggregate prefix, aka service provider common prefix) while each interface of MN can be assigned a /64 prefix. The /64 prefix is an extension of /48 one, for example, an AR's /48 prefix is 2001:DB8:0::/48, an interface of MN is assigned 2001:DB8:0:2::/64 prefix. The AR only broadcasts it's /48 or /32 prefix information to Internet.

This policy can be enforced as follows: DHCP Relay MUST set the IPv6 Prefix field in IA_PD Prefix option in IA_PD option in the Relay Forward message to the aggregate prefix (/48, /32, or /16 prefix assigned to the AR).

4. Prefix Delegation Using RADIUS and Diameter

In the initial network entry procedure Figure 2, AR as Remote Authentication Dial In User Service (RADIUS) client sends Access-Request message with MN information to RADIUS server. If the MN passes the authentication, the RADIUS server may send Access-Accept message with prefix information to the AR using Framed-IPv6-Prefix attribute. AAA server also provides routing information to be configured for MN on the AR using Framed-IPv6-Route attribute. Using such a process AR can handle initial prefix assignments to MNs but managing lifetime of the prefixes is totally left to the AR. Framed-IPv6-Prefix is not designed to support delegation of IPv6 prefixes. For this Delegated-IPv6-Prefix attribute can be used which is

discussed next.

[RFC4818] defines a RADIUS attribute Delegated-IPv6-Prefix that carries an IPv6 prefix to be delegated. This attribute is usable within either RADIUS or Diameter. [RFC4818] recommends the delegating router to use AAA server to receive the prefixes to be delegated using Delegated-IPv6-Prefix attribute/AVP.

DHCP server as the delegating router in Figure 2 MAY send an Access-Request packet containing Delegated-IPv6-Prefix attribute to the RADIUS server to request prefixes. In the Access-Request message, the delegating router MAY provide a hint that it would prefer a prefix, for example, a /48 prefix. The RADIUS server MAY delegate a /64 prefix which is an extension of the /48 prefix in an Access-Accept message containing Delegated-IPv6-Prefix attribute. The attribute can appear multiple times when RADIUS server delegates multiple prefixes to the delegating router. The delegating router sends the prefixes to the requesting router using IA_PD Option and AR as RR uses them for MN's as described in Section 3.

When Diameter is used, DHCP server as the delegating router in Figure 2 sends AA-Request message. AA-Request message MAY contain Delegated-IPv6-Prefix AVP. Diameter server replies with AA-Answer message. AA-Answer message MAY contain Delegated-IPv6-Prefix AVP. The AVP can appear multiple times when Diameter server assigns multiple prefixes to MN. The Delegated-IPv6-Prefix AVP MAY appear in an AA-Request packet as a hint by the AR to the Diameter server that it would prefer a prefix, for example, a /48 prefix. Diameter server MAY delegate in an AA-Answer message with a /64 prefix which is an extension of the /48 prefix. As in the case of RADIUS, the delegating router sends the prefixes to the requesting router using IA_PD Option and AR as RR uses them for MN's as described in Section 3.

5. Security Considerations

This draft introduces no additional messages. Comparing to [RFC3633], [RFC2865] and [RFC3588] there is no additional threats to be introduced. DHCPv6, RADIUS and Diameter security procedures apply.

6. IANA Considerations

None.

7. Acknowledgements

We are grateful to Suresh Krishnan, Hemant Singh, Qiang Zhao, Ole Troan, Qin Wu, Jouni Korhonen, Cameron Byrne and Jason Lin who provided in depth reviews of this document that have led to several improvements.

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.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
- [RFC3314] Wasserman, M., "Recommendations for IPv6 in Third Generation Partnership Project (3GPP) Standards", RFC 3314, September 2002.
- [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.
- [RFC3316] Arkko, J., Kuijpers, G., Soliman, H., Loughney, J., and J. Wiljakka, "Internet Protocol Version 6 (IPv6) for Some Second and Third Generation Cellular Hosts", RFC 3316, April 2003.
- [RFC3588] Calhoun, P., Loughney, J., Guttman, E., Zorn, G., and J. Arkko, "Diameter Base Protocol", RFC 3588, September 2003.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC4818] Salowey, J. and R. Droms, "RADIUS Delegated-IPv6-Prefix Attribute", RFC 4818, April 2007.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

8.2. Informative References

[I-D.ietf-v6ops-3gpp-eps]

Korhonen, J., Soininen, J., Patil, B., Savolainen, T.,
Bajko, G., and K. Iisakkila, "IPv6 in 3GPP Evolved Packet
System", draft-ietf-v6ops-3gpp-eps-01 (work in progress),
May 2011.

[I-D.ietf-v6ops-v6-in-mobile-networks]

Koodli, R., "Mobile Networks Considerations for IPv6
Deployment", draft-ietf-v6ops-v6-in-mobile-networks-05
(work in progress), May 2011.

[ThreeGPP23401]

"3GPP TS 23.401 V10.3.0, General Packet Radio Service
(GPRS) enhancements for Evolved Universal Terrestrial
Radio Access Network (E-UTRAN) access (Release 10).",
2011.

Authors' Addresses

Behcet Sarikaya
Huawei USA
5340 Legacy Dr.
Plano, TX 75074

Email: sarikaya@ieee.org

Frank Xia
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075

Phone: +1 972-509-5599
Email: xiayangsong@huawei.com

Ted Lemon
Nominum
2000 Seaport Blvd
Redwood City, CA 94063

Phone:
Email: mellon@nominum.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 13, 2012

B. Sarikaya
F. Xia
Huawei USA
T. Lemon
Nominum
February 10, 2012

DHCPv6 Prefix Delegation in Long Term Evolution (LTE) Networks
draft-sarikaya-v6ops-prefix-delegation-11.txt

Abstract

As interest on IPv6 deployment is increasing in cellular networks several migration issues are being raised and IPv6 prefix management is the one addressed in this document. Based on the idea that DHCPv6 servers can manage prefixes, we address prefix management issues such as the access router offloading delegation and release tasks of the prefixes to a DHCPv6 server using DHCPv6 Prefix Delegation. The access router first requests a prefix for an incoming mobile node from the DHCPv6 server. The access router may next do stateless or stateful address allocation to the mobile node, e.g. with a Router Advertisement or using DHCP. We also describe prefix management using Authentication Authorization and Accounting servers.

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 13, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
2. Terminology and Acronyms	4
3. Prefix Delegation Using DHCPv6	5
3.1. Prefix Request Procedure for Stateless Address Configuration	5
3.2. Prefix Request Procedure for Stateful Address Configuration	7
3.3. MN as Requesting Router in Prefix Delegation	8
3.4. Prefix Release Procedure	8
3.5. Miscellaneous Considerations	9
3.5.1. How to Generate IAID	9
3.5.2. Policy to Delegate Prefixes	10
4. Prefix Delegation Using RADIUS and Diameter	10
5. Security Considerations	11
6. IANA Considerations	11
7. Acknowledgements	11
8. Informative References	12
Authors' Addresses	13

1. Introduction

Figure 1 illustrates the key elements of a typical cellular access network. In a Long Term Evolution (LTE) network, access router is the packet data network (PDN) gateway [ThreeGPP23401].

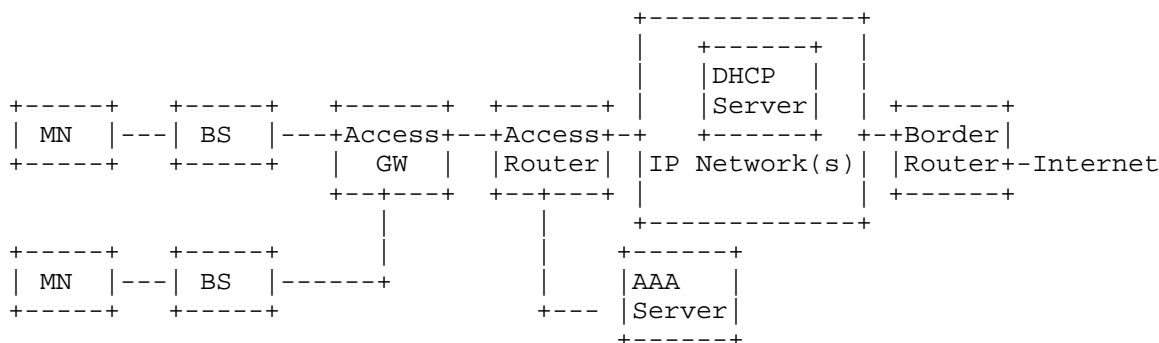


Figure 1: Key elements of a typical cellular network

Mobile node (MN) attaches to a base station (BS) through LTE air interface. A BS manages connectivity of UEs and extends connections to an Access Gateway (GW), e.g. the serving gateway (S-GW) in an LTE network. The access gateway and the Access Router (AR) are connected with an IP network. The access router is the first hop router of MNs and it is in charge of address/prefix management.

Access router is connected to an IP network which is owned by the operator which is connected to the public Internet via a Border Router. The network contains servers for subscriber management including Quality of Service, billing and accounting as well as Dynamic Host Configuration Protocol (DHCP) server [RFC6342].

As to IPv6 addressing, because mobile network links are point-to-point (p2p) Per-MN interface prefix model is used [RFC3314], [RFC3316]. In Per-MN interface prefix model, prefix management is an issue.

When an MN attaches an AR, the AR requests one or more prefixes for the MN. When the MN detaches the AR, the prefixes should be released. When the MN becomes idle, the AR should hold the prefixes allocated.

This document describes how to use DHCPv6 Prefix Delegation (PD) in mobile networks such as networks based on standards developed by the 3rd Generation Partnership Project (3GPP) and it could easily be adopted to Worldwide Interoperability for Microwave Access (WiMAX)

Forum as well. In view of migration to IPv6, the number of mobile nodes connected to the network at a given time may become very high. Traditional techniques such as prefix pools are not scalable. In such cases DHCPv6 PD becomes the viable approach to take.

The techniques described in this document have not been approved either by the IETF or by 3GPP, except what is described below in Section 3.3. This document is not a standard or best current practice. This document is published only as a possibility for consideration by operators.

This document is useful when address space needs to be managed by DHCPv6-PD. There are obviously other means of managing address space, including having the AR track internally what address space is used by what mobile.

2. Terminology and Acronyms

3GPP 3rd Generation Partnership Project

AAA Authentication Authorization and Accounting

AR Access Router

BS Base Station

DHCP Dynamic Host Control Protocol

E-UTRAN Evolved Universal Terrestrial Radio Access Network

GPRS General Packet Radio Service

LTE Long Term Evolution

MN Mobile node

PDN Packet data network

PD Prefix Delegation

p2p Point-to-point

Serving Gateway S-GW

WiMAX Worldwide Interoperability for Microwave Access

3. Prefix Delegation Using DHCPv6

Access router refers to the cellular network entity that has DHCP Client. According to [ThreeGPP23401] DHCP Client is located in PDN Gateway. So AR is the PDN Gateway in LTE architecture.

3.1. Prefix Request Procedure for Stateless Address Configuration

There are two function modules in the AR, DHCP Client and DHCP Relay. DHCP messages should be relayed if the AR and a DHCP server are not connected directly, otherwise DHCP relay function in the AR is not necessary. Figure 2 illustrates the scenario that the AR and the DHCP Server aren't connected directly:

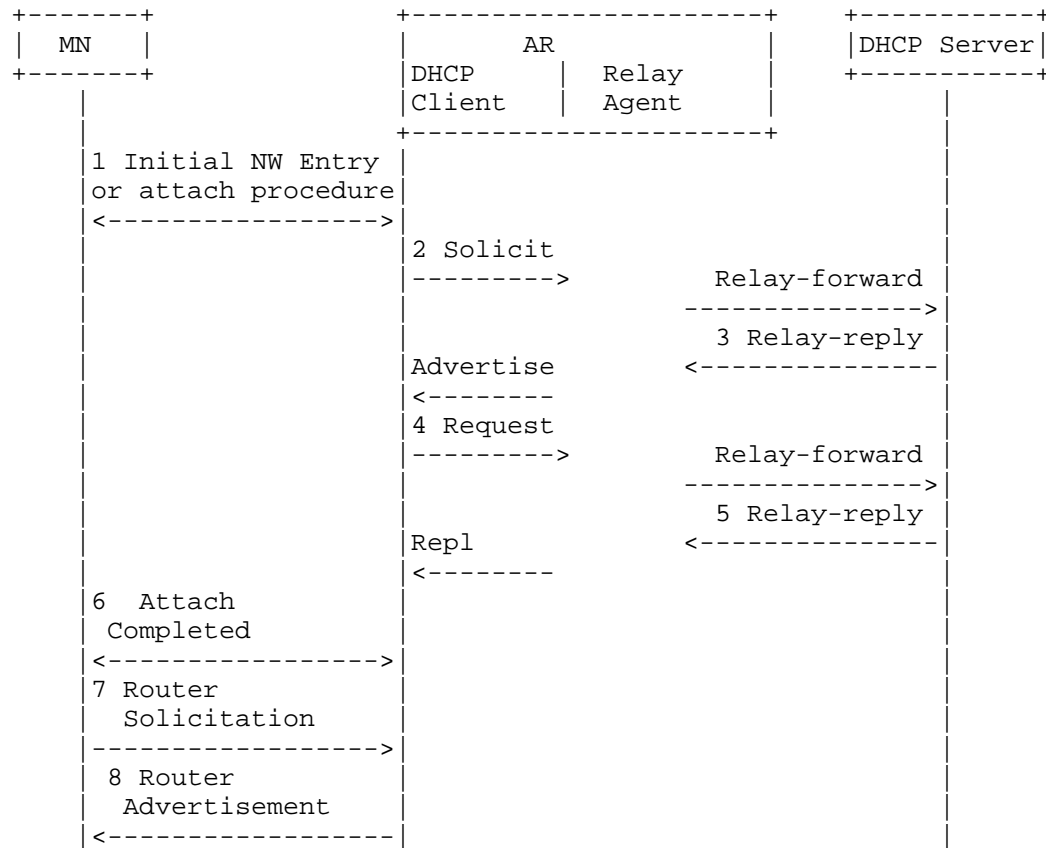


Figure 2: Prefix request

1. An MN (UE=User Equipment in 3GPP) performs initial network entry and authentication procedures, a.k.a. attach procedure.
2. On successful completion of Step 1, the AR initiates DHCP Solicit procedure to request prefixes for the MN. The DHCP Client in AR creates and transmits a Solicit message as described in sections 17.1.1, "Creation of Solicit Messages" and 17.1.2, "Transmission of Solicit Messages" of [RFC3315]. The DHCP Client in AR that supports DHCPv6 Prefix Delegation [RFC3633] creates an Identity Association for Prefix Delegation (IA_PD) and assigns it an Identity Association Identifier (IAID). The client must include the IA_PD option in the Solicit message. DHCP Client as Requesting Router must set prefix-length field to a value less than, e.g. 48 or equal to 64 to request a /64 prefix. Next, the Relay Agent in AR sends Relay-Forward message to the DHCP Server encapsulating Solicit message.
3. The DHCP server sends an Advertise message to the AR in the same way as described in section 17.2.2, "Creation and transmission of Advertise messages" of [RFC3315]. Advertise message with IA_PD shows that the DHCP server is capable of delegating prefixes. This message is received encapsulated in Relay-Reply message by the Relay Agent in AR and sent as Advertise message to the DHCP Client in AR.
4. The AR (DHCP Client and Relay Agent) uses the same message exchanges as described in section 18, "DHCP Client-Initiated Configuration Exchange" of [RFC3315] and [RFC3633] to obtain or update prefixes from the DHCP server. The AR (DHCP Client and Relay Agent) and the DHCP server use the IA_PD Prefix option to exchange information about prefixes in much the same way as IA Address options are used for assigned addresses. This is accomplished by the AR sending a DHCP Request message and the DHCP server sending a DHCP Reply message.
5. AR stores the prefix information it received in the Reply message.
6. A connection between MN and AR is established and the link becomes active. This step completes the PDP Context Activation Procedure in UMTS and PDN connection establishment in LTE networks.
7. The MN may send a Router Solicitation message to solicit the AR to send a Router Advertisement message.
8. The AR advertises the prefixes received in IA_PD option to MN with router advertisement (RA) once the PDP Context/PDN connection is established or in response to Router Solicitation message sent from the MN.

4-way exchange between AR as requesting router (RR) and DHCP server as delegating router (DR) in Figure 2 may be reduced into a two message exchange using the Rapid Commit option [RFC3315]. DHCP Client in AR acting as RR includes a Rapid Commit option in the

Solicit message. DR then sends a Reply message containing one or more prefixes.

3.2. Prefix Request Procedure for Stateful Address Configuration

Stateful address configuration requires a different architecture than shown in Figure 2. There are two function modules in the AR, DHCP Server and DHCP Client.

After the initial attach is completed, a connection to the AR is established for the MN. DHCP Client function at the AR as requesting router and DHCP server as delegating router follow Steps 2 through 5 of the procedure shown in Figure 2 to get the new prefix for this interface of MN from IA_PD Option exchange defined in [RFC3633].

DHCPv6 client at the MN sends DHCP Request to AR. DHCP Server function at the AR must use the IA_PD option received in DHCP PD exchange to assign an address to MN. IA_PD option must contain the prefix. AR sends DHCP Reply message to MN containing IA address option (IAADDR). Figure 3 shows the message sequence.

MN configures its interface with the address assigned by DHCP server in DHCP Reply message.

In Figure 3 AR may be the home gateway of a fixed network to which MN gets connected during MN's handover.

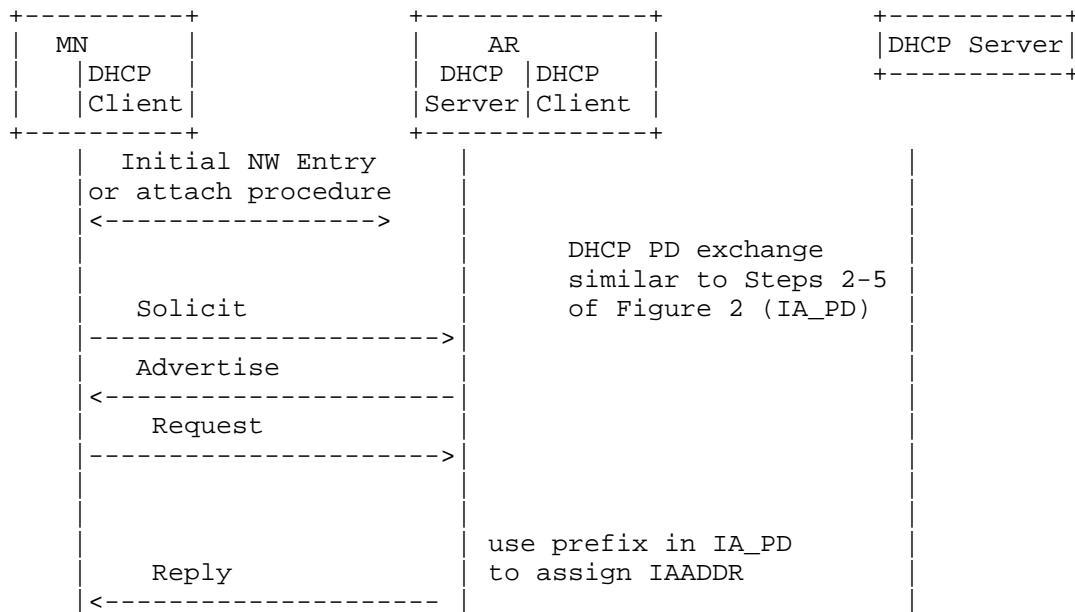


Figure 3: Stateful Address Configuration Following PD

3.3. MN as Requesting Router in Prefix Delegation

AR may use DHCPv6 prefix delegation exchange to get a delegated prefix shorter than /64 by setting prefix-length field to a value less than 64, e.g. 56 to get a /56 prefix. Each newly attaching MN first goes through the steps in Figure 2 in which AR requests a shorter prefix to establish a default connection with the MN.

MN may next request additional prefixes (/64 or shorter) from the AR using DHCPv6 prefix delegation where MN is the requesting router and AR is the delegating router [RFC6459], Section 5.3.1.2.6 in [ThreeGPP23401]. In this case the call flow is similar to Figure 3. Solicit message must include the IA_PD option with prefix-length field set to 64. MN may request more than one /64 prefixes. AR as delegating router must delegate these prefixes excluding the prefix assigned to the default connection.

3.4. Prefix Release Procedure

Prefixes can be released in two ways, prefix aging or DHCP release procedure. In the former way, a prefix should not be used by an MN when the prefix ages, and the DHCP Server can delegate it to another MN. A prefix lifetime is delivered from the DHCPv6 server to the MN

through DHCP IA_PD Prefix option [RFC3633] and RA Prefix Information option [RFC4861]. Figure 4 illustrates how the AR releases prefixes to a DHCP Server which isn't connected directly:

1. An MN detachment signaling, such as switch-off or handover, triggers prefix release procedure.
2. The AR initiates a Release message to give back the prefixes to the DHCP server.
3. The server responds with a Reply message, and then the prefixes can be reused by other MNs.

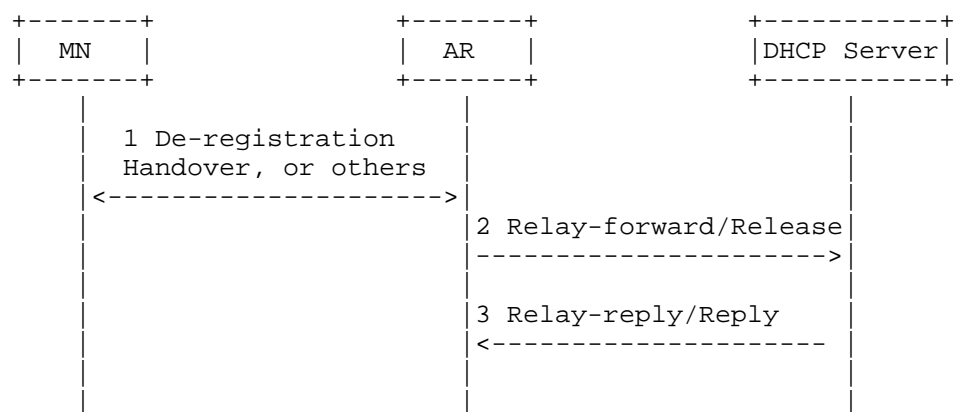


Figure 4: Prefix Release

3.5. Miscellaneous Considerations

3.5.1. How to Generate IAID

IAID is 4 bytes in length and should be unique in an AR scope. Prefix table should be maintained. Prefix table contains IAID, MAC address and the prefix(es) assigned to MN. In LTE networks, International Mobile station Equipment Identity (IMEI) uniquely identifies MN's interface and thus corresponds to the MAC address. MAC address of the interface should be stored in the prefix table and this field is used as the key for searching the table.

IAID should be set to Start_IAID, an integer of 4 octets. The following IAID generation algorithm is used:

1. Set this IAID value in IA_PD Prefix Option. Request prefix for this MN as in Section 3.1 or Section 3.2.
2. Store IAID, MAC address and the prefix(es) received in the next entry of the prefix table.

3. Increment IAID.

Prefix table entry for an MN that hands over to another AR must be removed. IAID value is released to be reused.

3.5.2. Policy to Delegate Prefixes

In point-to-point links, if /64 prefixes of all the MNs connected to one or more ARs are broadcast dynamically upstream as the route information this causes high routing protocol traffic (IGP, OSPF, etc.) due to Per-MN interface prefixes. There are two solutions this problem. One is to use static configuration, which would be preferable in many cases. No routing protocols are needed, because each AR has a known piece of address space. If the DHCP servers know this space, too, then they will assign from that space to a particular AR.

The other method is to use route aggregation. For example, each AR can be assigned a /48 or /32 prefix (aggregate prefix, aka service provider common prefix) while each interface of MN can be assigned a /64 prefix. The /64 prefix is an extension of /48 one, for example, an AR's /48 prefix is 2001:DB8:0::/48, an interface of MN is assigned 2001:DB8:0:2::/64 prefix. The border router (BR) in Figure 1 may be manually configured to broadcast only individual AR's /48 or /32 prefix information to Internet.

4. Prefix Delegation Using RADIUS and Diameter

In the initial network entry procedure Figure 2, AR as Remote Authentication Dial In User Service (RADIUS) client sends Access-Request message with MN information to RADIUS server. If the MN passes the authentication, the RADIUS server may send Access-Accept message with prefix information to the AR using Framed-IPv6-Prefix attribute. AAA server also provides routing information to be configured for MN on the AR using Framed-IPv6-Route attribute. Using such a process AR can handle initial prefix assignments to MNs but managing lifetime of the prefixes is totally left to the AR. Framed-IPv6-Prefix is not designed to support delegation of IPv6 prefixes. For this Delegated-IPv6-Prefix attribute can be used which is discussed next.

[RFC4818] defines a RADIUS attribute Delegated-IPv6-Prefix that carries an IPv6 prefix to be delegated. This attribute is usable within either RADIUS or Diameter. [RFC4818] recommends the delegating router to use AAA server to receive the prefixes to be delegated using Delegated-IPv6-Prefix attribute/AVP.

DHCP server as the delegating router in Figure 2 may send an Access-Request packet containing Delegated-IPv6-Prefix attribute to the RADIUS server to request prefixes. In the Access-Request message, the delegating router may provide a hint that it would prefer a prefix, for example, a /48 prefix. As the RADIUS server is not required to honor the hint, the server may delegate longer prefix, e.g. /56 or /64 in an Access-Accept message containing Delegated-IPv6-Prefix attribute [RFC4818]. The attribute can appear multiple times when RADIUS server delegates multiple prefixes to the delegating router. The delegating router sends the prefixes to the requesting router using IA_PD Option and AR as RR uses them for MN's as described in Section 3.

When Diameter is used, DHCP server as the delegating router in Figure 2 sends AA-Request message. AA-Request message may contain Delegated-IPv6-Prefix AVP. Diameter server replies with AA-Answer message. AA-Answer message may contain Delegated-IPv6-Prefix AVP. The AVP can appear multiple times when Diameter server assigns multiple prefixes to MN. The Delegated-IPv6-Prefix AVP may appear in an AA-Request packet as a hint by the AR to the Diameter server that it would prefer a prefix, for example, a /48 prefix. Diameter server may delegate in an AA-Answer message with a /64 prefix which is an extension of the /48 prefix. As in the case of RADIUS, the delegating router sends the prefixes to the requesting router using IA_PD Option and AR as RR uses them for MN's as described in Section 3.

5. Security Considerations

This draft introduces no additional messages. Comparing to [RFC3633], [RFC2865] and [RFC3588] there is no additional threats to be introduced. DHCPv6, RADIUS and Diameter security procedures apply.

6. IANA Considerations

None.

7. Acknowledgements

We are grateful to Suresh Krishnan, Hemant Singh, Qiang Zhao, Ole Troan, Qin Wu, Jouni Korhonen, Cameron Byrne, Brian Carpenter, Jari Arkko and Jason Lin who provided in depth reviews of this document that have led to several improvements.

8. Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
- [RFC3314] Wasserman, M., "Recommendations for IPv6 in Third Generation Partnership Project (3GPP) Standards", RFC 3314, September 2002.
- [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.
- [RFC3316] Arkko, J., Kuijpers, G., Soliman, H., Loughney, J., and J. Wiljakka, "Internet Protocol Version 6 (IPv6) for Some Second and Third Generation Cellular Hosts", RFC 3316, April 2003.
- [RFC3588] Calhoun, P., Loughney, J., Guttman, E., Zorn, G., and J. Arkko, "Diameter Base Protocol", RFC 3588, September 2003.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC4818] Salowey, J. and R. Droms, "RADIUS Delegated-IPv6-Prefix Attribute", RFC 4818, April 2007.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC6342] Koodli, R., "Mobile Networks Considerations for IPv6 Deployment", RFC 6342, August 2011.
- [RFC6459] Korhonen, J., Soininen, J., Patil, B., Savolainen, T., Bajko, G., and K. Iisakkila, "IPv6 in 3rd Generation Partnership Project (3GPP) Evolved Packet System (EPS)", RFC 6459, January 2012.
- [ThreeGPP23401] "3GPP TS 23.401 V11.0.0, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (Release 11).",

2011.

Authors' Addresses

Behcet Sarikaya
Huawei USA
5340 Legacy Dr.
Plano, TX 75074

Email: sarikaya@ieee.org

Frank Xia
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075

Phone: +1 972-509-5599
Email: xiayangsong@huawei.com

Ted Lemon
Nominum
2000 Seaport Blvd
Redwood City, CA 94063

Phone:
Email: mellon@nominum.com

v6ops
Internet-Draft
Intended status: Informational
Expires: January 12, 2012

Q. Sun
C. Xie
Q. Liu
China Telecom
X. Li
Tsinghua University
J. Qin
ZTE
D. Liu
BII Group
July 11, 2011

Rapid Transition of IPv4 contents to be IPv6-accessible
draft-sunq-v6ops-contents-transition-01

Abstract

This document describes one deployment model of NAT64, aiming at rapidly increasing the amount of IPv6 accessible contents for users from IPv6 Internet.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 2012.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
1.1. Requirements Language	4
2. Motivations	4
2.1. Transition As A Service	5
3. Deployment Model	5
4. The Implementation	7
4.1. Address Mapping	7
4.2. DNS implementations	8
4.3. Fragmentation	8
4.4. Examples	8
4.5. Logging and Statistics	9
5. Acknowledgements	9
6. IANA Considerations	9
7. Security Considerations	9
8. References	10
8.1. Normative References	10
8.2. Informative References	10
Authors' Addresses	10

1. Introduction

The global IPv4 address depletion becomes a reality. Although the IPv4 to IPv6 transition is considered inevitable, deployments of IPv6 are still quite limited as this document is written. Facing the pressure of IPv4 address shortage, the operators may like to provide services through IPv6 in some ways. However, compared to the readiness of operators' infrastructures, the IPv6 transition on the content provider and end user sides moves even more slowly. The lack of IPv6-reachable contents becomes one of the main obstacles.

This document describes one deployment model of the stateful IPv4/IPv6 translation [RFC6146],[RFC6052],[RFC6144],[RFC6145] , aiming at rapidly increasing the amount of IPv6-reachable contents with lower cost at the early stage of transition, for users from IPv6 Internet. The contents can be still accessible through IPv4. While this would be very helpful for CP/SPs to achieve rapid transition, the native transition of contents (by "Dual-Stack") should always be recommended.

1.1. Requirements Language

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

2. Motivations

There have been statements from several popular content providers that they have turned on, or planned to turn on IPv6 soon, which do have a beneficial effect on encouraging end users' transition to IPv6. While given the operational cost, the risk to the continuity of service delivery and compared to the number of active IPv6 users currently, it is difficult to convince much more content providers (especially the great many ones of small-to-medium size) to immediately enable IPv6 natively and make their publically-facing services accessible through IPv6. 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.

2.1. Transition As A Service

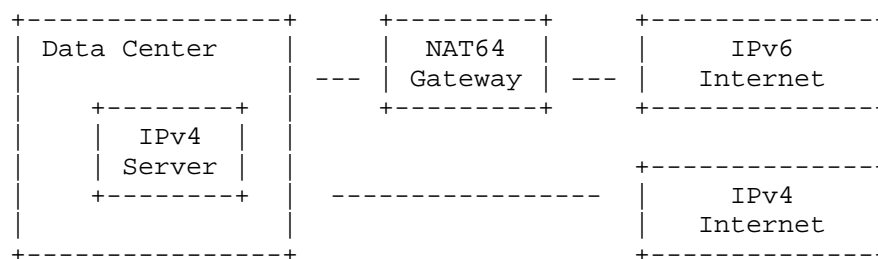
The deployment model of the stateful IPv4/IPv6 translation [RFC6146],[RFC6052],[RFC6144],[RFC6145] described in this document can be regarded as a transition service offered by the operators, to small-to-medium size content providers (e.g. , those who rent servers from the operators). By this means, they could make the publically-facing services to be IPv6-accessible shortly with lower cost, compared to the employment of "Dual-Stack" approach where the software or codes may have to be upgraded, which requires additional investment and expertise right away. Moreover, in this deployment model, we can still make use of current IPv4 security infrastructures in data centers, e.g. firewalls, IPS, etc.

For larger content providers (e.g. those who manage servers, or even the Data Centers of their own), this deployment model can also be attractive at the very early stage of transition (considering risks to the service continuity, and the costs). If there are load balancing devices deployed already, the NAT64 functional elements are likely to be co-located on these boxes naturally.

But it should be noted that the purpose of this deployment model is to encourage the IPv6 transition with economic justification within given transition period. The Dual-Stack mode which is the most straightforward approach should still be recommended to customers from the very beginning if the costs and risks are acceptable to them.

3. Deployment Model

The NAT64 gateway is deployed between the IPv6 Internet and the IPv4 servers[RFC6144]. See the following as an example.



In this deployment model, the Stateful NAT64 is performed to translate IPv6 packets to IPv4 and vice versa. The guidance in [RFC6146],[RFC6052],[RFC6144],[RFC6145] should be followed. The

communications are initiated from the IPv6 side. The IPv6 node will firstly get A/AAAA addresses of the server from DNS, and then the communication will follow the path to NAT64 Gateway. When an IPv6 packet arrives at NAT64 Gateway, 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.

(1) Mapping and Addressing

The Stateful NAT64 can be operated in either of the two mapping modes:

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

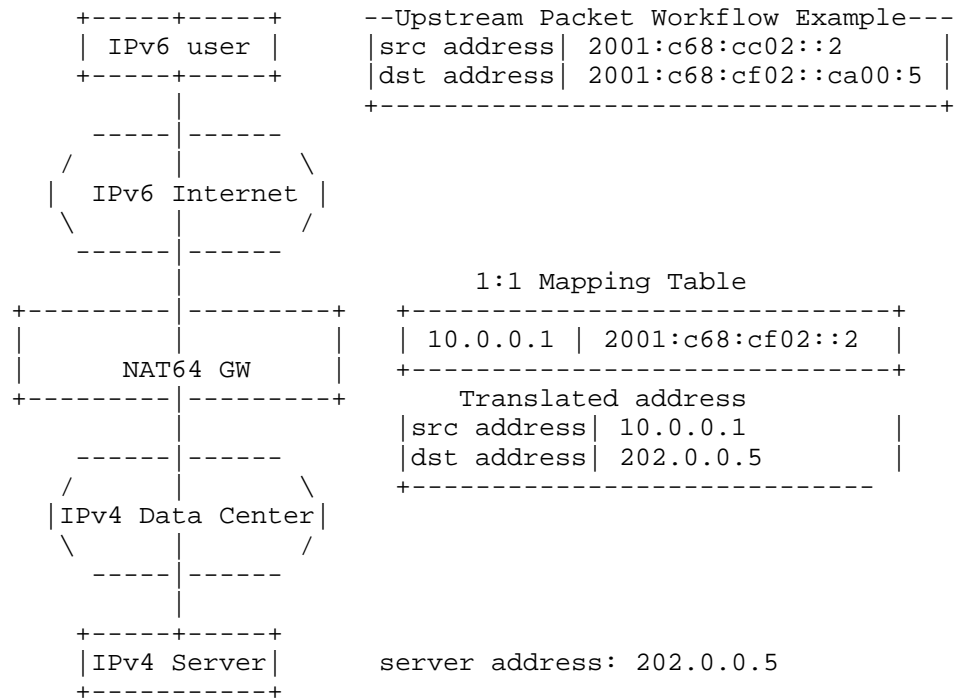
To save global IPv4 addresses which become scarce resources, private blocks, for instance 10.0.0.0/8 may be used for the Stateful NAT64. In addition, an IPv6 prefix is needed to represent the IPv4 server, and the route of the prefix should be advertised to the IPv6 Internet. The IPv4 address of the server can be embedded in the IPv6 prefix following the algorithm specified in [RFC6052].

(2) DNS

Before initiating a session, generally an IPv6 user will generate a DNS lookup to get the AAAA records and learn the addresses of the hosts to access. In this case, the IPv6 addresses learned through AAAA records are those translated from the IPv4 addresses of the server.

Note that the connections may fail in case of IPv4 address literals. Refer to [I-D.wing-behave-http-ip-address-literals] for more details.

The workflow of this model is depicted in the following Figure.



4. The Implementation

The deployments of Stateful NAT64 on the server side are different from those on the client side:

- o The traffic accessing servers come from IPv6 Internet, so the source IPv6 addresses do not belong to prefixes of any special Service Provider's;
- o It is possible to use private IPv4 blocks for the Stateful NAT64;
- o The DNS implementation.

4.1. Address Mapping

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 elected for the Stateful NAT64. By this means, 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.

Additionally, an IPv6 prefix from the Service Provider's space is assigned to represent the servers and form the IPv4-translated AAAA records.

4.2. DNS implementations

To make sure the addresses of servers can be retrieved by IPv6 users before initiating sessions, the AAAA records which are formed through IPv4-translated addresses should be added on the domain's authoritative DNS. The AAAA records under one domain name could be converted from the corresponding A records.

Please note that if the authoritative DNS of given Content Providers' 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 Content Providers must make sure the authoritative AAAA records can be added.

4.3. 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.

4.4. Examples

See below for some sites that have migrated through the approach aforementioned, and are IPv6 accessible:

Content Provider	Categories
www.2118.com.cn	News, BBS, E-commerce, Video
www.5460.net	BBS, Album
www.118326.com	News, Video
www.hnradio.com	Video
www.voc.com.cn	News, BBS, E-commerce, Video
www.chinatelecom.com.cn	News, BBS, Recruitment

4.5. Logging and Statistics

Up to now, there are more than 15 thousands different IPv6 users ever accessing the above six Content Providers through the NAT64 box totally, with 6000 to 7000 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.

The IPv6 users aforementioned are located worldwide. More than 91 percent come from CERNET2, and the rest are from China telecom, USA, Australia, Finland, etc. The total percentage of 6to4 users accounts for approximately 3.2%.

5. Acknowledgements

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

6. IANA Considerations

This document includes no request to IANA.

7. Security Considerations

The security issues and considerations discussed in [RFC6146] apply to the deployment model described in this document.

8. References

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

8.2. Informative References

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

Authors' Addresses

Qiong Sun
China Telecom
Room 708 No.118, Xizhimenneidajie
Beijing, 100035
P.R.China

Phone: +86 10 5855 2923
Email: sunqiong@ctbri.com.cn

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

Phone: +86 10 5855 2116
Email: xiechf@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

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
ZTE
Shanghai,
China

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

Dong Liu
BII Group
Beijing 100028
P.R.China

Phone: +86 138 0103 2487
Email: dliu@biigroup.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.

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 13, 2012.

Copyright Notice

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

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

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	4
2. Requirements Language	4
3. Motivations	5
3.1. Transition As A Service	5
3.2. Guiding the traffic to IPv6 network	6
4. Deployment practice one: Communication from IPv6 users to IPv4 server	6
4.1. Deployment scenario	6
4.2. Mapping and Addressing	7
4.3. DNS	8
4.4. Fragmentation	8
4.5. Logging	8
4.6. Geographically aware services	9
4.7. ALG issues	9
4.8. High Availability	10
4.9. Security	10
4.10. Deployment practices	10
5. Deployment practice two: communications from IPv4 users to IPv6 server	11
5.1. Deployment scenario	11
5.2. Mapping and Addressing	11
5.3. DNS	12
5.4. Logging	12
5.5. Geographically aware services	12
5.6. ALG issues	12
5.7. High Availability	12
5.8. Security	12
5.9. Deployment practices	13
6. Additional Author List	13
7. IANA Considerations	14
8. Acknowledgements	14
9. References	14
9.1. Normative References	14
9.2. Informative References	15
Authors' Addresses	15

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 widely 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.

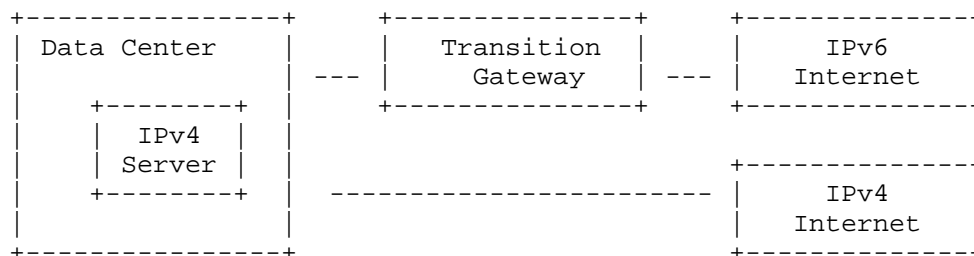


Figure 1: Deployment Model 1

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:

- o 1:1, one IPv6 address is mapped to one IPv4 address (exclusively for given lifetime);
- o 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 than 15 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].

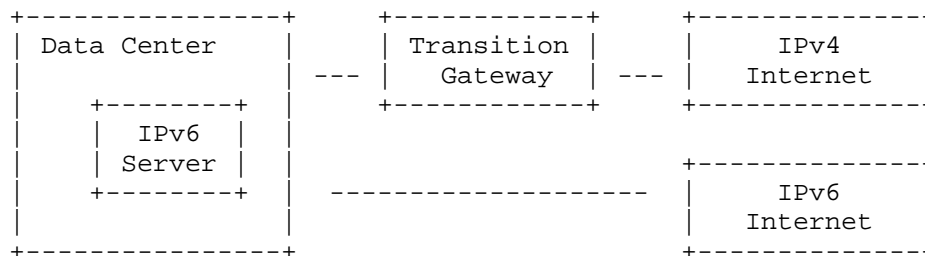


Figure 2: Deployment Model2

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

