

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 5, 2013

JP. Dionne
S. Perreault
Viagenie
T. Tsou
Huawei Technologies (USA)
July 4, 2012

Gap Analysis for IPv4 Sunset
draft-dionne-sunset4-v4gapanalysis-00

Abstract

Sunsetting IPv4 refers to the process of turning off IPv4 definitively. It can be seen as the final phase of the migration to IPv6. This memo analyses difficulties arising when sunseting IPv4, and identifies the gaps resulting in additional work.

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, 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. Related Work	3
3. Remotely Disabling IPv4	3
3.1. Indicating that IPv4 connectivity is unavailable	3
3.2. Disabling IPv4 in the LAN	3
4. Client Connection Establishment Behavior	4
5. Disabling IPv4 in Operating System and Applications	5
6. IANA Considerations	5
7. Security Considerations	5
8. Acknowledgements	5
9. Informative References	5
Authors' Addresses	7

1. Introduction

The final phase of the migration to IPv6 is the sunset of IPv4, that is turning off IPv4 definitively on the attached networks and on the upstream networks.

Some current implementations behavior make it hard to sunset IPv4. Additionally, some new features could be added to IPv4 to make its sunsetting easier. This document analyzes the current situation and proposes new work in this area.

2. Related Work

[RFC3789], [RFC3790], [RFC3791], [RFC3792], [RFC3793], [RFC3794], [RFC3795] and [RFC3796] contain surveys of IETF protocols with their IPv4 dependencies.

3. Remotely Disabling IPv4

3.1. Indicating that IPv4 connectivity is unavailable

PROBLEM 1: When an IPv4 node boots and requests an IPv4 address (e.g., using DHCP), it typically interprets the absence of a response as a failure condition even when it is not.

PROBLEM 2: Home router devices often identify themselves as default routers in DHCP responses that they send to requests coming from the LAN, even in the absence of IPv4 connectivity on the WAN.

One way to address these issues is to send a signal to an dual-stack node that IPv4 connectivity is unavailable. Given that IPv4 shall be off, the message must be delivered through IPv6.

3.2. Disabling IPv4 in the LAN

PROBLEM 3: IPv4-enabled hosts inside an IPv6-only LAN can auto-configure IPv4 addresses [RFC3927] and enable various protocols over IPv4 such as mDNS [I-D.cheshire-dnsext-multicastdns] and LLmNR [RFC4795]. This can be undesirable for operational or security reasons, since in the absence of IPv4, no monitoring or logging of IPv4 will be in place.

PROBLEM 4: IPv4 can be completely disabled on a link by filtering it on the L2 switching device. However, this may not be possible in all cases or complex to deploy. For example, an ISP is often not able to control the L2 switching device in the subscriber home network.

One way to address these issues is to send a signal to an dual-stack node that auto-configuration of IPv4 addresses is undesirable, or that direct IPv4 communications between nodes on the same link should not take place.

This problem was described in [RFC2563], which standardized a DHCP option to disable IPv4 address auto-configuration. However, using this option requires running an IPv4 DHCP server, which is contrary to the goal of IPv4 sunsetting. An equivalent way of signalling this over IPv6 is necessary,, using either Router Advertisements or DHCPv6.

Furthermore, it could be useful to have L2 switches snoop this signalling and automatically start filtering IPv4 traffic as a consequence.

Finally, it could be useful to publish guidelines on how to safely block IPv4 on an L2 switch.

4. Client Connection Establishment Behavior

PROBLEM 5: Happy Eyeballs [RFC6555] refers to the multiple approaches to dual-stack client implementations that try to reduce connection setup delays by trying both IPv4 and IPv6 paths simultaneously. Some implementations introduce delays which provide an advantage to IPv6, while others do not [Huston2012]. The latter will pick the fastest path, no matter whether it is over IPv4 or IPv6, directing more traffic over IPv4 than the other kind of implementations. This can prove problematic in the context of IPv4 sunsetting, especially for Carrier-Grade NAT phasing out.

PROBLEM 6: `getaddrinfo()` [RFC3493] sends DNS queries for both A and AAAA records regardless of the state of IPv4 or IPv6 availability. The `AI_ADDRCONFIG` flag can be used to change this behavior, but it relies on programmers using the `getaddrinfo()` function to always pass this flag to the function. The current situation is that in an IPv6-only environment, many useless A queries are made.

Recommendations on client connection establishment behavior that would facilitate IPv4 sunsetting is therefore appropriate.

5. Disabling IPv4 in Operating System and Applications

PROBLEM 7: Completely disabling IPv4 at runtime often reveals implementation bugs. Hard-coded dependencies on IPv4 abound, such as on the 127.0.0.1 address assigned to the loopback interface. It is therefore often operationally impossible to completely disable IPv4 on individual nodes.

PROBLEM 8: In an IPv6-only world, legacy IPv4 code in operating systems and applications incurs a maintenance overhead and can present security risks.

It is possible to completely remove IPv4 support from an operating system as has been shown by the work of Bjoern Zeeb on FreeBSD. [Zeeb] Removing IPv4 support in the kernel revealed many IPv4 dependencies in libraries and applications.

It would be useful for the IETF to provide guidelines to programmers on how to avoid creating dependencies on IPv4, how to discover existing dependencies, and how to eliminate them. Having programs and operating systems that behave well in an IPv6-only environment is a prerequisite for IPv4 sunsetting.

6. IANA Considerations

None.

7. Security Considerations

TODO

8. Acknowledgements

TODO

9. Informative References

[Huston2012]

Huston, G. and G. Michaelson, "RIPE 64: Analysing Dual

Stack Behaviour and IPv6 Quality", April 2012.

- [I-D.cheshire-dnsext-multicastdns]
Cheshire, S. and M. Krochmal, "Multicast DNS",
draft-cheshire-dnsext-multicastdns-15 (work in progress),
December 2011.
- [RFC2563] Troll, R., "DHCP Option to Disable Stateless Auto-
Configuration in IPv4 Clients", RFC 2563, May 1999.
- [RFC3493] Gilligan, R., Thomson, S., Bound, J., McCann, J., and W.
Stevens, "Basic Socket Interface Extensions for IPv6",
RFC 3493, February 2003.
- [RFC3789] Nesser, P. and A. Bergstrom, "Introduction to the Survey
of IPv4 Addresses in Currently Deployed IETF Standards
Track and Experimental Documents", RFC 3789, June 2004.
- [RFC3790] Mickles, C. and P. Nesser, "Survey of IPv4 Addresses in
Currently Deployed IETF Internet Area Standards Track and
Experimental Documents", RFC 3790, June 2004.
- [RFC3791] Olvera, C. and P. Nesser, "Survey of IPv4 Addresses in
Currently Deployed IETF Routing Area Standards Track and
Experimental Documents", RFC 3791, June 2004.
- [RFC3792] Nesser, P. and A. Bergstrom, "Survey of IPv4 Addresses in
Currently Deployed IETF Security Area Standards Track and
Experimental Documents", RFC 3792, June 2004.
- [RFC3793] Nesser, P. and A. Bergstrom, "Survey of IPv4 Addresses in
Currently Deployed IETF Sub-IP Area Standards Track and
Experimental Documents", RFC 3793, June 2004.
- [RFC3794] Nesser, P. and A. Bergstrom, "Survey of IPv4 Addresses in
Currently Deployed IETF Transport Area Standards Track and
Experimental Documents", RFC 3794, June 2004.
- [RFC3795] Sofia, R. and P. Nesser, "Survey of IPv4 Addresses in
Currently Deployed IETF Application Area Standards Track
and Experimental Documents", RFC 3795, June 2004.
- [RFC3796] Nesser, P. and A. Bergstrom, "Survey of IPv4 Addresses in
Currently Deployed IETF Operations & Management Area
Standards Track and Experimental Documents", RFC 3796,
June 2004.
- [RFC3927] Cheshire, S., Aboba, B., and E. Guttman, "Dynamic

Configuration of IPv4 Link-Local Addresses", RFC 3927,
May 2005.

[RFC4795] Aboba, B., Thaler, D., and L. Esibov, "Link-local
Multicast Name Resolution (LLMNR)", RFC 4795,
January 2007.

[RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with
Dual-Stack Hosts", RFC 6555, April 2012.

[Zeeb] "FreeBSD Snapshots without IPv4 support",
<<http://wiki.freebsd.org/IPv6Only>>.

Authors' Addresses

Jean-Philippe Dionne
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: jean-philippe.dionne@viagenie.ca
URI: <http://viagenie.ca>

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: simon.perreault@viagenie.ca
URI: <http://viagenie.ca>

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 12, 2013

C. Donley
C. Grundemann
V. Sarawat
K. Sundaresan
CableLabs
July 11, 2012

Deterministic Address Mapping to Reduce Logging in Carrier Grade NAT
Deployments
draft-donley-behave-deterministic-cgn-04

Abstract

Abuse response in a Carrier Grade NAT environment requires Service Providers to be able to map a subscriber's inside address with the address used on the public Internet. Unfortunately, many Carrier Grade NAT abuse-response solutions require per-connection logging. Research indicates that such logging is not scalable to many residential broadband services. This document suggests a way to manage Carrier Grade NAT translations in such a way as to significantly reduce the amount of logging required while providing traceability for abuse response. While the authors acknowledge that IPv6 is a preferred solution, Carrier Grade NAT is a reality in many networks, and is needed in situations where either customer equipment or Internet content only supports IPv4; this approach should in no way slow the deployment of IPv6.

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, 2013.

Copyright Notice

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

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

Table of Contents

1. Introduction	4
2. Deterministic Port Ranges	5
2.1. Stability and Load-Balancing Considerations	8
2.2. IPv4 Port Utilization Efficiency	8
2.3. Planning & Dimensioning	9
2.4. Deterministic CGN Example	9
3. Additional Logging Considerations	10
4. Impact on the IPv6 Transition	11
5. IANA Considerations	11
6. Security Considerations	11
7. Acknowledgements	11
8. References	11
8.1. Normative References	11
8.2. Informative References	12
Authors' Addresses	13

1. Introduction

The world is rapidly running out of unallocated IPv4 addresses. To ensure IPv4 service continuity under the growing demands from new subscribers, devices, and service types, some ISPs will be forced to share a single public IPv4 address among multiple subscribers using techniques such as Carrier Grade Network Address Translation (CGN) [RFC6264] (e.g., NAT444 [I-D.shirasaki-nat444], DS-Lite [RFC6333], NAT64 [RFC6146]etc.). However, address sharing poses additional challenges to ISPs in responding to public safety requests or attack/abuse reports [RFC6269]. In order to respond to such requests to identify a specific user associated with an IP address, an ISP will need to map a subscriber's internal source IP address and source port with the global public IP address and source port provided by the CGN for every connection initiated by the user.

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

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

The volume of CGN logging can be reduced by assigning port ranges instead of individual ports. Using this method, only the assignment of a new port range is logged. This may massively reduce logging

volume. The log reduction may vary depending on the length of the assigned port range, whether the port range is static or dynamic, etc. This has been acknowledged in [RFC6269]:

"Address sharing solutions may mitigate these issues to some extent by pre-allocating groups of ports. Then only the allocation of the group needs to be recorded, and not the creation of every session binding within that group. There are trade-offs to be made between the sizes of these port groups, the ratio of public addresses to subscribers, whether or not these groups timeout, and the impact on logging requirements and port randomization security (RFC6056) [RFC6056]."

However, the existing solution still poses an impact on ISPs and public safety officials for logging and searching. Instead, CGNs could be designed and/or configured to deterministically map internal addresses to {external address + port range} in such a way as to be able to algorithmically calculate the mapping. Only inputs and configuration of the algorithm need to be logged. This approach reduces both logging volume and subscriber identification times.

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

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

2. Deterministic Port Ranges

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

then process it to identify a subscriber.

The algorithmic mapping can be expressed as:

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

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

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

Deterministic Port Range allocation requires configuration of the following variables:

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

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

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

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

The algorithm is not designed to retrieve an internal host among those sharing the same internal IP address (e.g., in a DS-Lite context, only an IPv6 address/prefix can be retrieved using the algorithm while the internal IPv4 address used for the encapsulated IPv4 datagram is lost).

Several address assignment algorithms are possible. Using predefined algorithms, such as those that follow, simplifies the process of reversing the algorithm when needed. However, additional algorithms can also be supported. Subscribers could be restricted to ports from a single IPv4 address, or could be allocated ports across all addresses in a pool, for example. The following algorithms and corresponding values of A are as follow:

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

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

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

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

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

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

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

The CGN then reserves ports as follows:

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

3. When a subscriber initiates a connection, the CGN creates a translation mapping between the subscriber's inside local IP address/port and the CGN outside global IP address/port. The CGN MUST use one of the ports allocated in step 2 for the translation as long as such ports are available. The CGN MUST use the preallocated port range from step 2 for Port Control Protocol (PCP, [I-D.ietf-pcp-base]) reservations as long as such ports are available. While the CGN maintains its mapping table, it need not generate a log entry for translation mappings created in this step.
4. The CGN will have a pool of ports left for dynamic assignment. If a subscriber uses more than the range of ports allocated in step 2 (but fewer than the configured maximum ports M), the CGN uses a port from the dynamic assignment range for such a connection or for PCP reservations. The CGN MUST log dynamically assigned ports to facilitate subscriber-to-address mapping. The CGN SHOULD manage dynamic ports as described in [I-D.tsou-behave-natx4-log-reduction].
5. Configuration of reserved ports (e.g., system ports) is left to operator configuration.

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

2.1. Stability and Load-Balancing Considerations

Using the procedure defined in this document assumes a deterministic distribution of customers among deployed CGN devices. Balancing the traffic among several CGNs based on their actual load may not be supported because of the potential conflict of enforced algorithmic mapping rule. When CGN redundancy group is used, the same mapping rule, including in particular the external IP address, MUST be used. Furthermore, traffic oscillation MUST be avoided (because, unless state synchronization is used, the actual NAT state may not be instantiated in the redundancy group).

2.2. IPv4 Port Utilization Efficiency

For Service Providers requiring an aggressive address sharing ration, the use of the algorithmic mapping may impact the efficiency of the address sharing. Using a dynamic port range scheme, dynamic port assignment or a mix of static mapping and dynamic port assignment is more suitable for those SPs.

2.3. Planning & Dimensioning

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

2.4. Deterministic CGN Example

To illustrate the use of deterministic NAT, let's consider a simple example. The operator configures an inside address range (I) of 100.64.0.0/28 and outside address (O) of 203.0.113.1. The dynamic address pool factor (D) is set to '2'. Thus, the total compression ratio is $1:(14+2) = 1:16$. Only the system ports (e.g. ports < 1024) are reserved. This configuration causes the CGN to preallocate $((65536-1024)/16 =)$ 4032 TCP and 4032 UDP ports per inside IPv4 address. For the purposes of this example, let's assume that they are allocated sequentially, where 100.64.0.1 maps to 203.0.113.1 ports 1024-5055, 100.64.0.2 maps to 203.0.113.1 ports 5056-9087, etc. The dynamic port range thus contains ports 57472-65535 (port allocation illustrated in the table below). Finally, the maximum ports/subscriber is set to 5040.

Inside Address / Pool	Outside Address & Port
Reserved	203.0.113.1:0-1023
100.64.0.1	203.0.113.1:1024-5055
100.64.0.2	203.0.113.1:5056-9087
100.64.0.3	203.0.113.1:9088-13119
100.64.0.4	203.0.113.1:13120-17151
100.64.0.5	203.0.113.1:17151-21183
100.64.0.6	203.0.113.1:21184-25215
100.64.0.7	203.0.113.1:25216-29247
100.64.0.8	203.0.113.1:29248-33279
100.64.0.9	203.0.113.1:33280-37311
100.64.0.10	203.0.113.1:37312-41343
100.64.0.11	203.0.113.1:41344-45375
100.64.0.12	203.0.113.1:45376-49407
100.64.0.13	203.0.113.1:49408-53439
100.64.0.14	203.0.113.1:53440-57471
Dynamic	203.0.113.1:57472-65535

When subscriber 1 using 100.64.0.1 initiates a low volume of connections (e.g. < 4032 concurrent connections), the CGN maps the outgoing source address/port to the preallocated range. These

translation mappings are not logged.

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

If a public safety agency reports abuse from 203.0.113.1, port 2001, the operator can reverse the mapping algorithm to determine that the internal IP address subscriber 1 has been assigned generated the traffic without consulting CGN logs (by correlating the internal IP address with DHCP/PPP lease connection records). If a second abuse report comes in for 203.0.113.1, port 58204, the operator will determine that port 58204 is within the dynamic pool range, consult the log file, correlate with connection records, and determine that subscriber 2 generated the traffic (assuming that the public safety timestamp matches the operator timestamp. As noted in RFC6292 [RFC6292], accurate time-keeping (e.g., use of NTP or Simple NTP) is vital).

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

3. Additional Logging Considerations

In order to be able to identify a subscriber based on observed external IPv4 address, port, and timestamp, an operator needs to know how the CGN was configured with regards to internal and external IP addresses, dynamic address pool factor, maximum ports per user, and reserved port range at any given time. Therefore, the CGN **MUST** generate a log message any time such variables are changed. Also, the CGN **SHOULD** generate such a log message once per day to facilitate quick identification of the relevant configuration in the event of an abuse notification.

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

[Wed Oct 11 14:32:52 2000]:100.64.0.0:28:203.0.113.0:32:2:5040:0:1-1023,5004,5060.

4. Impact on the IPv6 Transition

The solution described in this document is applicable to Carrier Grade NAT transition technologies (e.g. NAT444, DS-Lite, and NAT64). As discussed in [I-D.donley-nat444-impacts], the authors acknowledge that native IPv6 will offer subscribers a better experience than CGN. However, many CPE devices only support IPv4. Likewise, as of July 2012, only approximately 4% of the top 1 million websites were available using IPv6. Accordingly, deterministic CGN should in no way be understood as making CGN a replacement for IPv6 service. The authors encourage device manufacturers to consider [RFC6540] and include IPv6 support. In the interim, however, CGN has already been deployed in some ISP networks. Deterministic CGN will provide ISPs with the ability to quickly respond to public safety requests without requiring excessive infrastructure, operations, and bandwidth to support per-connection logging.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

The security considerations applicable to NAT operation for various protocols as documented in, for example, RFC 4787 [RFC4787] and RFC 5382 [RFC5382] also apply to this document.

7. Acknowledgements

The authors would like to thank the following people for their feedback: Bobby Flaim, Lee Howard, Wes George, Jean-Francois Tremblay, Mohammed Boucadair, Alain Durand, David Miles, Andy Anchev.

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.

- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC6264] Jiang, S., Guo, D., and B. Carpenter, "An Incremental Carrier-Grade NAT (CGN) for IPv6 Transition", RFC 6264, June 2011.
- [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.

8.2. Informative References

- [I-D.donley-nat444-impacts]
Donley, C., Howard, L., Kuarsingh, V., Berg, J., and U. Colorado, "Assessing the Impact of Carrier-Grade NAT on Network Applications", draft-donley-nat444-impacts-04 (work in progress), May 2012.
- [I-D.ietf-pcp-base]
Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-25 (work in progress), May 2012.
- [I-D.miles-behave-l2nat]
Miles, D. and M. Townsley, "Layer2-Aware NAT", draft-miles-behave-l2nat-00 (work in progress), March 2009.
- [I-D.shirasaki-nat444]
Yamagata, I., Shirasaki, Y., Nakagawa, A., Yamaguchi, J., and H. Ashida, "NAT444", draft-shirasaki-nat444-05 (work in progress), January 2012.
- [I-D.tsou-behave-natx4-log-reduction]
ZOU, T., Li, W., and T. Taylor, "Port Management To Reduce Logging In Large-Scale NATs", draft-tsou-behave-natx4-log-reduction-02 (work in progress), September 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.
- [RFC6292] Hoffman, P., "Requirements for a Working Group Charter Tool", RFC 6292, June 2011.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.
- [RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", BCP 165, RFC 6335, August 2011.
- [RFC6431] Boucadair, M., Levis, P., Bajko, G., Savolainen, T., and T. Tsou, "Huawei Port Range Configuration Options for PPP IP Control Protocol (IPCP)", RFC 6431, November 2011.
- [RFC6540] George, W., Donley, C., Liljenstolpe, C., and L. Howard, "IPv6 Support Required for All IP-Capable Nodes", BCP 177, RFC 6540, April 2012.

Authors' Addresses

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

Email: c.donley@cablelabs.com

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

Email: c.grundemann@cablelabs.com

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

Email: v.sarawat@cablelabs.com

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

Email: k.sundaresan@cablelabs.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: June 14, 2015

C. Donley
CableLabs
C. Grundemann
Internet Society
V. Sarawat
K. Sundaresan
CableLabs
O. Vautrin
Juniper Networks
December 11, 2014

Deterministic Address Mapping to Reduce Logging in Carrier Grade NAT
Deployments
draft-donley-behave-deterministic-cgn-09

Abstract

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

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute

working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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

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

Copyright Notice

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

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

Table of Contents

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

1. Introduction

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

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

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

approached the authors to identify a way to reduce this impact while improving abuse response.

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

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

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

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

2. Deterministic Port Ranges

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

operator from having to transport and store massive amounts of session data from the CGN and then process it to identify a subscriber.

The algorithmic mapping can be expressed as:

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

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

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

Deterministic Port Range allocation requires configuration of the following variables:

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

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

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

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

The algorithm is not designed to retrieve an internal host among those sharing the same internal IP address (e.g., in a DS-Lite context, only an IPv6 address/prefix can be retrieved using the

algorithm while the internal IPv4 address used for the encapsulated IPv4 datagram is lost).

Several address assignment algorithms are possible. Using predefined algorithms, such as those that follow, simplifies the process of reversing the algorithm when needed. However, the CGN MAY support additional algorithms. Also, the CGN is not required to support all algorithms described below. Subscribers could be restricted to ports from a single IPv4 address, or could be allocated ports across all addresses in a pool, for example. The following algorithms and corresponding values of A are as follow:

- 0: Sequential (e.g. the first block goes to address 1, the second block to address 2, etc.)
- 1: Staggered (e.g. for every n between 0 and $((65536-R)/(C+D))-1$, address 1 receives ports $n*C+R$, address 2 receives ports $(1+n)*C+R$, etc.)
- 2: Round robin (e.g. the subscriber receives the same port number across a pool of external IP addresses. If the subscriber is to be assigned more ports than there are in the external IP pool, the subscriber receives the next highest port across the IP pool, and so on. Thus, if there are 10 IP addresses in a pool and a subscriber is assigned 1000 ports, the subscriber would receive a range such as ports 2000-2099 across all 10 external IP addresses).
- 3: Interlaced horizontally (e.g. each address receives every Cth port spread across a pool of external IP addresses).
- 4: Cryptographically random port assignment (Section 2.2 of RFC6431 [RFC6431]). If this algorithm is used, the Service Provider needs to retain the keying material and specific cryptographic function to support reversibility.
- 5: Vendor-specific. Other vendor-specific algorithms may also be supported.

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

The CGN then reserves ports as follows:

1. The CGN removes reserved ports (R) from the port candidate list (e.g., 0-1023 for TCP and UDP). At a minimum, the CGN SHOULD remove system ports (RFC6335) [RFC6335] from the port candidate list reserved for deterministic assignment.

2. The CGN calculates the total compression ratio (C+D), and allocates $1/(C+D)$ of the available ports to each internal IP address. Specific port allocation is determined by the algorithm (A) configured on the CGN. Any remaining ports are allocated to the dynamic pool.

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

3. When a subscriber initiates a connection, the CGN creates a translation mapping between the subscriber's inside local IP address/port and the CGN outside global IP address/port. The CGN MUST use one of the ports allocated in step 2 for the translation as long as such ports are available. The CGN SHOULD allocate ports randomly within the port range assigned by the deterministic algorithm. This is to increase subscriber privacy. The CGN MUST use the preallocated port range from step 2 for Port Control Protocol (PCP, [RFC6887]) reservations as long as such ports are available. While the CGN maintains its mapping table, it need not generate a log entry for translation mappings created in this step.
4. If $D > 0$, the CGN will have a pool of ports left for dynamic assignment. If a subscriber uses more than the range of ports allocated in step 2 (but fewer than the configured maximum ports M), the CGN assigns a block of ports from the dynamic assignment range for such a connection or for PCP reservations. The CGN MUST log dynamically assigned port blocks to facilitate subscriber-to-address mapping. The CGN SHOULD manage dynamic ports as described in [I-D.tsou-behave-natx4-log-reduction].
5. Configuration of reserved ports (e.g., system ports) is left to operator configuration.

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

2.1. IPv4 Port Utilization Efficiency

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

2.2. Planning & Dimensioning

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

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

2.3. Deterministic CGN Example

To illustrate the use of deterministic NAT, let's consider a simple example. The operator configures an inside address range (I) of 198.51.100.0/28 [RFC6598] and outside address (O) of 192.0.2.1. The dynamic address pool factor (D) is set to '2'. Thus, the total compression ratio is $1:(14+2) = 1:16$. Only the system ports (e.g. ports < 1024) are reserved (R). This configuration causes the CGN to preallocate $((65536-1024)/16 =)$ 4032 TCP and 4032 UDP ports per inside IPv4 address. For the purposes of this example, let's assume that they are allocated sequentially, where 198.51.100.1 maps to 192.0.2.1 ports 1024-5055, 198.51.100.2 maps to 192.0.2.1 ports 5056-9087, etc. The dynamic port range thus contains ports 57472-65535 (port allocation illustrated in the table below). Finally, the maximum ports/subscriber is set to 5040.

Inside Address / Pool	Outside Address & Port
Reserved	192.0.2.1:0-1023
198.51.100.1	192.0.2.1:1024-5055
198.51.100.2	192.0.2.1:5056-9087
198.51.100.3	192.0.2.1:9088-13119
198.51.100.4	192.0.2.1:13120-17151
198.51.100.5	192.0.2.1:17152-21183
198.51.100.6	192.0.2.1:21184-25215
198.51.100.7	192.0.2.1:25216-29247
198.51.100.8	192.0.2.1:29248-33279
198.51.100.9	192.0.2.1:33280-37311
198.51.100.10	192.0.2.1:37312-41343
198.51.100.11	192.0.2.1:41344-45375
198.51.100.12	192.0.2.1:45376-49407
198.51.100.13	192.0.2.1:49408-53439
198.51.100.14	192.0.2.1:53440-57471
Dynamic	192.0.2.1:57472-65535

When subscriber 1 using 198.51.100.1 initiates a low volume of connections (e.g. < 4032 concurrent connections), the CGN maps the outgoing source address/port to the preallocated range. These translation mappings are not logged.

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

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

records, and determine that subscriber 2 generated the traffic (assuming that the public safety timestamp matches the operator timestamp. As noted in RFC6292 [RFC6292], accurate time-keeping (e.g., use of NTP or Simple NTP) is vital).

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

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

3. Additional Logging Considerations

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

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

```
[Wed Oct 11 14:32:52
2000]:198.51.100.0:28:192.0.2.0:32:2:5040:0:1-1023,5004,5060.
```

3.1. Failover Considerations

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

CGN 2, an operator can identify the subscriber based on outside source address and port information.

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

4. Impact on the IPv6 Transition

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

5. Privacy Considerations

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

6. IANA Considerations

This document makes no request of IANA.

7. Security Considerations

The security considerations applicable to NAT operation for various protocols as documented in, for example, RFC 4787 [RFC4787] and RFC 5382 [RFC5382] also apply to this document.

Note that with the possible exception of cryptographically-based port allocations, attackers could reverse-engineer algorithmically-derived port allocations to either target a specific subscriber or to spoof traffic to make it appear to have been generated by a specific subscriber. However, this is exactly the same level of security that the subscriber would experience in the absence of CGN. CGN is not intended to provide additional security by obscurity.

8. Acknowledgements

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

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC6264] Jiang, S., Guo, D., and B. Carpenter, "An Incremental Carrier-Grade NAT (CGN) for IPv6 Transition", RFC 6264, June 2011.
- [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.

9.2. Informative References

- [I-D.miles-behave-l2nat] Miles, D. and M. Townsley, "Layer2-Aware NAT", draft-miles-behave-l2nat-00 (work in progress), March 2009.

- [I-D.shirasaki-nat444]
Yamagata, I., Shirasaki, Y., Nakagawa, A., Yamaguchi, J.,
and H. Ashida, "NAT444", draft-shirasaki-nat444-06 (work
in progress), July 2012.
- [I-D.sivakumar-behave-nat-logging]
Sivakumar, S. and R. Penno, "IPFIX Information Elements
for logging NAT Events", draft-sivakumar-behave-nat-
logging-06 (work in progress), January 2013.
- [I-D.tsou-behave-natx4-log-reduction]
Tsou, T., Li, W., Taylor, T., and J. Huang, "Port
Management To Reduce Logging In Large-Scale NATs", draft-
tsou-behave-natx4-log-reduction-04 (work in progress),
July 2013.
- [RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful
NAT64: Network Address and Protocol Translation from IPv6
Clients to IPv4 Servers", RFC 6146, April 2011.
- [RFC6292] Hoffman, P., "Requirements for a Working Group Charter
Tool", RFC 6292, June 2011.
- [RFC6302] Durand, A., Gashinsky, I., Lee, D., and S. Sheppard,
"Logging Recommendations for Internet-Facing Servers", BCP
162, RFC 6302, June 2011.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-
Stack Lite Broadband Deployments Following IPv4
Exhaustion", RFC 6333, August 2011.
- [RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S.
Cheshire, "Internet Assigned Numbers Authority (IANA)
Procedures for the Management of the Service Name and
Transport Protocol Port Number Registry", BCP 165, RFC
6335, August 2011.
- [RFC6431] Boucadair, M., Levis, P., Bajko, G., Savolainen, T., and
T. Tsou, "Huawei Port Range Configuration Options for PPP
IP Control Protocol (IPCP)", RFC 6431, November 2011.
- [RFC6598] Weil, J., Kuarsingh, V., Donley, C., Liljenstolpe, C., and
M. Azinger, "IANA-Reserved IPv4 Prefix for Shared Address
Space", BCP 153, RFC 6598, April 2012.
- [RFC6887] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P.
Selkirk, "Port Control Protocol (PCP)", RFC 6887, April
2013.

[RFC7021] Donley, C., Howard, L., Kuarsingh, V., Berg, J., and J. Doshi, "Assessing the Impact of Carrier-Grade NAT on Network Applications", RFC 7021, September 2013.

Authors' Addresses

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

Email: c.donley@cablelabs.com

Chris Grundemann
Internet Society
Denver, CO
US

Email: cgrundemann@gmail.com

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

Email: v.sarawat@cablelabs.com

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

Email: k.sundaresan@cablelabs.com

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

Email: olivier@juniper.net

Internet Engineering Task Force
Internet-Draft
Intended status: Best Current Practice
Expires: April 3, 2015

W. George
L. Howard
Time Warner Cable
September 30, 2014

IPv6 Support Within IETF work
draft-george-ipv6-support-03

Abstract

This document recommends that the IETF formally require its standards work to be IP version agnostic or to explicitly include support for IPv6, with some exceptions, to ensure that it is possible to operate without dependencies on IPv4.

Status of This Memo

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

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

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

This Internet-Draft will expire on April 3, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. Requirements Language	3
2. IPv6-only operation	3
2.1. Functional Parity with IPv4	3
2.2. IPv4 Sunset	3
3. Requirements and Recommendations	4
4. Acknowledgements	5
5. IANA Considerations	5
6. Security Considerations	5
7. References	5
7.1. Normative References	5
7.2. Informative References	5
Authors' Addresses	5

1. Introduction

[RFC6540] gives guidance to implementers that in order to ensure interoperability and proper function after IPv4 exhaustion, IP-capable devices need to support IPv6, and cannot be reliant on IPv4, because global IPv4 exhaustion creates many circumstances where the use of IPv6 will no longer be optional. Since this is an IETF Best Current Practice recommendation, it is imperative that the results of IETF efforts enable implementers to follow that recommendation. This document provides recommendations and guidance as to how IETF itself should handle future work as it relates to Internet Protocol versions.

When considering support for IPv4 vs IPv6 within IETF work, the general goal is to provide tools that enable networks and applications to operate seamlessly in any combination of IPv4-only, dual-stack, or IPv6-only as their needs dictate. However, as the IPv4 to IPv6 transition continues, it will become increasingly difficult to ensure interoperability and backward compatibility with IPv4-only networks and applications. As IPv6 deployment grows, IETF will naturally focus on features and protocols that enhance and extend IPv6, along with continuing work on items that are IP version agnostic. New features and protocols will not typically be introduced for use as IPv4-only. However, as of this document's writing, there is no formal requirement for all IETF work to support IPv6, either implicitly by being network-layer agnostic or explicitly by having an IPv6-specific implementation.

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. IPv6-only operation

At this document's writing, IPv6 has seen significant deployment. Most of these deployments are dual-stack, with IPv4 and IPv6 coexisting on the same networks. However, dual-stack is a waypoint in the transition from IPv4 to IPv6. The eventual end state is networks and end points that are IPv6-only. Some operators may take a long time to turn off IPv4, if they ever do, but the IETF MUST ensure that its standards can be deployed by even the first operators to turn off IPv4. Problems (and solutions) need to be identified before they are encountered by the earliest adopters.

2.1. Functional Parity with IPv4

In order for IPv6-only operation to be realistic, IPv6 MUST have at least functional parity with IPv4. "Functional parity" means that any function that IPv4 enables MUST also be enabled by IPv6. This does not mean that every feature that exists in IPv4 will exist in IPv6; different features may enable the same function. For instance, IPv4 supports some features that are no longer in use. In some cases it has not been practical to remove them in IPv4, or even to declare them historic, but it is unnecessary to carry them forward into IPv6. IPv6 also eliminates the need for some features that exist in IPv4; no effort to create unneeded features is required. Functional parity does not mean that all functions in IPv6 must also be possible in IPv4. Indeed, with IPv6 becoming the predominant protocol, new functionality should be developed in IPv6, and IETF effort SHOULD NOT be spent retrofitting features into the legacy protocol.

2.2. IPv4 Sunset

Somewhat distinct from identifying the needed features for IPv6-only functional parity is the effort to identify what is necessary to disable or sunset IPv4 in a given network. Since many of the protocols in use today were designed to be fault-tolerant and very robust, actually removing them from a network once they are no longer needed is sometimes complex. Many implementations may not even have "off switches" because the assumption was that they would never be switched off in a normal network implementation. The Sunset4 Working Group was chartered to address these issues:

"The Working Group will point out specific areas of concern, provide recommendations, and standardize protocols that facilitate the graceful "sunsetting" of the IPv4 Internet in areas where IPv6 has been deployed. This includes the act of shutting down IPv4 itself, as well as the ability of IPv6-only portions of the Internet to continue to connect with portions of the Internet that remain IPv4-only. ... Disabling IPv4 in applications, hosts, and networks is new territory for much of the Internet today, and it is expected that problems will be uncovered including those related to basic IPv4 functionality, interoperability, as well as potential security concerns. The working group will report on common issues, provide recommendations, and, when necessary, protocol extensions in order to facilitate disabling IPv4 in networks where IPv6 has been deployed."

3. Requirements and Recommendations

Ongoing focus is required to ensure that future IETF work is capable of IPv6-only operation. This attention may take the form of IESG evaluation, individual document reviews, or future WG charters. Due to the existing operational base of IPv4, it is not realistic to completely bar further work on IPv4 within the IETF at this time, nor to formally declare it historic. Until the time when IPv4 is no longer in wide use and/or declared historic, the IETF needs to continue to update IPv4-only protocols and features for vital operational or security issues. Similarly, the IETF needs to complete the work related to IPv4-to-IPv6 transition tools for migrating more traffic to IPv6. As the transition to IPv6-capable networks accelerates, it is also likely that some changes may be necessary in IPv4 protocols to facilitate decommissioning IPv4 in a way that does not create unacceptable impact to applications or users. These sorts of IPv4-focused activities, in support of security, transition, and decommissioning, should continue, accompanied by problem statements based on operational experience. Generally the focus should move away from IPv4-only work.

The IESG SHOULD review working group charters to ensure that work will be capable of operating without IPv4, except in cases of IPv4 security, transition, and decommissioning work.

IETF SHOULD make updates to IPv4 protocols and features to facilitate IPv4 decommissioning

IETF work SHOULD explicitly support IPv6 or SHOULD be IP version agnostic (because it is implemented above the network layer), except IPv4-specific transition or address-sharing technologies.

IETF SHOULD NOT initiate new IPv4 extension technology development.

IETF work SHOULD function completely on IPv6-only nodes and networks, unless consensus exists that it is unnecessary to use a given feature or protocol on IPv6-only networks.

IETF SHOULD identify and update IPv4-only protocols and applications to support IPv6 unless consensus exists that it is unnecessary for a given feature or protocol.

4. Acknowledgements

Thanks to the following people for their comments: Jari Arkko, Ralph Droms, Scott Brim, Margaret Wasserman, Brian Haberman. Thanks also to Randy Bush, Mark Townsley, and Dan Wing for their discussion in IntArea WG at IETF 81 in Taipei, TW regarding transition technologies, IPv4 life extension, and IPv6 support.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

This document generates no new security considerations because it is not defining a new protocol. As existing work is analyzed for its ability to operate properly on IPv6-only networks, new security issues may be identified.

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.

7.2. Informative References

[RFC6540] George, W., Donley, C., Liljenstolpe, C., and L. Howard, "IPv6 Support Required for All IP-Capable Nodes", BCP 177, RFC 6540, April 2012.

Authors' Addresses

Wesley George
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA 20171
US

Phone: +1 703-561-2540
Email: wesley.george@twcable.com

Lee Howard
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA 20171
US

Phone: +1-703-345-3513
Email: lee.howard@twcable.com

Internet Engineering Task Force
Internet-Draft
Intended status: BCP
Expires: January 4, 2013

M. Mawatari
Japan Internet Exchange Co.,Ltd.
M. Kawashima
NEC AccessTechnica, Ltd.
C. Byrne
T-Mobile USA
July 3, 2012

464XLAT: Combination of Stateful and Stateless Translation
draft-ietf-v6ops-464xlat-05

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 4, 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. Requirements Language	3
3. Terminology	3
4. Motivation and Uniqueness of 464XLAT	4
5. Network Architecture	4
5.1. Wireline Network Architecture	4
5.2. Wireless 3GPP Network Architecture	5
6. Applicability	6
6.1. Wireline Network Applicability	6
6.2. Wireless 3GPP Network Applicability	7
7. Implementation Considerations	7
7.1. IPv6 Address Format	7
7.2. IPv4/IPv6 Address Translation Chart	7
7.2.1. Case of enabling only stateless XLATE on CLAT	7
7.2.2. Case of enabling NAT44 and stateless XLATE on CLAT	9
7.3. IPv6 Prefix Handling	11
7.3.1. Case of enabling only stateless XLATE on CLAT	11
7.3.2. Case of enabling NAT44 and stateless XLATE on CLAT	11
7.4. Traffic Treatment Scenarios	12
7.5. DNS Proxy Implementation	12
7.6. CLAT in a Gateway	12
7.7. CLAT to CLAT communications	12
8. Deployment Considerations	13
9. Security Considerations	13
10. IANA Considerations	13
11. Acknowledgements	14
12. References	14
12.1. Normative References	14
12.2. Informative References	14
Appendix A. Examples of IPv4/IPv6 Address Translation	16
Authors' Addresses	19

1. Introduction

The IANA unallocated IPv4 address pool was exhausted on February 3, 2011. Each RIR's unallocated IPv4 address pool will exhaust in the near future. It will be difficult for many networks to assign IPv4 addresses to end users, despite substantial IP connectivity growth required for fast growing edge networks.

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

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

By combining 464XLAT with BIH [RFC6535], it is also possible to provide single IPv4 to IPv6 translation service, which will be needed in the future case of IPv6-only servers and peers to be reached from legacy IPv4-only hosts across IPv6-only networks.

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. Terminology

- PLAT: PLAT is Provider side translator(XLAT) that complies with [RFC6146]. It translates N:1 global IPv6 packets to global IPv4 packets, and vice versa.
- CLAT: CLAT is Customer side translator(XLAT) that complies with [RFC6145]. It algorithmically translates 1:1 private IPv4 packets to global IPv6 packets, and vice versa. The CLAT function is applicable to a router or an end-node such as a mobile phone. CLAT SHOULD perform router function to facilitate packets forwarding through the stateless translation even if it is an end-node. In the case where the access network does not allow for a dedicated IPv6 prefix for translation, a NAT44 SHOULD be used between the router function and the stateless translator function. The CLAT as a common home router or 3G router is expected to perform gateway functions such as DHCP server and DNS proxy for local clients. The CLAT does not comply with the sentence "Both IPv4-translatable IPv6 addresses and IPv4-converted IPv6 addresses SHOULD use the same prefix." that is described on Section 3.3 in [RFC6052] due to using different IPv6 prefixes for CLAT-side and PLAT-side IPv4 addresses.

4. Motivation and Uniqueness of 464XLAT

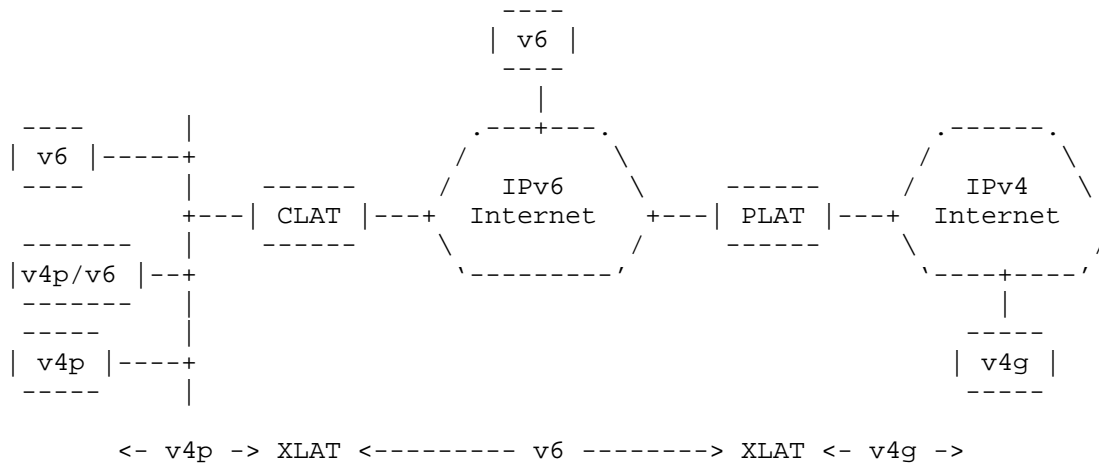
1. Minimal IPv4 resource requirements, maximum IPv4 efficiency through statistical multiplexing
2. No new protocols required, quick deployment
3. IPv6-only networks are simpler and therefore less expensive to operate

5. Network Architecture

464XLAT architecture is shown in the following figure.

5.1. Wireline Network Architecture

The private IPv4 host on this diagram can reach global IPv4 hosts via translation on both CLAT and PLAT. On the other hand, the IPv6 host can reach other IPv6 hosts on the Internet directly without translation. This means that the CPE can not only have the function of CLAT but also the function of IPv6 native router for IPv6 native traffic.

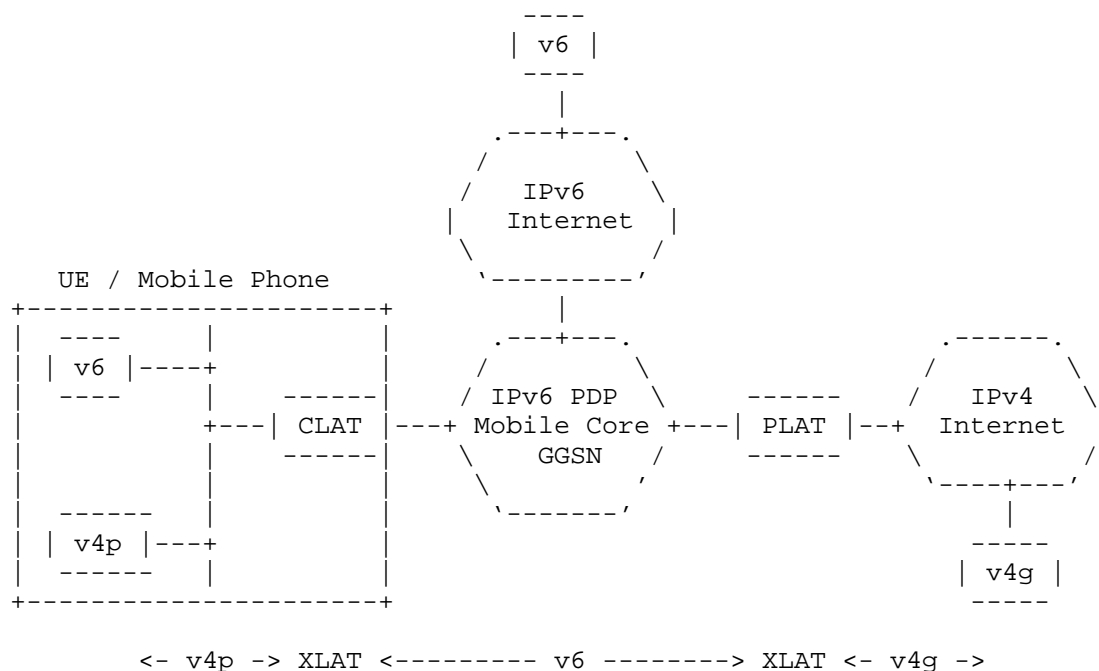


v6 : Global IPv6
v4p : Private IPv4
v4g : Global IPv4

Figure 1: Wireline Network Topology

5.2. Wireless 3GPP Network Architecture

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



```
v6  : Global IPv6
v4p : Private IPv4
v4g : Global IPv4
```

Figure 2: Wireless 3GPP Network Topology

6. Applicability

6.1. Wireline Network Applicability

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

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

Incidentally, the effectiveness of 464XLAT was confirmed in the WIDE camp Spring 2012. The result is described in

[I-D.hazeyama-widencamp-ipv6-only-experience].

6.2. Wireless 3GPP Network Applicability

The vast majority of mobile networks are compliant to Pre-Release 9 3GPP standards. In Pre-Release 9 3GPP networks, GSM and UMTS networks must signal and support both IPv4 and IPv6 Packet Data Protocol (PDP) attachments to access IPv4 and IPv6 network destinations [RFC6459]. Since there are 2 PDPs required to support 2 address families, this is double the number of PDPs required to support the status quo of 1 address family, which is IPv4.

464XLAT in combination with stateful translation [RFC6146] and DNS64 [RFC6147] allows 85% of the Android applications to continue to work with single translation or native IPv6 access. For the remaining 15% of applications that require IPv4 connectivity, the CLAT function on the UE provides a private IPv4 address and IPv4 default-route on the host for the applications to reference and bind to. Connections sourced from the IPv4 interface are immediately routed to the CLAT function and passed to the IPv6-only mobile network, destined to the PLAT. In summary, the UE has the CLAT function that does a stateless translation [RFC6145], but only when required. The mobile network has a PLAT that does stateful translation [RFC6146].

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

7. Implementation Considerations

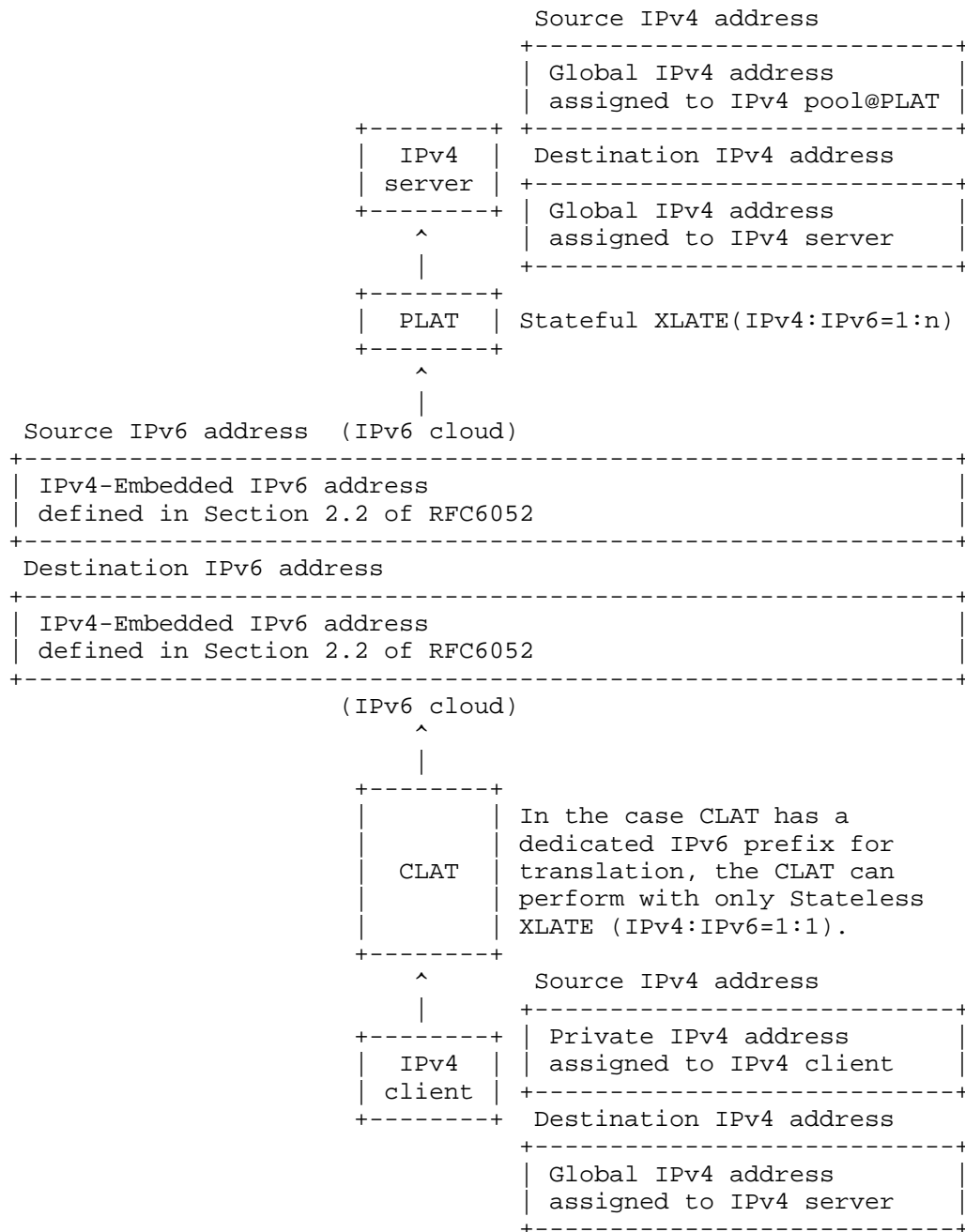
7.1. IPv6 Address Format

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

7.2. IPv4/IPv6 Address Translation Chart

7.2.1. Case of enabling only stateless XLATE on CLAT

This case should be used when a prefix delegation mechanism such as DHCPv6-PD [RFC3633] is available to assign a dedicated translation prefix to the CLAT.



Case of enabling only stateless XLATE on CLAT

7.2.2. Case of enabling NAT44 and stateless XLATE on CLAT

This case should be used when a prefix delegation mechanism is not available to assign a dedicated translation prefix to the CLAT. In this case, NAT44 SHOULD be used so that all IPv4 source addresses are mapped to a single IPv6 address.



Case of enabling NAT44 and stateless XLATE on CLAT

7.3. IPv6 Prefix Handling

7.3.1. Case of enabling only stateless XLATE on CLAT

From the delegated DHCPv6 [RFC3633] prefix, a /64 is dedicated to source and receive IPv6 packets associated with the stateless translation [RFC6145].

The CLAT MAY discover the Pref64::/n of the PLAT via some method such as DHCPv6 option, TR-069, DNS APL RR [RFC3123] or [I-D.ietf-behave-nat64-discovery-heuristic].

7.3.2. Case of enabling NAT44 and stateless XLATE on CLAT

In the case that DHCPv6-PD [RFC3633] is not available, the CLAT does not have dedicated IPv6 prefix for translation. If the CLAT does not have a dedicated IPv6 prefix for translation, the CLAT can perform with NAT44 and stateless translation [RFC6145].

Incoming source IPv4 packets from the LAN of [RFC1918] addresses are NAT44 to the CLAT IPv4 host address. Then, the CLAT will do a stateless translation [RFC6145] so that the IPv4 packets from the CLAT IPv4 host address are translated to the CLAT WAN IPv6 address as described in [RFC6052].

Its subnet prefix is made of the delegated prefix, completed if needed to a /64 by a subnet ID = 0. Its interface ID is the 464XLAT interface ID (Section 10).

The CLAT MAY discover the Pref64::/n of the PLAT via some method such as TR-069, DNS APL RR [RFC3123] or [I-D.ietf-behave-nat64-discovery-heuristic].

7.4. Traffic Treatment Scenarios

Server	Application and Host	Traffic Treatment	Location of Translation
IPv6	IPv6	End-to-end IPv6	None
IPv4	IPv6	Stateful Translation	PLAT
IPv4	IPv4	464XLAT	PLAT/CLAT
IPv6	IPv4	Stateless Translation	CLAT

Traffic Treatment Scenarios

The above chart shows most common traffic types and traffic treatment.

7.5. DNS Proxy Implementation

The CLAT SHOULD implement a DNS proxy as defined in [RFC5625]. The case of an IPv4-only node behind CLAT querying an IPv4 DNS server is undesirable since it requires both stateful and stateless translation for each DNS lookup. The CLAT SHOULD set itself as the DNS server via DHCP or other means and proxy DNS queries for IPv4 and IPv6 clients. Using the CLAT enabled home router or UE as a DNS proxy is a normal consume gateway function and simplifies the traffic flow so that only IPv6 native queries are made across the access network. The CLAT SHOULD allow for a client to query any DNS server of its choice and bypass the proxy.

7.6. CLAT in a Gateway

The CLAT is a stateless translation feature which can be implemented in a common home router or mobile phone that has a mobile router feature. The router with CLAT function SHOULD provide common router services such as DHCP of [RFC1918] addresses, DHCPv6, and DNS service. The router SHOULD set itself as the DNS server advertised via DHCP or other means to the clients so that it may implement the DNS proxy function to avoid double translation of DNS request.

7.7. CLAT to CLAT communications

While CLAT to CLAT IPv4 communication may work when the client IPv4 subnets do not overlap, this traffic flow is out of scope. 464XLAT is

a hub and spoke architecture focused on enabling IPv4-only services over IPv6-only access networks.

8. Deployment Considerations

Even if the Internet access provider for consumers is different from the PLAT provider (e.g. another internet access provider), it can implement traffic engineering independently from the PLAT provider. Detailed reasons are below:

1. The Internet access provider for consumers can figure out IPv4 destination address from translated IPv6 packet header, so it can implement traffic engineering based on IPv4 destination address (e.g. traffic monitoring for each IPv4 destination address, packet filtering for each IPv4 destination address, etc.). The tunneling methods do not have such a advantage, without any deep packet inspection for processing the inner IPv4 packet of the tunnel packet.
2. If the Internet access provider for consumers can assign IPv6 prefix greater than /64 for each subscriber, this 464XLAT architecture can separate IPv6 prefix for native IPv6 packets and XLAT prefix for IPv4/IPv6 translation packets. Accordingly, it can identify the type of packets ("native IPv6 packets" and "IPv4/IPv6 translation packets"), and implement traffic engineering based on IPv6 prefix.

This 464XLAT architecture has two capabilities. One is a IPv4 -> IPv6 -> IPv4 translation for sharing global IPv4 addresses, another, if combined with BIH [RFC6535], is a IPv4 -> IPv6 translation for reaching IPv6-only servers from IPv4-only clients that can not support IPv6. IPv4-only clients must be support through the long period of global transition to IPv6.

9. Security Considerations

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

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

10. IANA Considerations

IANA is requested to reserve a Modified EUI-64 identifier for 464XLAT

according to section 2.2.2 of [RFC5342]. Its suggested value is 02-00-5E-00-00-00-00 to 02-00-5E-0F-FF-FF-FF-FF or 02-00-5E-10-00-00-00-00 to 02-00-5E-EF-FF-FF-FF-FF, depending on whether it should be taken in reserved or available values.

11. Acknowledgements

The authors would like to thank JPIX NOC members, JPIX 464XLAT trial service members, Seiichi Kawamura, Dan Drown, Brian Carpenter, Rajiv Asati, Washam Fan, Behcet Sarikaya, Jan Zorz, Tatsuya Oishi, Lorenzo Colitti, Erik Kline, Ole Troan, Maoke Chen, Gang Chen, Tom Petch, and Jouni Korhonen for their helpful comments. Special acknowledgments go to Remi Despres for his plentiful supports and suggestions, especially about using NAT44 with IANA's EUI-64 ID. We also would like to thank Fred Baker and Joel Jaeggli for their support.

12. References

12.1. Normative References

- [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.

12.2. Informative References

- [I-D.hazeyama-widencamp-ipv6-only-experience] Hazeyama, H., Hiromi, R., Ishihara, T., and O. Nakamura, "Experiences from IPv6-Only Networks with Transition Technologies in the WIDE Camp Spring 2012", draft-hazeyama-widencamp-ipv6-only-experience-01 (work in progress), March 2012.

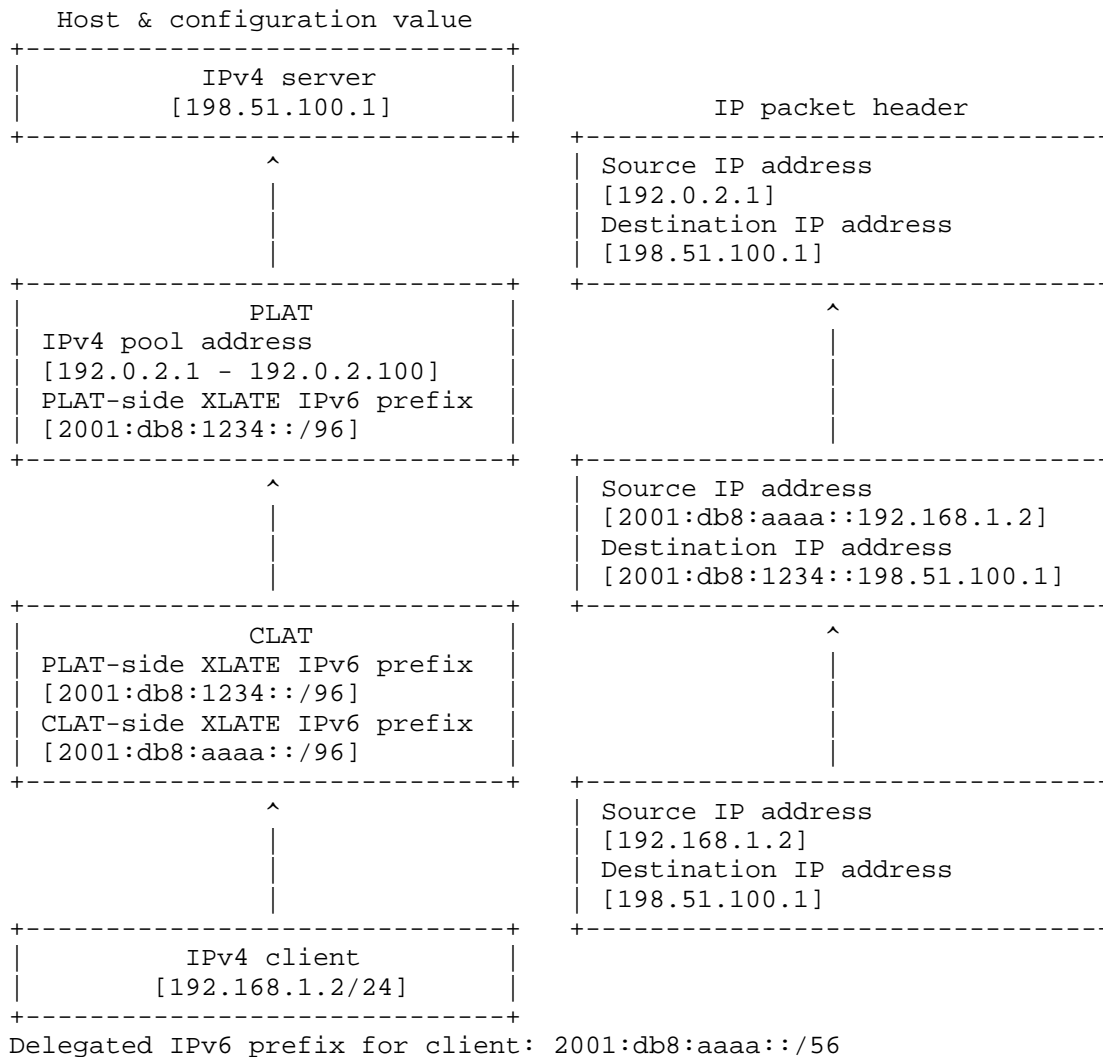
- [I-D.ietf-behave-nat64-discovery-heuristic]
Savolainen, T., Korhonen, J., and D. Wing, "Discovery of IPv6 Prefix Used for IPv6 Address Synthesis", draft-ietf-behave-nat64-discovery-heuristic-10 (work in progress), June 2012.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC3123] Koch, P., "A DNS RR Type for Lists of Address Prefixes (APL RR)", RFC 3123, June 2001.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC5342] Eastlake, D., "IANA Considerations and IETF Protocol Usage for IEEE 802 Parameters", BCP 141, RFC 5342, September 2008.
- [RFC5625] Bellis, R., "DNS Proxy Implementation Guidelines", BCP 152, RFC 5625, August 2009.
- [RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service", RFC 6092, January 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.
- [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.
- [RFC6535] Huang, B., Deng, H., and T. Savolainen, "Dual-Stack Hosts Using "Bump-in-the-Host" (BIH)", RFC 6535, February 2012.
- [TS.23203] 3GPP, "Policy and charging control architecture", 3GPP TS 23.203 10.7.0, June 2012.

Appendix A. Examples of IPv4/IPv6 Address Translation

The following are examples of IPv4/IPv6 Address Translation on the 464XLAT architecture.

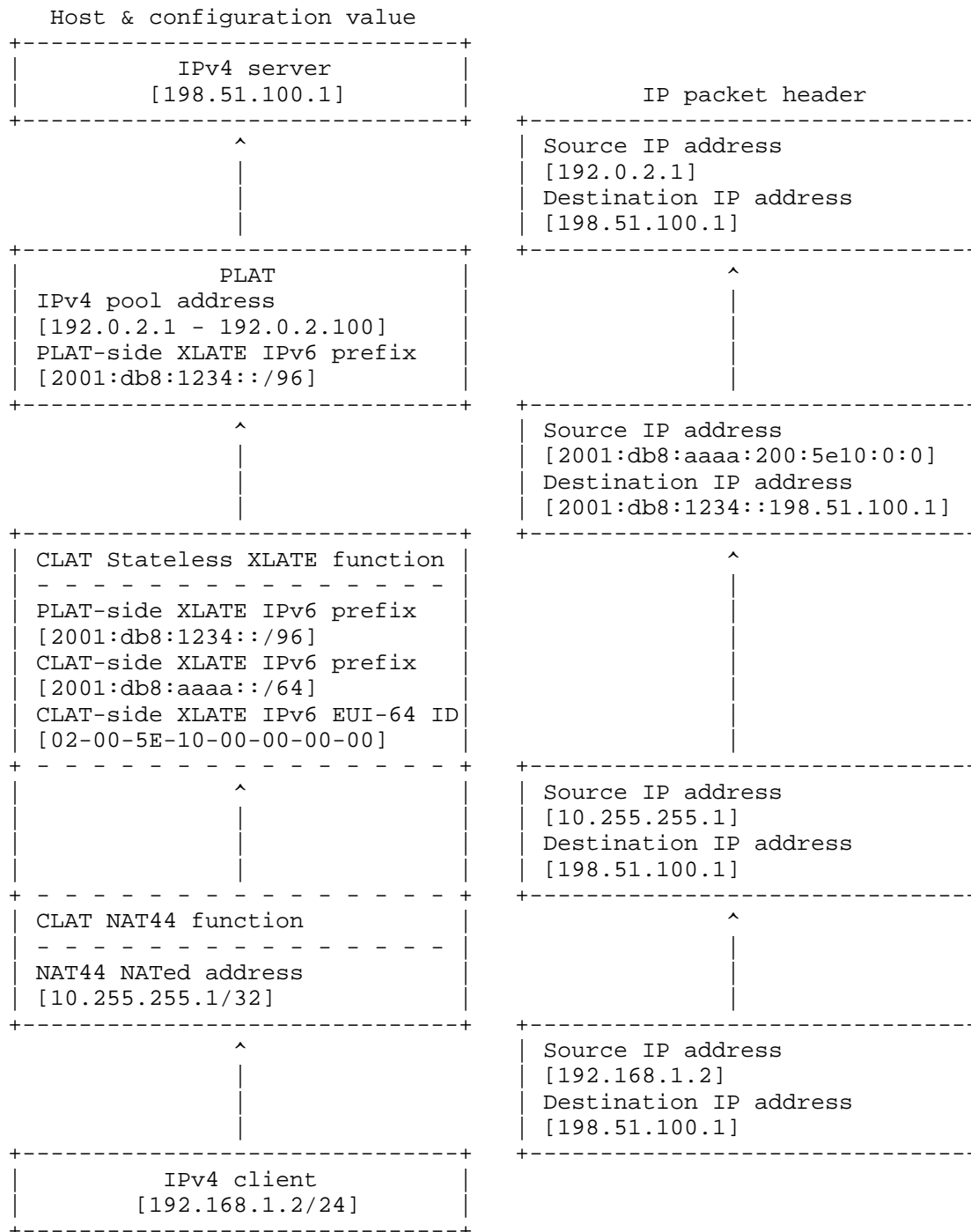
Example 1. (Case of enabling only stateless XLATE on CLAT)

In the case that IPv6 prefix greater than /64 is assigned to end users by such as DHCPv6-PD [RFC3633], only the function of Stateless XLATE should be enabled on CLAT. Because the CLAT can use dedicated a /64 from the assigned IPv6 prefix for Stateless XLATE.



Example 2. (Case of enabling NAT44 and stateless XLATE on CLAT)

In the case that IPv6 prefix /64 is assigned to end users, the function of NAT44 and Stateless XLATE should be enabled on CLAT. Because the CLAT does not have dedicated IPv6 prefix for translation.



Delegated IPv6 prefix for client: 2001:db8:aaaa::/64

Authors' Addresses

Masataka Mawatari
Japan Internet Exchange Co.,Ltd.
KDDI Otemachi Building 19F, 1-8-1 Otemachi,
Chiyoda-ku, Tokyo 100-0004
JAPAN

Phone: +81 3 3243 9579
Email: mawatari@jpix.ad.jp

Masanobu Kawashima
NEC AccessTechnica, Ltd.
800, Shimomata
Kakegawa-shi, Shizuoka 436-8501
JAPAN

Phone: +81 537 23 9655
Email: kawashimam@vx.jp.nec.com

Cameron Byrne
T-Mobile USA
Bellevue, Washington 98006
USA

Email: cameron.byrne@t-mobile.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 27, 2013

M. Mawatari
Japan Internet Exchange Co.,Ltd.
M. Kawashima
NEC AccessTechnica, Ltd.
C. Byrne
T-Mobile USA
February 23, 2013

464XLAT: Combination of Stateful and Stateless Translation
draft-ietf-v6ops-464xlat-10

Abstract

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

Status of this Memo

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

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

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

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

Copyright Notice

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

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

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

Table of Contents

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

1. Introduction

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

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

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

2. Terminology

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

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

communications between a local IPv4-only node and an IPv6-only node on the Internet.

3. Motivation and Uniqueness of 464XLAT

1. Minimal IPv4 resource requirements, maximum IPv4 efficiency through statistical multiplexing.
2. No new protocols required, quick deployment.
3. IPv6-only networks are simpler and therefore less expensive to operate than dual-stack networks.
4. Consistent native IP based monitoring, traffic engineering, and capacity planning techniques can be applied without the indirection or obfuscation of a tunnel.

4. Network Architecture

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

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

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

4.1. Wireline Network Architecture

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

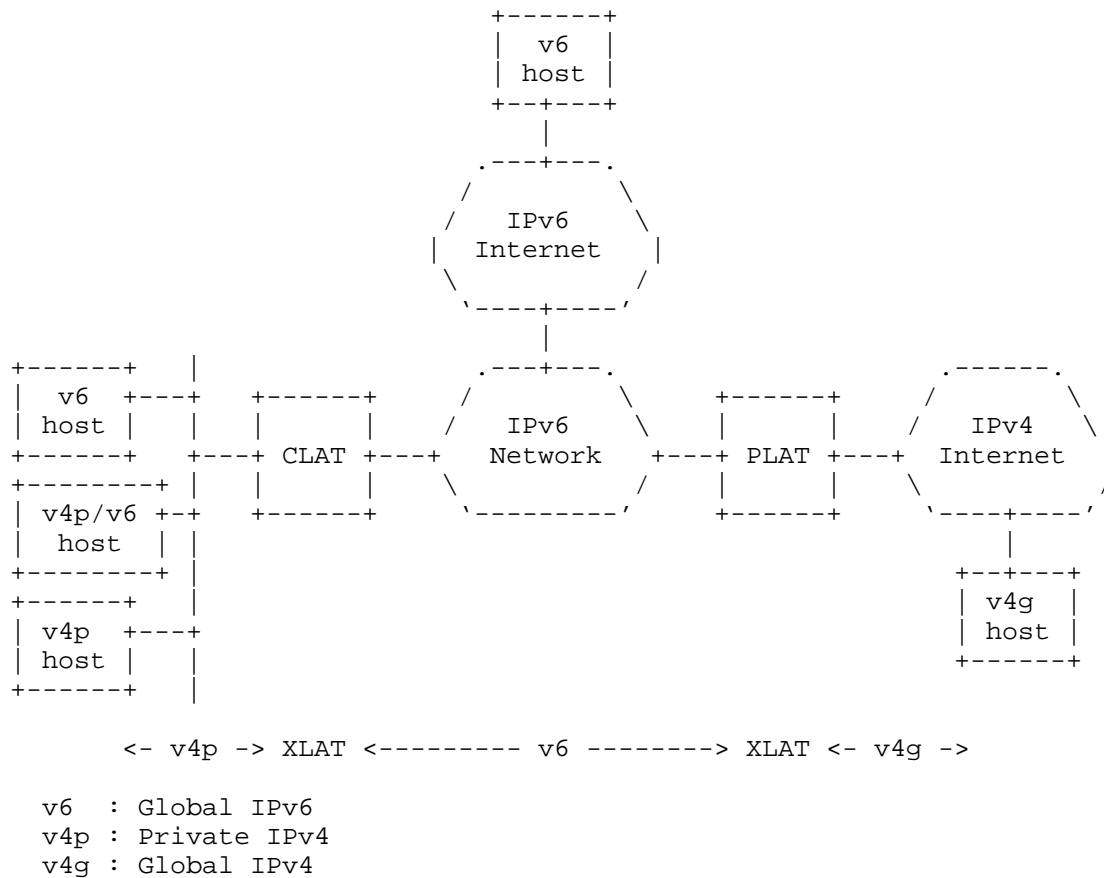


Figure 1: Wireline Network Topology

4.2. Wireless 3GPP Network Architecture

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

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

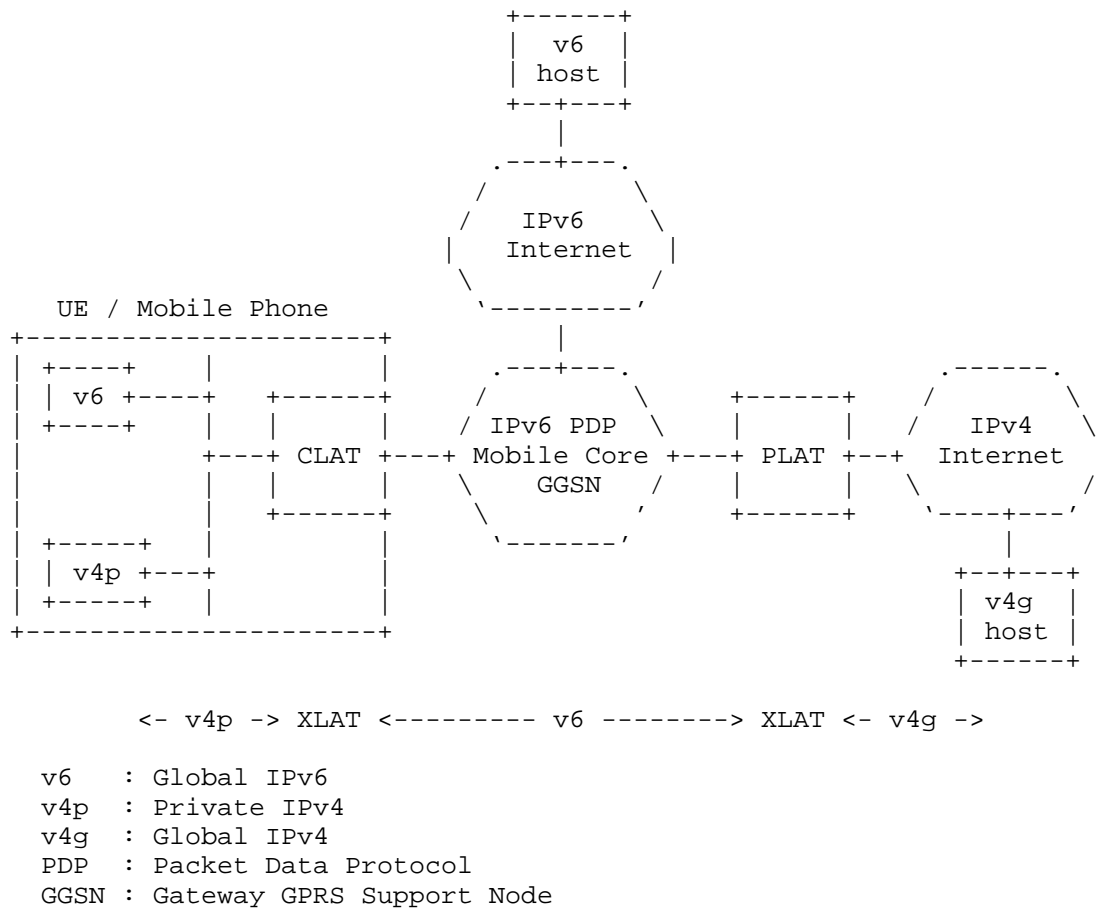


Figure 2: Wireless 3GPP Network Topology

5. Applicability

5.1. Wireline Network Applicability

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

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

highly scalable IPv6-only environment.

5.2. Wireless 3GPP Network Applicability

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

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

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

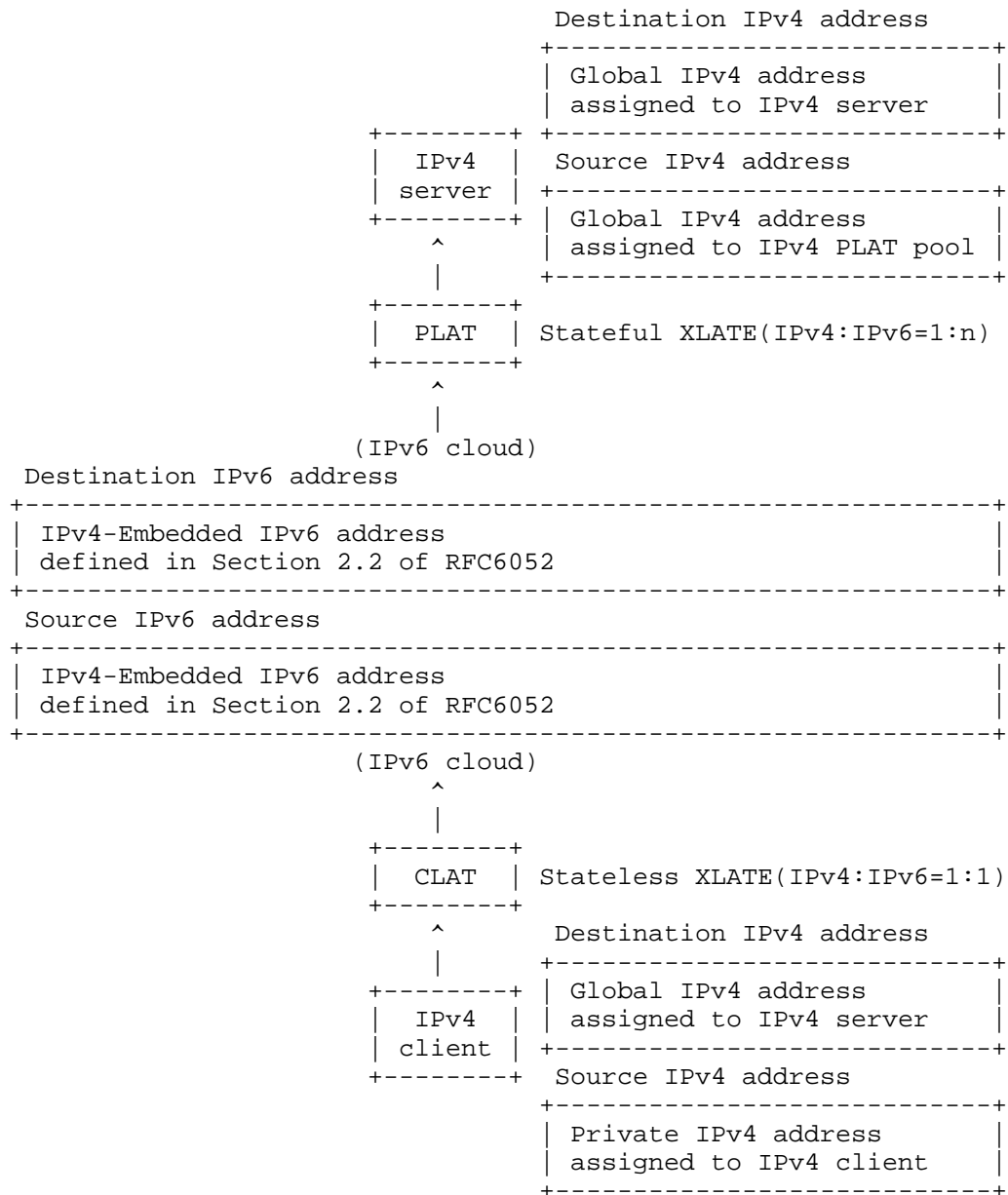
6. Implementation Considerations

6.1. IPv6 Address Format

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

6.2. IPv4/IPv6 Address Translation Chart

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



Case of enabling only stateless XLATE on CLAT

6.3. IPv6 Prefix Handling

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

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

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

6.4. DNS Proxy Implementation

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

6.5. CLAT in a Gateway

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

6.6. CLAT to CLAT communications

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

support peer-to-peer communication within a 464XLAT network.

7. Deployment Considerations

7.1. Traffic Engineering

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

1. The ISP for end users can figure out IPv4 destination address from translated IPv6 packet header, so it can implement traffic engineering based on IPv4 destination address (e.g. traffic monitoring for each IPv4 destination address, packet filtering for each IPv4 destination address, etc.). The tunneling methods do not have such an advantage, without any deep packet inspection for processing the inner IPv4 packet of the tunnel packet.
2. If the ISP for end users can assign an IPv6 prefix greater than /64 to each subscriber, this 464XLAT architecture can separate IPv6 prefix for native IPv6 packets and the XLAT prefixes for IPv4/IPv6 translation packets. Accordingly, it can identify the type of packets ("native IPv6 packets" and "IPv4/IPv6 translation packets"), and implement traffic engineering based on the IPv6 prefix.

7.2. Traffic Treatment Scenarios

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

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

Server	Application and Host	Traffic Treatment	Location of Translation
IPv6	IPv6	End-to-end IPv6	None
IPv4	IPv6	Stateful Translation	PLAT
IPv4	IPv4	464XLAT	PLAT/CLAT

Traffic Treatment Scenarios

8. Security Considerations

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

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

9. IANA Considerations

This document has no actions for IANA.

10. Acknowledgements

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

11. References

11.1. Normative References

- [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.

11.2. Informative References

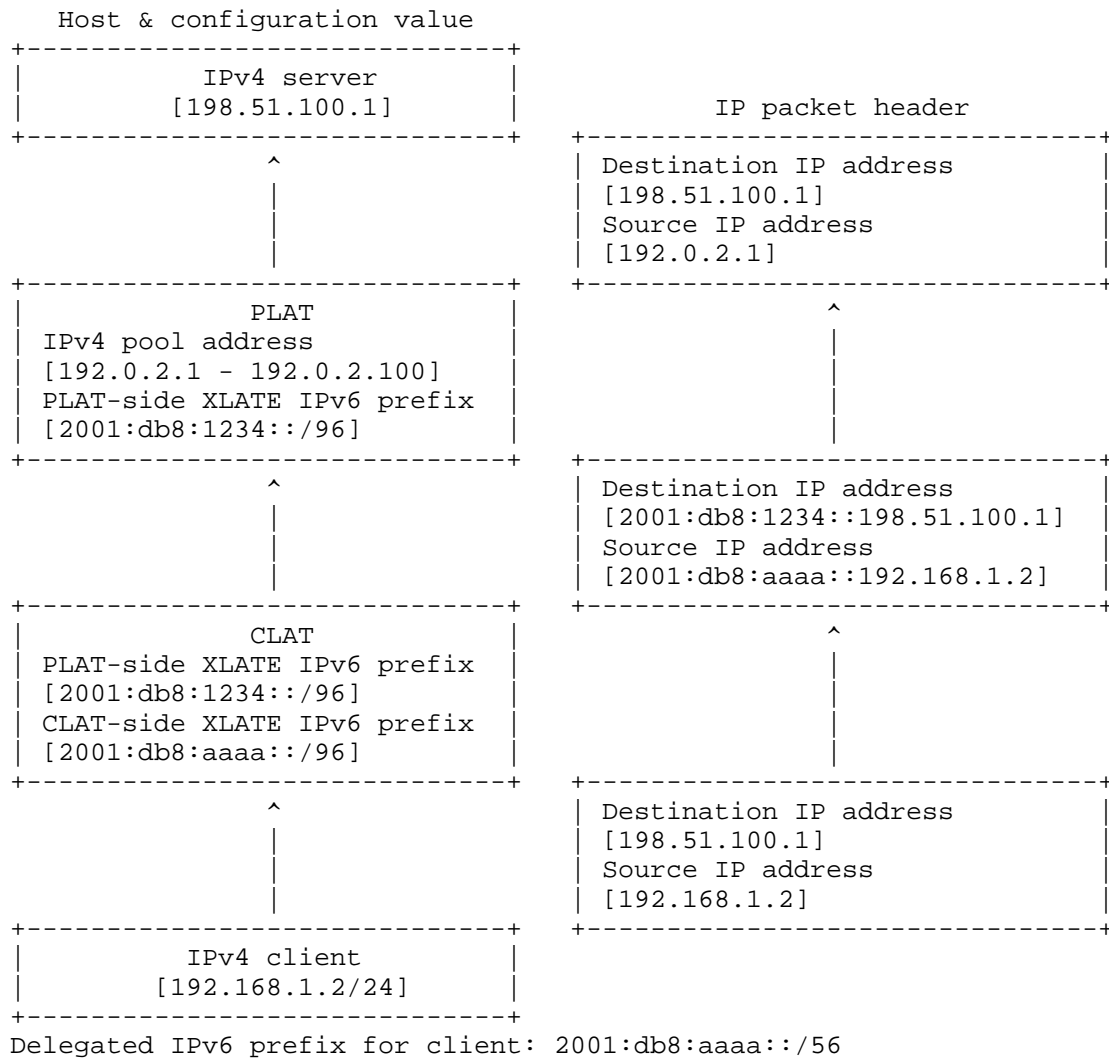
- [I-D.ietf-behave-nat64-discovery-heuristic] Savolainen, T., Korhonen, J., and D. Wing, "Discovery of the IPv6 Prefix Used for IPv6 Address Synthesis", draft-ietf-behave-nat64-discovery-heuristic-13 (work in progress), November 2012.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC5245] Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", RFC 5245, April 2010.
- [RFC5625] Bellis, R., "DNS Proxy Implementation Guidelines", BCP 152, RFC 5625, August 2009.
- [RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service", RFC 6092, January 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.
- [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.

[TS.23203] 3GPP, "Policy and charging control architecture", 3GPP
TS 23.203 10.7.0, June 2012.

Appendix A. Examples of IPv4/IPv6 Address Translation

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

In the case that an IPv6 prefix greater than /64 is assigned to an
end user by such as DHCPv6-PD [RFC3633], the CLAT can use a dedicated
/64 from the assigned IPv6 prefix.



Authors' Addresses

Masataka Mawatari
Japan Internet Exchange Co.,Ltd.
KDDI Otemachi Building 19F, 1-8-1 Otemachi,
Chiyoda-ku, Tokyo 100-0004
JAPAN

Phone: +81 3 3243 9579
Email: mawatari@jpix.ad.jp

Masanobu Kawashima
NEC AccessTechnica, Ltd.
800, Shimomata
Kakegawa-shi, Shizuoka 436-8501
JAPAN

Phone: +81 537 22 8274
Email: kawashimam@vx.jp.nec.com

Cameron Byrne
T-Mobile USA
Bellevue, Washington 98006
USA

Email: cameron.byrne@t-mobile.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 17, 2013

Z. Li
H. Guo
C. Liu
China Telecom
W. Liu
Z. Zhang
Huawei Technologies
July 16, 2012

Experience from NAT44 Translation Testing
draft-li-behave-nat444-test-01

Abstract

This document describes the testing result of CGN device in Wuxi Branch of China Telecom, by providing an overview of support situation of CGN for getting applications through NAT. The CGN device is from Huawei and the test environment is a real network in Wuxi China.

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 .

Status of this Memo

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

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

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

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

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. 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. Testbed Overview	4
3.1. A general topology for NAT444 testing	5
3.2. Testbed Description	7
4. Applications Testing Overview	8
4.1. Instant message applications	8
4.1.1. Microsoft Messenger	8
4.1.2. skype	8
4.1.3. Other IM	9
4.2. Web browsing	9
4.2.1. www.google.com	9
4.2.2. Other web browsings	10
4.3. Online gaming	10
4.3.1. QQ online gaming	10
4.3.2. Other online gaming	11
4.4. Downloading	11
4.4.1. HTTP downloading	11
4.4.2. FTP downloading	12
4.4.3. Bittorrent/eMule downloading	13
4.4.4. Xunlei downloading	14
4.5. Internet Video/music	15
4.5.1. PPStream	15
4.5.2. Other Internet Video/music	16
4.6. Email	16
4.6.1. Outlook/Outlook express	16
4.6.2. Other Email softwares	17
4.7. Other applications	17
4.7.1. Telnet	17
4.7.2. SSH	18
4.7.3. Traceroute	19
4.7.4. Remote desktop	20
4.8. VPN	21
4.8.1. iAccess	21
4.9. Shopping online	22

4.9.1. Taobao	22
4.10. Bank	23
4.10.1. China Merchants Bank	23
4.11. Negotiable securities	24
4.11.1. United securities	24
4.12. Map	25
4.12.1. google map	25
5. Applications Testing with same public IP address	26
5.1. Instant message applications	26
5.1.1. Microsoft Messenger	26
5.2. Online gaming	27
5.2.1. QQ online gaming	27
5.3. Internet Video/music	28
5.3.1. Youku	28
5.4. Shopping online	29
5.4.1. Taobao	29
5.5. Bank	30
5.5.1. Industrial and Commercial Bank of China	30
6. Effect analysis	31
6.1. User experience	31
6.2. Testing summary	31
7. Security Considerations	32
8. Acknowledgments	32
9. IANA Considerations	32
10. Informative References	32
Authors' Addresses	32

1. Introduction

This testing is based on specification of IP device from China Telecom. The main purpose is to know the states that CGN supports the applications translating the NAT device. The testing is done on a real network of China Telecom Wuxi branch where the CGN is a centralized device for NAT translation.

Base on testing result we know which applications could adapt to the NAT device and the time delay after translation, whether there is echo for video and audio services.

The CGN devices include BRAS, SR, CR which can support NAT444 by adding a CGN board or connecting a CGN device. The access devices include LSW, DSLAM, OLT, MxU. CPE devices can be HGW, ONT which support router/bridge model. Other devices such as Network management servers, log servers, AAA servers, user action analysis server, FTP/HTTP server are also included in the system.

2. Terminology

This document makes use of the following terms:

- NAT: Network Address Translation
- CGN : Carrier Grade NAT
- BRAS: Broadband Remote Access Server
- SR: Service Router
- CR: Core Router
- LSW: LAN Switching
- DSLAM: Digital Subscriber Line Access Multiplexer
- OLT: Optical Line Terminal
- CPE: Customer premises equipment
- HGW: Home Gateway
- ONT: Optical Network Terminal
- FTP: File Transfer Protocol
- HTTP: Hypertext Transfer Protocol
- ALG: Application Layer Gateway
- PCP: Port Control Protocol
- VPN: Virtual Private Network
- SSH: Secure Shell

3. Testbed Overview

3.1. A general topology for NAT444 testing

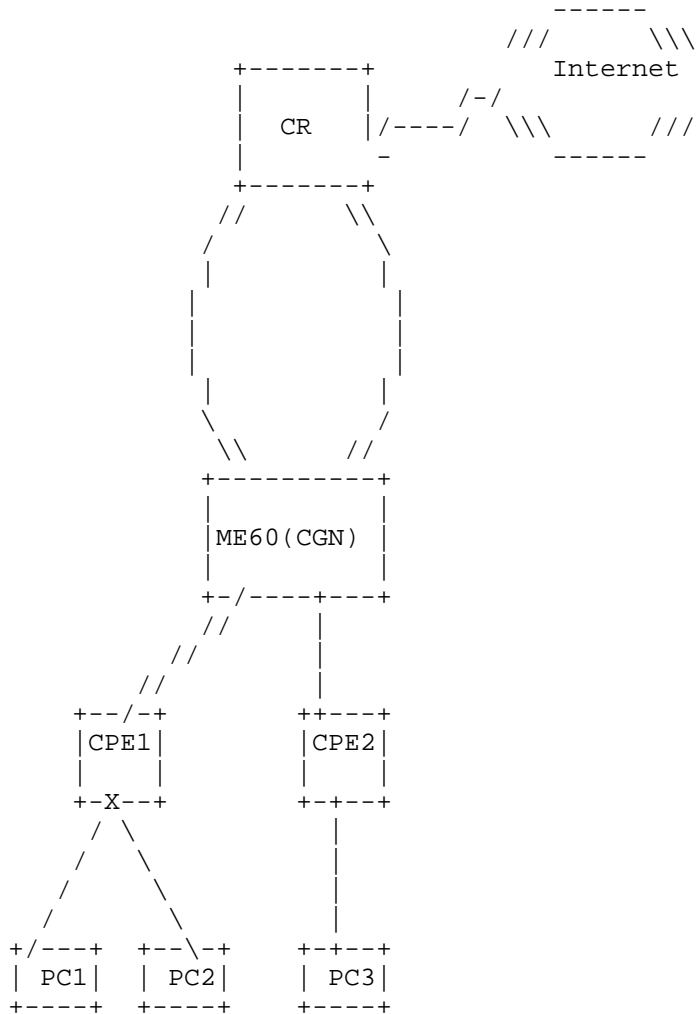


Figure 1: Distributed CGN topology for NAT444 testing

In figure 1 CPE1 and CPE2 have NAT function, and NE60 is a BRAS device with a embedded CGN . There are two scenarios in figure 1. Scenario 1: Communication between PC1 and PC2; Scenario 2: Communication between PC2 and PC3 .

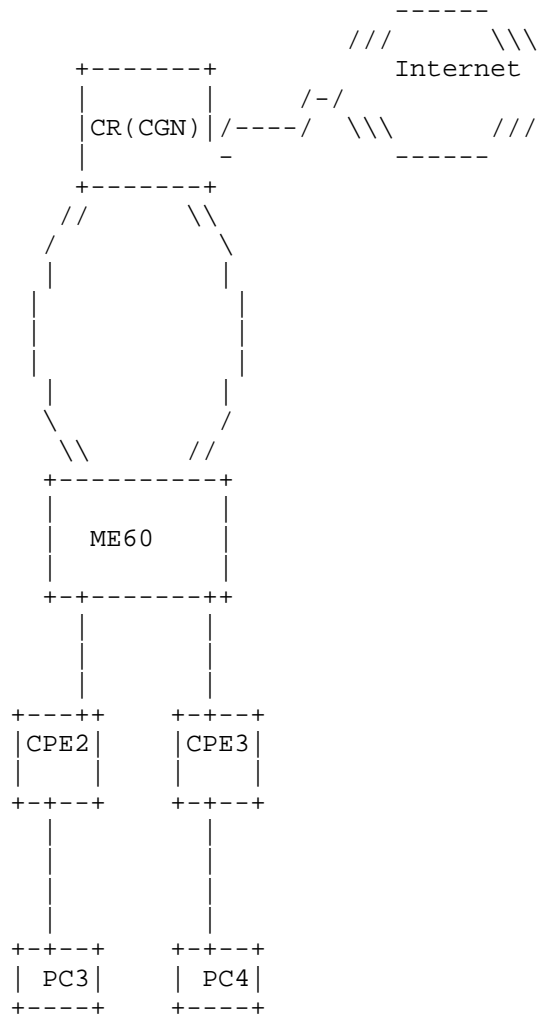


Figure 2:Centralized CGN topology for NAT444 testing

In figure 2 CPE2 and CPE3 have NAT function, and ME60 is a BRAS device without embedded CGN . There is an embedded CGN in CR device. This is scenario 3: Communication between PC3 and PC4.

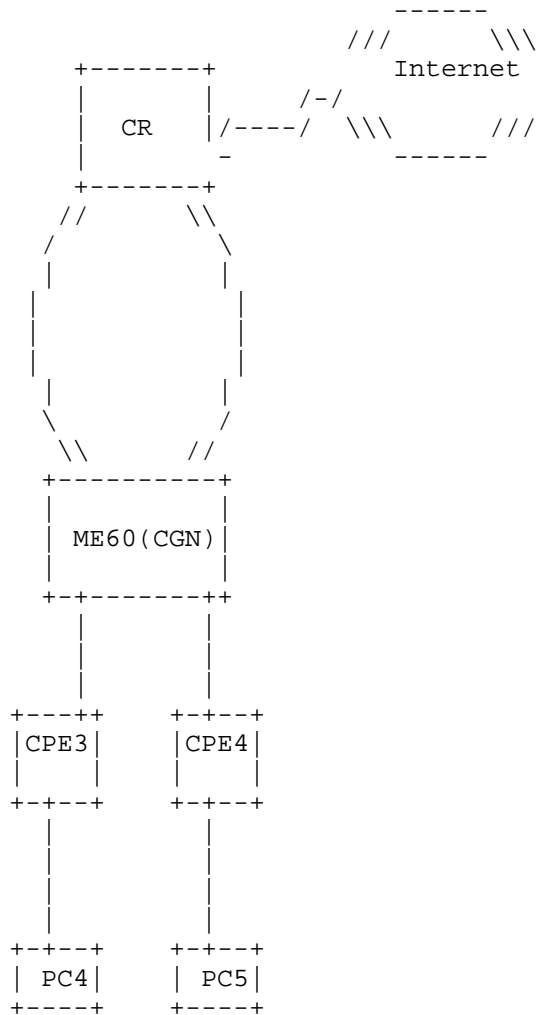


Figure 3:Public user and private user interworking

In figure 3 CPE3 has NAT function and accesses a private IP address from NE60; CPE4 has NAT function and accesses a public IPv4 address by PPP from NE60. NE60 is a BRAS device with a embedded CGN. This is scenario 4: Communication between PC4 and PC5.

3.2. Testbed Description

During the testing ALG function can be closed and open. So we tested based on: Activation ALG and three-tuple(Index NAT entries by source IP, source port, protocol) ; Deactivation ALG and tree-tuple;

Activation single ALG and three-tuple; Activation ALG and Five-tuple(Index NAT entries by source IP, source port, protocol, destined IP, destined port) ; Deactivation ALG and five-tuple;

4. Applications Testing Overview

This section describes testing result for all kinds applications.

4.1. Instant message applications

4.1.1. Microsoft Messenger

Test Item	IM
Sub-Item	Microsoft Messenger
Test Objective	Check whether Microsoft Messenger can work under NAT44.Voice, Video, Webcam,File transfer are tested
Test Scenario	Scenario:1, 2,3,4
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install MSN in PC 3.Check whether MSN user can register 4.Check whether users can communicate normally 5.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	MSN user can register Two user can communicate with MSN Under four scenarios two user can communicate
Actual Result	Passed
Remarks	Independent ALG

4.1.2. skype

Test Item	IM
Sub-Item	Skype
Test Objective	Check whether skype can used under NA44. Voice, Video, Webcam, File transfer are tested
Test Scenario	Scenario:1, 2,3,4
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install skype in PC 3.Check whether skype user can register 4.Check whether users can communicate normally 5.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	Skype user can register Two user can communicate with skype Under four scenarios two user can communicate
Actual Result	Passed
Remarks	Independent ALG

4.1.3. Other IM

We tested other IM application in the same way and got the same result as MSN. Other IM application include Feixin, QQ, Miliao, aliwangwang, and they are all popular IM applications in china.

4.2. Web browsing

4.2.1. www.google.com

Test Item	Web browsing
Sub-Item	www.google.com
Test Objective	Check whether we can access www.google.com when there is NAT in the network.
Test Scenario	Scenario:1, 2,3,4 PCs can access web browsing
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open browsing and access www.google.com in PC 3.Check whether PC can access the Web normally. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	PC can access the web.
Actual Result	Passed
Remarks	Independent ALG

4.2.2. Other web browsings

We tested other web browsings in the same way and got the same result as google web. Other web browsings include www.baidu.com, www.yahoo.com, www.sohu.com, www.renren.com, www.sina.com, www.tianya.cn, www.qq.com, www.163.com, www.ifeng.com, www.chinanews.com, and they are all popular web sites in china. We also access web by HTTPS,we access <https://chatmodels.dmm.co.jp/login/top> and it runs smoothly.

4.3. Online gaming

4.3.1. QQ online gaming

Test Item	Online gaming
Sub-Item	QQ Online gaming
Test Objective	Check whether PC can register QQ online gaming room.
Test Scenario	Scenario:1, 2,3,4 PCs can access online gaming room.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install QQ online gaming client on PC 3.Check whether PC can entry game room and play. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	QQ game user can entry game room and play.
Actual Result	Passed
Remarks	Independent ALG

4.3.2. Other online gaming

We tested other online gamings in the same way and got the same result as QQ online gaming. Other online gamings include World of Warcraft , QQ farm, ourgame, Kaixin network, and they are all popular online game in china.

4.4. Downloading

4.4.1. HTTP downloading

Test Item	Downloading
Sub-Item	HTTP downloading
Test Objective	Check whether PC can download by HTTP with NAT444 on the networks.
Test Scenario	Scenario:1, 2,3,4 PCs can download by HTTP.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open any software or MP3 file download page. 3.Check whether PC can download the by HTTP. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can download files by HTTP.
Actual Result	Passed
Remarks	Independent ALG

4.4.2. FTP downloading

Test Item	Downloading
Sub-Item	FTP downloading
Test Objective	Check whether PC can download by FTP with NAT444 on the networks.
Test Scenario	Scenario:1, 2,3,4 PCs can download by FTP.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Input a FTP address:FTP://debian.bjlx.org.cn. 3.Check whether PC can connect to FTP server and download by FTP. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can download files by FTP.
Actual Result	Passed but dependent ALG
Remarks	Not testing when FTP server is in private network

4.4.3. Bittorrent/eMule downloading

Test Item	Downloading
Sub-Item	Bittorrent/eMule
Test Objective	Check whether PC can download by Bittorrent/eMule
Test Scenario	Scenario:1, 2,3,4 PCs can download by Bittorrent/eMule
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install Bittorrent or eMule client on PC. 3.Check whether PC can download by Bittorrent/eMule. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can download files by Bittorrent. User can download files by eMule.
Actual Result	Passed and Independent ALG
Remarks	No testing When Bittorrent server in private network No testing When eMule server in private network. CGN not support PCP

Remark: PCP([draft-ietf-pcp-base-26]) is not actived in CGN. When eMule/Bittorrent server is behind in CGN, we didn't test.

+--+

4.4.4. Xunlei downloading

Test Item	Downloading
Sub-Item	Xunlei downloading
Test Objective	Check whether PC can download by Xunlei when it is in a private network.
Test Scenario	Scenario:1, 2,3,4 PCs can download by Xunlei.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install Xunlei client on PC. 3.Open a file in Xunlei and check whether PC can download by Xunlei. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can download files by Xunlei.
Actual Result	Passed and Independent ALG
Remarks	

4.5. Internet Video/music

4.5.1. PPStream

Test Item	Internet Video/music
Sub-Item	PPStream
Test Objective	Check whether PC with PPStream client can play video/music programme.
Test Scenario	Scenario:1, 2,3,4 PCs can play video/music programme
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install PPStream client on PC. 3.Check whether PC can play programmes on PPStream. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can see the film or listen to music with PPStream client.
Actual Result	Passed
Remarks	Independent ALG

4.5.2. Other Internet Video/music

We tested other Internet Video/music software in the same way and got the same result as PPStream. Other Internet Video/music software include PPlive, Youku, Qiyi, Xunleikankan, Tudou, Baidu video, Sohu video, 163 video, and they are all popular video/music used in china.

Youtube can't be accessed by Chinese user and do not pass the test.

4.6. Email

4.6.1. Outlook/Outlook express

Test Item	Email
Sub-Item	Outlook/Outlook express
Test Objective	Check whether PC with Outlook/Outlook express can receive and send mail from mail server.
Test Scenario	Scenario:1, 2,3,4 PCs can receive/send mail.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Set Outlook/Outlook express on PC. 3.Check whether PC can use Outlook/Outlook express. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can see the film or listen to music with PPStream client.
Actual Result	Passed
Remarks	Independent ALG

4.6.2. Other Email softwares

We tested other Email software in the same way and got the same result as Outlook/Outlook express. Other Email softwares include QQ mail, 163 mail, sina mail, and they are all popular mail used in china.

4.7. Other applications

4.7.1. Telnet

Test Item	Telnet
Sub-Item	Telnet
Test Objective	Check whether PC can telnet a device within NAT environment.
Test Scenario	Scenario:1, 2,3,4 PCs can Telnet.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Configure the Telnet on a PC. 3.Check whether PC can build telnet. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can build the telnet connection.
Actual Result	Passed
Remarks	Independent ALG

4.7.2. SSH

Test Item	SSH
Sub-Item	SSH
Test Objective	Check whether PC can build SSH connection within NAT environment.
Test Scenario	Scenario:1, 2,3,4 PCs can Build SSH connection.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Configure the SHH on a router in network 3.Check whether PC can build SSH connection 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can build the SHH connection.
Actual Result	Passed
Remarks	Independent ALG

4.7.3. Traceroute

Test Item	Traceroute
Sub-Item	Traceroute (using ICMP)
Test Objective	Check whether two PCs behind NAT can traceroute. NAT environment.
Test Scenario	Scenario:1, 2,3,4 .
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Traceroute from a PC to another PC. 3.Check whether two PC can traceroute. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	Two users can traceroute.
Actual Result	Passed
Remarks	Independent ALG

4.7.4. Remote desktop

Test Item	Remote desktop
Sub-Item	Remote desktop
Test Objective	Check whether a PC behind NAT can remote desktop to another PC behind NAT or to a public PC.
Test Scenario	Scenario:1, 2,3,4 .
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Remote desktop from a PC to another PC. 3.Check whether two PC can remotedesktop successfully 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User behind CGN can remote desktop to another CGN user or a public IP user.
Actual Result	Passed
Remarks	Independent ALG

4.8. VPN

4.8.1. iAccess

Test Item	VPN
Sub-Item	iAccess
Test Objective	Check whether a PC behind NAT can remote desktop to another PC behind NAT or to a public PC.
Test Scenario	Scenario:1, 2,3,4 .
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Get a iAccess user and password from company. 3.Check whether public PC can access the company. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can access company resource from public network by iAccess user and password.
Actual Result	Passed
Remarks	Independent ALG; not test PPTP,L2TP

4.9. Shopping online

4.9.1. Taobao

Test Item	Shopping online
Sub-Item	Taobao
Test Objective	Check whether user can shop by Taobao within NAT environment.
Test Scenario	Scenario:1, 2,3,4 PC can access Taobao. .
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open browsing and input Taobao address. 3.Check whether user can access Taobao web site. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can shop in Taobao and do all kind of operation in web site.
Actual Result	Passed
Remarks	Independent ALG

4.10. Bank

4.10.1. China Merchants Bank

Test Item	Bank
Sub-Item	China Merchants Bank
Test Objective	Check whether user can use online bank web within NAT environment.
Test Scenario	Scenario:1, 2,3,4 PC can access online bank. .
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open browsing and input China Merchants Bank Addr 3.Check whether user can use online bank. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can use online bank on web site.
Actual Result	Passed
Remarks	Independent ALG

4.11. Negotiable securities

4.11.1. United securities

Test Item	Negotiable securities
Sub-Item	United securities
Test Objective	Check whether user can entry securities exchange centre and trade.
Test Scenario	Scenario:1, 2,3,4 PC can access securities web.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install United securities client. 3.Check whether user can entry the securities exchange centre and trade 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can entry securities exchange centre and trade.
Actual Result	Passed
Remarks	Independent ALG

4.12. Map

4.12.1. google map

Test Item	MAP
Sub-Item	Google map
Test Objective	Check whether user can use google map for search Within the NAT environment.
Test Scenario	Scenario:1, 2,3,4 PC can use google map.
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open google map. 3.Check whether user can goole map for search. Check the session entries on CGN. 4.Test Activation/Deactivation/Single ALG+tree-tuple
Expected Result	User can use google map for search.
Actual Result	Passed
Remarks	Independent ALG

We tested Baidu map in the same way and got the same result .

5. Applications Testing with same public IP address

This section describes testing result when different CPEs use same public IP address. The purpose of testing is make sure the application can also be used when different users use same external public IP address.

This section include three scenarios. Scenario 1: in figure 1 PC1 and PC2 use same external public IP address; Scenario 2: in figure1 PC2 and PC3 use same external public IP address; Scenario 3: in figure 3 PC4 are CGN user and PC5 are public user;

5.1. Instant message applications

5.1.1. Microsoft Messenger

Test Item	IM
Sub-Item	Microsoft Messenger
Test Objective	Check when ALG active or deactive whether MSN has same communication flow in three scenarios.
Test Scenario	Scenario:1, 2,3
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install MSN in PC 3.Check whether MSN user can register 4.Active ALG and see the communication flow by grasping packets in three scenarios.
Expected Result	MSN user can communicate in three scenarios.
Actual Result	Passed
Remarks	

5.2. Online gaming

5.2.1. QQ online gaming

Test Item	Online gaming
Sub-Item	QQ Online gaming
Test Objective	Check whether QQ online game has the same flow when ALG active or deactive.
Test Scenario	Scenario:1, 2,3
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Install QQ online gaming client on PC 3.Check whether PC can entry game room and play. 4.Grasp packets when ALG active or deactive.
Expected Result	QQ game user can entry game room and play.
Actual Result	Failed
Remarks	same public IP user can't entry the same game room.

5.3. Internet Video/music

5.3.1. Youku

Test Item	Internet Video/music
Sub-Item	Youku
Test Objective	Check whether Youku has the same flow when ALG active or deactive.
Test Scenario	Scenario:1, 2,3
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Go to Youku web site and view video. 3.Grasp packets when ALG active or deactive and analyse the flow.
Expected Result	User can see the film or listen to music in Youku web site.
Actual Result	Passed
Remarks	

5.4. Shopping online

5.4.1. Taobao

Test Item	Shopping online
Sub-Item	Taobao
Test Objective	Check whether Taobao user has the same flow when NAT actives or deactives.
Test Scenario	Scenario:1, 2,3
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open browsing and input Taobao address. 3.Check whether user can shop on Taobao web site. 4.Grasp packets when ALG actives or deactives to see whether the flow are same or not.
Expected Result	User can shop in Taobao.
Actual Result	Passed
Remarks	

5.5. Bank

5.5.1. Industrial and Commercial Bank of China

Test Item	Bank
Sub-Item	Industrial and Commercial Bank of China(ICBC)
Test Objective	Check when user can use online ICBC bank web the service flow is same when activating/deactivating ALG.
Test Scenario	Scenario:1, 2,3
Test Procedure	1.Configure user IP pool in BRAS. Configure NAT444 and IPv4 public pool in CGN. 2.Open browsing and input ICBC Bank address. 3.Check whether user can use online bank to transfer 4.Grasp the packets to analyse the flow when ALG actives or deactives.
Expected Result	User can use online bank on web site.
Actual Result	Passed
Remarks	

6. Effect analysis

6.1. User experience

User experience can't be quantified and we get the result only by subjective experience. Time delay, echo, fluency in video and audio are almost same as without NAT444 on network. Communications between CGN users and CGN user with public user are always normal. As a result, NAT444 has no affection on the users' experience in the tests we have run.

6.2. Testing summary

In all the applications aforementioned only FTP depends on ALG. We only test two levels NAT.

QQ online gaming does not permit two users use the same external public IP address in the same game room. When two users use the same

external public IP address, QQ online gaming considers they come from the same subscriber. If they are in the same game room, they are regarded as cribbers.

We only tested a bank account to use online bank since we only have one account.

We didn't test when eMule, Bittorrent work as internal server. This needs support of PCP.

When there is two levels NAT, users can't set internal server, such as FTP server, in home network.

Communication between CGN user and public IP user belonging to the same CGN is not processed by service board.

7. Security Considerations

8. Acknowledgments

9. IANA Considerations

10. Informative References

[draft-ietf-pcp-base-26]
IETF, "Port Control Protocol (PCP)", June 2012,
<<http://tools.ietf.org/html/draft-ietf-pcp-base-26>>.

Authors' Addresses

Zhongchao Li
China Telecom
Nanjing,
P.R. China

Email: 15301588336@189.cn

Hongwei Guo
China Telecom
Nanjing,
P.R. China

Email: 15306188213@189.cn

Chunlin Liu
China Telecom
Nanjing,
P.R. China

Email: liuchunlin@jsptpd.com

Will Liu
Huawei Technologies
Bantian, Longgang DIST
Shenzhen 518129
P.R. China

Phone: +86 755 28972315
Email: liushucheng@huawei.com

Zhongjian Zhang
Huawei Technologies
Bantian, Longgang DIST
Shenzhen,
P.R. China

Email: zhangzhongjian@huawei.com

Port Control Protocol
Internet-Draft
Intended status: BCP
Expires: January 16, 2013

R. Penno
S. Perreault
Cisco
S. Kamiset

M. Boucadair
France Telecom
July 15, 2012

Network Address Translation (NAT) Behavioral Requirements Updates
draft-penno-behave-rfc4787-5382-5508-bis-03

Abstract

This document clarifies and updates several requirements of RFC4787, RFC5382 and RFC5508 based on operational and development experience. The focus of this document is NAPT44.

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 16, 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. Terminology	3
2. Introduction	3
2.1. Scope	3
3. TCP Session Tracking	3
3.1. TCP Transitory Connection Idle-Timeout	4
3.2. TCP RST	5
4. Address Pooling Paired (APP)	5
5. EIF Security	5
6. EIF Protocol Independence	5
7. EIF Mapping Refresh	6
7.1. Outbound Mapping Refresh and Error Packets	6
8. EIM Protocol Independence	6
9. Port Parity	6
10. Port Randomization	7
11. IP Identification (IP ID)	7
12. ICMP Query Mappings Timeout	7
13. Hairpinning Support for ICMP Packets	7
14. IANA Considerations	8
15. Security Considerations	8
16. Acknowledgements	8
17. References	8
17.1. Normative References	8
17.2. Informative References	9
Authors' Addresses	9

1. Terminology

The reader should be familiar with all terms defined in RFC2663 [RFC2663], RFC4787 [RFC4787], RFC5382 [RFC5382], RFC5508 [RFC5508]

2. Introduction

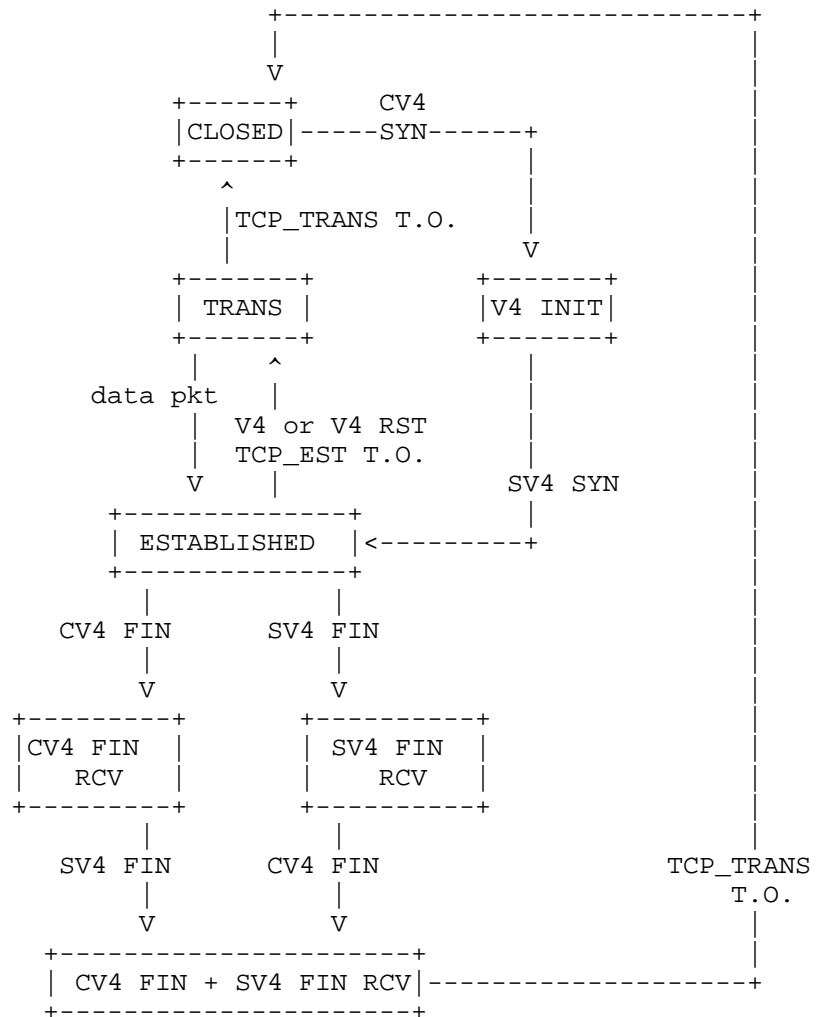
[RFC4787], [RFC5382] and [RFC5508] greatly advanced NAT interoperability and conformance. But with widespread deployment and evolution of NAT more development and operational experience was acquired some areas of the original documents need further clarification or updates. This documents provides such clarifications and updates.

2.1. Scope

This document focuses solely on NAPT44 and its goal is to clarify, fill gaps or update requirements of [RFC4787], [RFC5382] and [RFC5508]. It is out of the scope of this document the creation of completely new requirements not associated with the documents cited above. New requirements would be better served elsewhere and if they are CGN specific in [I-D.ietf-behave-lsn-requirements]

3. TCP Session Tracking

[RFC5382] specifies TCP timers associated with various connection states but does not specify the TCP state machine a NAPT44 should use as a basis to apply such timers. The TCP state machine below, adapted from [RFC6146], provides guidance on how TCP session tracking could be implemented - it is non-normative.



(postamble)

3.1. TCP Transitory Connection Idle-Timeout

[RFC5382]:REQ-5 The transitory connection idle-timeout is defined as the minimum time a TCP connection in the partially open or closing phases must remain idle before the NAT considers the associated session a candidate for removal. But the document does not clearly states if these can be configured separately. This document clarifies that a NAT device SHOULD provide different knobs for configuring the open and closing idle timeouts. This document further acknowledges that most TCP flows are very short (less than 10 seconds) [FLOWRATE][TCPWILD] and therefore a partially open timeout

of 4 minutes might be excessive if security is a concern. Therefore it MAY be configured to be less than 4 minutes in such cases.

There are other initiatives to reduce reclaim state at NAT devices faster [I-D.naito-nat-resource-optimizing-extension]

3.2. TCP RST

[RFC5382] leaves the handling of TCP RST packets unspecified. This document does not try standardize such behavior but clarifies based on operational experience that a NAT that receives a TCP RST for an active mapping and performs session tracking MAY immediately delete the sessions and remove any state associated with it. If the NAT device that performs TCP session tracking receives a TCP RST for the first session that created a mapping, it MAY remove the session and the mapping immediately.

4. Address Pooling Paired (APP)

[RFC4787]: REQ-2 [RFC5382]:ND Address Pooling Paired behavior for NAT is recommended in previous documents but behavior when a public IPv4 run out of ports is left undefined. This document clarifies that if APP is enabled new sessions from a subscriber that already has a mapping associated with a public IP that ran out of ports SHOULD be dropped. The administrator MAY provide a knob that allows a NAT device to starting using ports from another public IP when the one that anchored the APP mapping ran out of ports. This is trade-off between subscriber service continuity and APP strict enforcement. (NE: It is sometimes referred as 'soft-APP')

5. EIF Security

[RFC4787]:REQ-8 and [RFC5382]:REQ-3 End-point independent filtering could potentially result in security attacks from the public realm. In order to handle this, when possible there MUST be strict filtering checks in the inbound direction. A knob SHOULD be provided to limit the number of inbound sessions and a knob SHOULD be provided to enable or disable EIF on a per application basis.

6. EIF Protocol Independence

[RFC4787]:REQ-8 and[RFC5382]: REQ-3 Current RFCs do not specify whether EIF mappings are protocol independent. In other words, if a outbound TCP SYN creates a mapping it is left undefined whether inbound UDP can create sessions and packets are forwarded. EIF

mappings SHOULD be protocol independent in order allow inbound packets for protocols that multiplex TCP and UDP over the same IP: port through the NAT and maintain compatibility with stateful NAT64 RFC6146 [RFC6146]. But the administrator MAY provide a configuration knob to make it protocol dependent.

7. EIF Mapping Refresh

[RFC4787]: REQ-6 [RFC5382]: ND The NAT mapping Refresh direction MAY have a "NAT Inbound refresh behavior" of "True" but it does not clarify how this applies to EIF mappings. The issue in question is whether inbound packets that match an EIF mapping but do not create a new session due to a security policy should refresh the mapping timer. This document clarifies that even when a NAT device has a inbound refresh behavior of TRUE, that such packets SHOULD NOT refresh the mapping. Otherwise a simple attack of a packet every 2 minutes can keep the mapping indefinitely.

7.1. Outbound Mapping Refresh and Error Packets

In the case of NAT outbound refresh behavior there might be certain types of packets that should not refresh the mapping. For example, if the mapping is kept alive by ICMP Error or TCP RST outbound packets sent as response to inbound packets, these SHOULD NOT refresh the mapping.

8. EIM Protocol Independence

[RFC4787] [RFC5382]: REQ-1 Current RFCs do not specify whether EIM are protocol independent. In other words, if a outbound TCP SYN creates a mapping it is left undefined whether outbound UDP can reuse such mapping and create session. On the other hand, Stateful NAT64 [RFC6146] clearly specifies three binding information bases (TCP, UDP, ICMP). This document clarifies that EIM mappings SHOULD be protocol dependent. A knob MAY be provided in order allow protocols that multiplex TCP and UDP over the same source IP and port to use a single mapping.

9. Port Parity

A NAT devices MAY disable port parity preservation for dynamic mappings. Nevertheless, A NAT SHOULD support means to explicitly request to preserve port parity (e.g., [I-D.boucadair-pcp-rtcp]).

10. Port Randomization

A NAT SHOULD follow the recommendations specified in Section 4 of [RFC6056] especially: "A NAT that does not implement port preservation [RFC4787] [RFC5382] SHOULD obfuscate selection of the ephemeral port of a packet when it is changed during translation of that packet. A NAT that does implement port preservation SHOULD obfuscate the ephemeral port of a packet only if the port must be changed as a result of the port being already in use for some other session. A NAT that performs parity preservation and that must change the ephemeral port during translation of a packet SHOULD obfuscate the ephemeral ports. The algorithms described in this document could be easily adapted such that the parity is preserved (i.e., force the lowest order bit of the resulting port number to 0 or 1 according to whether even or odd parity is desired)."

11. IP Identification (IP ID)

A NAT SHOULD handle the Identification field of translated IPv4 packets as specified in Section 9 of [I-D.ietf-intarea-ipv4-id-update].

12. ICMP Query Mappings Timeout

Section 3.1 of [RFC5508] says that ICMP Query Mappings are to be maintained by NAT device. However, RFC doesn't discuss about the Query Mapping timeout values. Section 3.2 of that RFC only discusses about ICMP Query Session Timeouts. ICMP Query Mappings MAY be deleted once the last the session using the mapping is deleted.

13. Hairpinning Support for ICMP Packets

[RFC5508]:REQ-7 This requirement specifies that NAT devices enforcing Basic NAT MUST support traversal of hairpinned ICMP Query sessions. This implicitly means that address mappings from external address to internal address (similar to Endpoint Independent Filters) MUST be maintained to allow inbound ICMP Query sessions. If an ICMP Query is received on an external address, NAT device can then translate to an internal IP. [RFC5508]:REQ-7 This requirement specifies that all NAT devices (i.e., Basic NAT as well as NAT devices) MUST support the traversal of hairpinned ICMP Error messages. This too requires NAT devices to maintain address mappings from external IP address to internal IP address in addition to the ICMP Query Mappings described in section 3.1 of that RFC.

14. IANA Considerations

TBD

15. Security Considerations

In the case of EIF mappings due to high risk of resource crunch, a NAT device MAY provide a knob to limit the number of inbound sessions spawned from a EIF mapping.

16. Acknowledgements

Thanks to Dan Wing, Suresh Kumar, Mayuresh Bakshi, Rajesh Mohan and Senthil Sivamular for review and discussions

17. References

17.1. Normative References

- [I-D.ietf-pcp-base]
Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-26 (work in progress), June 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.
- [RFC3605] Huitema, C., "Real Time Control Protocol (RTCP) attribute in Session Description Protocol (SDP)", RFC 3605, October 2003.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC5508] Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT Behavioral Requirements for ICMP", BCP 148, RFC 5508,

April 2009.

- [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.

17.2. Informative References

- [FLOWRATE] Zhang, Y., Breslau, L., Paxson, V., and S. Shenker, "On the Characteristics and Origins of Internet Flow Rates".
- [I-D.boucadair-pcp-rtp-rtcp] Boucadair, M. and S. Sivakumar, "Reserving N and N+1 Ports with PCP", draft-boucadair-pcp-rtp-rtcp-04 (work in progress), April 2012.
- [I-D.ietf-behave-lsn-requirements] Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common requirements for Carrier Grade NATs (CGNs)", draft-ietf-behave-lsn-requirements-08 (work in progress), July 2012.
- [I-D.naito-nat-resource-optimizing-extension] Kengo, K. and A. Matsumoto, "NAT resource optimizing extension", draft-naito-nat-resource-optimizing-extension-01 (work in progress), March 2012.
- [TCPWILD] Qian, F., Subhabrata, S., Spatscheck, O., Morley Mao, Z., and W. Willinger, "TCP Revisited: A Fresh Look at TCP in the Wild".

Authors' Addresses

Reinaldo Penno
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, California 95134
USA

Email: repenno@cisco.com

Simon Perreault
Cisco Systems, Inc.
2875 boul. Laurier, suite D2-630
Quebec, QC G1V 2M2
Canada

Email: simon.perreault@viagenie.ca

Sarat Kamiset
California

Phone:
Fax:

Mohamed Boucadair
France Telecom
Rennes, 35000
France

Email: mohamed.boucadair@orange.com

Port Control Protocol
Internet-Draft
Intended status: BCP
Expires: July 11, 2013

R. Penno
Cisco
S. Perreault
Viagenie
S. Kamiset
Consultant
M. Boucadair
France Telecom
K. Naito
NTT
January 07, 2013

Network Address Translation (NAT) Behavioral Requirements Updates
draft-penno-behave-rfc4787-5382-5508-bis-04

Abstract

This document clarifies and updates several requirements of RFC4787, RFC5382 and RFC5508 based on operational and development experience. The focus of this document is NAPT44.

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 July 11, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1. Terminology	3
2. Introduction	3
2.1. Scope	3
3. TCP Session Tracking	3
3.1. TCP Transitory Connection Idle-Timeout	4
3.1.1. Port resources limited case	5
3.1.2. Proposal: Apply RFC6191 and PAWS to NAT	6
3.2. TCP RST	9
4. Port Overlapping behavior	9
5. Address Pooling Paired (APP)	10
6. EIF Security	10
7. EIF Protocol Independence	10
8. EIF Mapping Refresh	10
8.1. Outbound Mapping Refresh and Error Packets	11
9. EIM Protocol Independence	11
10. Port Parity	11
11. Port Randomization	11
12. IP Identification (IP ID)	12
13. ICMP Query Mappings Timeout	12
14. Hairpinning Support for ICMP Packets	12
15. IANA Considerations	12
16. Security Considerations	12
17. Acknowledgements	13
18. References	13
18.1. Normative References	13
18.2. Informative References	14
Authors' Addresses	14

1. Terminology

The reader should be familiar with all terms defined in RFC2663 [RFC2663], RFC4787 [RFC4787], RFC5382 [RFC5382], RFC5508 [RFC5508]

2. Introduction

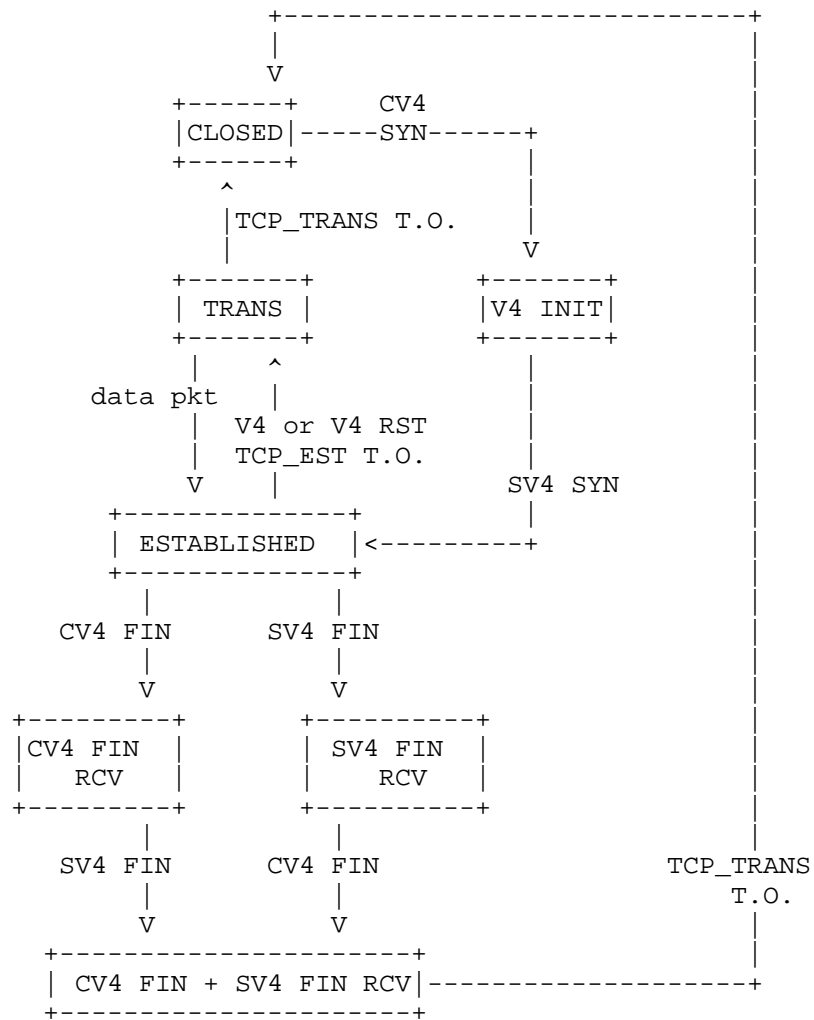
[RFC4787], [RFC5382] and [RFC5508] greatly advanced NAT interoperability and conformance. But with widespread deployment and evolution of NAT more development and operational experience was acquired some areas of the original documents need further clarification or updates. This documents provides such clarifications and updates.

2.1. Scope

This document focuses solely on NAPT44 and its goal is to clarify, fill gaps or update requirements of [RFC4787], [RFC5382] and [RFC5508]. It is out of the scope of this document the creation of completely new requirements not associated with the documents cited above. New requirements would be better served elsewhere and if they are CGN specific in [I-D.ietf-behave-lsn-requirements]

3. TCP Session Tracking

[RFC5382] specifies TCP timers associated with various connection states but does not specify the TCP state machine a NAPT44 should use as a basis to apply such timers. The TCP state machine below, adapted from [RFC6146], provides guidance on how TCP session tracking could be implemented - it is non-normative.



(postamble)

3.1. TCP Transitory Connection Idle-Timeout

[RFC5382]:REQ-5 The transitory connection idle-timeout is defined as the minimum time a TCP connection in the partially open or closing phases must remain idle before the NAT considers the associated session a candidate for removal. But the document does not clearly states if these can be configured separately. This document clarifies that a NAT device SHOULD provide different knobs for configuring the open and closing idle timeouts. This document further acknowledges that most TCP flows are very short (less than 10 seconds) [FLOWRATE][TCPWILD] and therefore a partially open timeout

of 4 minutes might be excessive if security is a concern. Therefore it MAY be configured to be less than 4 minutes in such cases. There also may be a case that timeout of 4 minutes might be excessive. The case and the solution are written below.

3.1.1.1. Port resources limited case

After IPv4 addresses run out, IPv4 address resources will be further restricted site-by-site. If global IPv4 address are shared between several clients, assignable port resources at each client will be limited.

NAT is a tool that is widely used to deal with this IPv4 address shortage problem. However, the demand for resources to provide Internet access to users and devices will continue to increase. IPv6 is a fundamental solution to this problem, but the deployment of IPv6 will take time.

In some cases, e.g. browsing a dynamic web page for a map service, a lot of sessions are used by the browser, and a number of ports are eaten up in a short time. What is worse is that when a NAT is between a PC and a server, TIME_WAIT state of each TCP connection is kept for certain period, typically for four minutes, which consumes port resources. Therefore, new connections cannot be established.

This problem is caused or worsened by the following behavior.

TIME_WAIT state assigned for a TCP connection remains active for 2MSL after the last ACK to the last FIN is transferred.

To reuse resources effectively, reducing TIME_WAIT without making any bad effect is important. To reduce TIME_WAIT, [RFC6191] is proposed for clients and remote hosts. To prevent bad effects, there is a PAWS mechanism, which prevent the old duplicate problem. We propose mechanisms adopting to NAT, to change the TIME_WAIT behavior that make it possible to save addresses and ports resources.

3.1.1.1.1. RFC6191 Reducing the TIME-WAIT State Using TCP Timestamps

[RFC6191] defines a mechanism for reducing the TIME_WAIT state using TCP timestamps and sequence numbers. When a connection request is received with a four-tuple that is in the TIME-WAIT state, the connection request may be accepted if the sequence number or the timestamp of the incoming SYN segment is greater than the last sequence number seen on the previous incarnation of the connection

3.1.1.2. TCP TIME_WAIT

The TCP TIME_WAIT state is described in [RFC0793]. The TCP TIME_WAIT state needs to be kept for 2MSL before a connection is CLOSED, for the reasons below.

- 1: In the event that packets from a session are delayed in the in-between network, and delivered to the end relatively later, we should prevent the packets from being transferred and interpreted as a packet that belongs to a new session.
- 2: If the remote TCP has not received the acknowledgment of its connection termination request, it will re-send the FIN packet several times.

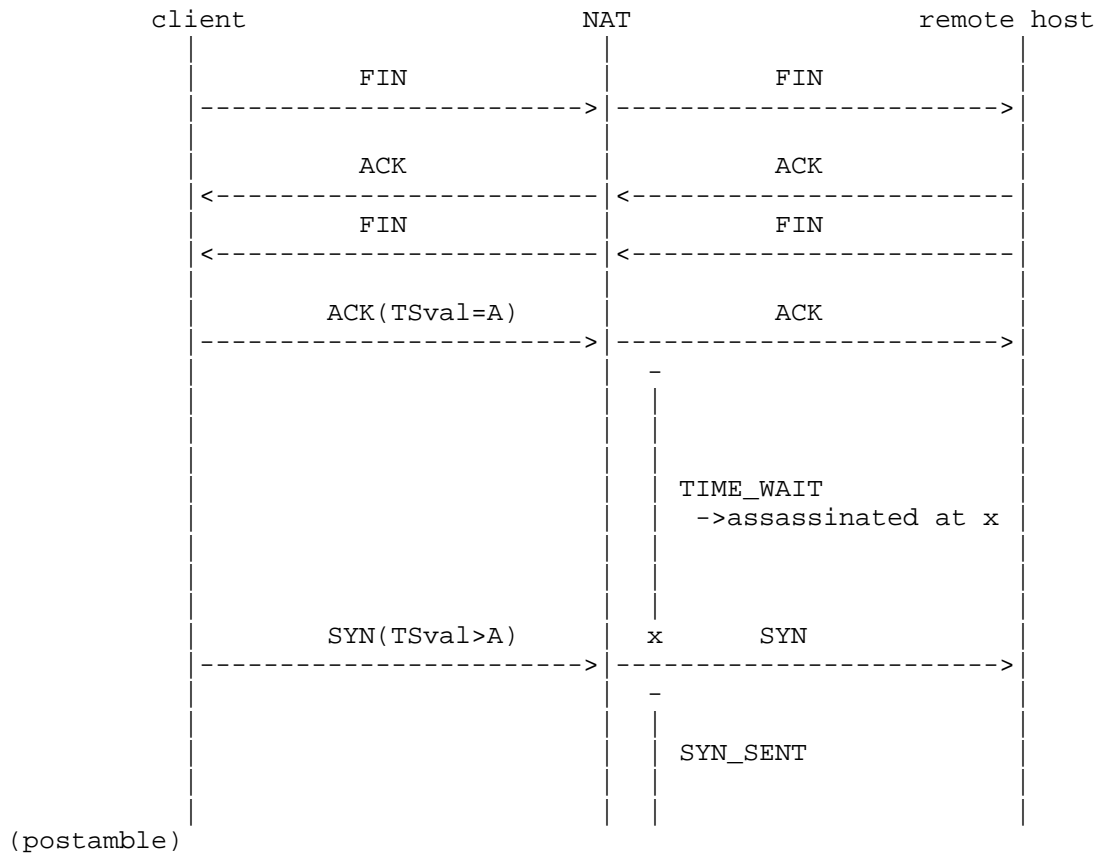
These points are important for the TCP to work without problems.

3.1.1.3. Protect Against Wrapped Sequence numbers (PAWS)

The TCP sequence number wraps frequently especially in a high bandwidth session. PAWS is used to prevent old duplicate packets that occurred in a previous session from being transferred to the new session whose valid TCP sequence numbers happen to overlap with the old duplicate packets. This is implemented by introducing TCP timestamp option, and checking the timestamp option value of each packet. PAWS is described in [RFC1323].

3.1.2. Proposal: Apply RFC6191 and PAWS to NAT

This section proposes to apply [RFC6191] mechanism at NAT. This mechanism MAY be adopted for both clients' and remote hosts' TCP active close.



Also, PAWS works to discard old duplicate packets at NAT. A packet can be discarded as an old duplicate if it is received with a timestamp or sequence number value less than a value recently received on the connection.

To make these mechanisms work, we should concern the case that there are several clients with nonsuccessive timestamp or sequence number values are connected to a NAT device (i.e. not monotonically increasing among clients). Two mechanisms to solve this mechanism and applying [RFC6191] and PAWS to NAT are described below. These mechanisms are optional.

3.1.2.1. Rewrite timestamp and sequence number values at NAT

Rewrite timestamp and sequence number values of outgoings packets at NAT to be monotonically increasing. This can be done by adopting following mechanisms at NAT.

A: Store the newest rewritten value of timestamp and sequence number as the "max value at the time".

B: NAT rewrite timestamp and sequence number values of incoming packets to be monotonically increasing.

When packets come back as replies from remote hosts, NAT rewrite again the timestamp and sequence number values to be the original values. This can be done by adopting following mechanisms at NAT.

C: Store the values of original timestamp and sequence number of packets, and rewritten values of those.

3.1.2.2. Split an assignable number of port space to each client

Adopt following mechanisms at NAT.

A: Choose clients that can be assigned ports.

B: Split assignable port numbers between clients.

Packets from other clients which are not chosen by these mechanisms are rejected at NAT, unless there is unassigned port left.

3.1.2.3. Resend the last ACK to the resended FIN

We should concern another case to make RFC6191 work at NAT. In case the remote TCP could not receive the acknowledgment of its connection termination request, NAT, on behalf of clients, resends the last ACK packet when it receives an FIN packet of the previous connection, and when the state of the previous connection is deleted from the NAT. This mechanism MAY be used when clients starts closing process, and the remote host could not receive the last ACK.

3.1.2.4. Remote host behavior of several implementations

To solve the port shortage problem on the client side, the behavior of remote host should be compliant to [RFC6191] or the mechanism written in 4.2.2.13 of [RFC1122], since NAT may reuse the same 5 tuple for a new connection. We have investigated behaviors of OSes (e.g., Linux, FreeBSD, Windows, MacOS), and found that they implemented the server side behavior of the above two.

3.2. TCP RST

[RFC5382] leaves the handling of TCP RST packets unspecified. This document does not try standardize such behavior but clarifies based on operational experience that a NAT that receives a TCP RST for an active mapping and performs session tracking MAY immediately delete the sessions and remove any state associated with it. If the NAT device that performs TCP session tracking receives a TCP RST for the first session that created a mapping, it MAY remove the session and the mapping immediately.

4. Port Overlapping behavior

There may be another solution to the address resource restricted environment written in 3.1.1. Also NAT are required to be mapped endpoint-independent in [RFC4787] and [RFC5382] REQ-1, the mechanism below MAY be one optional implement to NAT.

If destination addresses and ports are different for outgoing connections started by local clients, NAT MAY assign the same external port as the source ports for the connections. The port overlapping mechanism manages mappings between external packets and internal packets by looking at and storing the 5-tuple (protocol, source address, source port, destination address, destination port) of them. This enables concurrent use of a single NAT external port for multiple transport sessions, which enables NAT to work correctly in IP address resource limited network.

Discussions:

[RFC4787]and[RFC5382] requires "endpoint-independent mapping" at NAT, and port overlapping NAT cannot meet the requirement. This mechanism can degrade the transparency of NAT in that its mapping mechanism is endpoint-dependent and makes NAT traversal harder. However, if a NAT adopts endpoint-independent mapping together with endpoint-dependent filtering, then the actual behavior of the NAT will be the same as port overlapping NAT. It should also be noted that a lot of existing NAT devices(e.g., SEIL, FITElnet Series) adopted this port overlapping mechanism.

A: Reference URL for SEIL -> www.seil.jp

B: Reference URL for FITElnet -> www.furukawa.co.jp/fitelnet

The netfilter, which is a popular packet filtering mechanism for

Linux, also adopts port overlapping behavior.

5. Address Pooling Paired (APP)

[RFC4787]: REQ-2 [RFC5382]:ND Address Pooling Paired behavior for NAT is recommended in previous documents but behavior when a public IPv4 run out of ports is left undefined. This document clarifies that if APP is enabled new sessions from a subscriber that already has a mapping associated with a public IP that ran out of ports SHOULD be dropped. The administrator MAY provide a knob that allows a NAT device to starting using ports from another public IP when the one that anchored the APP mapping ran out of ports. This is trade-off between subscriber service continuity and APP strict enforcement. (NE: It is sometimes referred as 'soft-APP')

6. EIF Security

[RFC4787]:REQ-8 and [RFC5382]:REQ-3 End-point independent filtering could potentially result in security attacks from the public realm. In order to handle this, when possible there MUST be strict filtering checks in the inbound direction. A knob SHOULD be provided to limit the number of inbound sessions and a knob SHOULD be provided to enable or disable EIF on a per application basis. This is specially important in the case of Mobile networks where such attacks can consume radio resources and count against the user quota.

7. EIF Protocol Independence

[RFC4787]:REQ-8 and[RFC5382]: REQ-3 Current RFCs do not specify whether EIF mappings are protocol independent. In other words, if a outbound TCP SYN creates a mapping it is left undefined whether inbound UDP packets create sessions and are forwarded. EIF mappings SHOULD be protocol independent in order allow inbound packets for protocols that multiplex TCP and UDP over the same IP: port through the NAT and maintain compatibility with stateful NAT64 RFC6146 [RFC6146]. But the administrator MAY provide a configuration knob to make it protocol dependent.

8. EIF Mapping Refresh

[RFC4787]: REQ-6 [RFC5382]: ND The NAT mapping Refresh direction MAY have a "NAT Inbound refresh behavior" of "True" but it does not clarifies how this applies to EIF mappings. The issue in question is whether inbound packets that match an EIF mapping but do not create a

new session due to a security policy should refresh the mapping timer. This document clarifies that even when a NAT device has a inbound refresh behavior of TRUE, that such packets SHOULD NOT refresh the mapping. Otherwise a simple attack of a packet every 2 minutes can keep the mapping indefinitely.

8.1. Outbound Mapping Refresh and Error Packets

In the case of NAT outbound refresh behavior there might be certain types of packets that should not refresh the mapping. For example, if the mapping is kept alive by ICMP Error or TCP RST outbound packets sent as response to inbound packets, these SHOULD NOT refresh the mapping.

9. EIM Protocol Independence

[RFC4787] [RFC5382]: REQ-1 Current RFCs do not specify whether EIM are protocol independent. In other words, if a outbound TCP SYN creates a mapping it is left undefined whether outbound UDP can reuse such mapping and create session. On the other hand, Stateful NAT64 [RFC6146] clearly specifies three binding information bases (TCP, UDP, ICMP). This document clarifies that EIM mappings SHOULD be protocol dependent. A knob MAY be provided in order allow protocols that multiplex TCP and UDP over the same source IP and port to use a single mapping.

10. Port Parity

A NAT devices MAY disable port parity preservation for dynamic mappings. Nevertheless, A NAT SHOULD support means to explicitly request to preserve port parity (e.g., [I-D.boucadair-pcp-rtp-rtcp]).

11. Port Randomization

A NAT SHOULD follow the recommendations specified in Section 4 of [RFC6056] especially: "A NAPT that does not implement port preservation [RFC4787] [RFC5382] SHOULD obfuscate selection of the ephemeral port of a packet when it is changed during translation of that packet. A NAPT that does implement port preservation SHOULD obfuscate the ephemeral port of a packet only if the port must be changed as a result of the port being already in use for some other session. A NAPT that performs parity preservation and that must change the ephemeral port during translation of a packet SHOULD obfuscate the ephemeral ports. The algorithms described in this document could be easily adapted such that the parity is preserved

(i.e., force the lowest order bit of the resulting port number to 0 or 1 according to whether even or odd parity is desired)."

12. IP Identification (IP ID)

A NAT SHOULD handle the Identification field of translated IPv4 packets as specified in Section 9 of [I-D.ietf-intarea-ipv4-id-update].

13. ICMP Query Mappings Timeout

Section 3.1 of [RFC5508] says that ICMP Query Mappings are to be maintained by NAT device. However, RFC doesn't discuss about the Query Mapping timeout values. Section 3.2 of that RFC only discusses about ICMP Query Session Timeouts. ICMP Query Mappings MAY be deleted once the last the session using the mapping is deleted.

14. Hairpinning Support for ICMP Packets

[RFC5508]:REQ-7 This requirement specifies that NAT devices enforcing Basic NAT MUST support traversal of hairpinned ICMP Query sessions. This implicitly means that address mappings from external address to internal address (similar to Endpoint Independent Filters) MUST be maintained to allow inbound ICMP Query sessions. If an ICMP Query is received on an external address, NAT device can then translate to an internal IP. [RFC5508]:REQ-7 This requirement specifies that all NAT devices (i.e., Basic NAT as well as NAPT devices) MUST support the traversal of hairpinned ICMP Error messages. This too requires NAT devices to maintain address mappings from external IP address to internal IP address in addition to the ICMP Query Mappings described in section 3.1 of that RFC.

15. IANA Considerations

TBD

16. Security Considerations

In the case of EIF mappings due to high risk of resource crunch, a NAT device MAY provide a knob to limit the number of inbound sessions spawned from a EIF mapping.

[TCP-Security] contains a detailed discussion of the security

implications of TCP Timestamps and of different timestamp generation algorithms.

17. Acknowledgements

Thanks to Dan Wing, Suresh Kumar, Mayuresh Bakshi, Rajesh Mohan and Senthil Sivamular for review and discussions

18. References

18.1. Normative References

- [I-D.ietf-pcp-base]
Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-29 (work in progress), November 2012.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.
- [RFC1122] Braden, R., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, October 1989.
- [RFC1323] Jacobson, V., Braden, B., and D. Borman, "TCP Extensions for High Performance", RFC 1323, May 1992.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.
- [RFC3605] Huitema, C., "Real Time Control Protocol (RTCP) attribute in Session Description Protocol (SDP)", RFC 3605, October 2003.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC5508] Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT

Behavioral Requirements for ICMP", BCP 148, RFC 5508, April 2009.

- [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.
- [RFC6191] Gont, F., "Reducing the TIME-WAIT State Using TCP Timestamps", BCP 159, RFC 6191, April 2011.

18.2. Informative References

- [FLOWRATE] Zhang, Y., Breslau, L., Paxson, V., and S. Shenker, "On the Characteristics and Origins of Internet Flow Rates".
- [I-D.boucadair-pcp-rtp-rtcp] Boucadair, M. and S. Sivakumar, "Reserving N and N+1 Ports with PCP", draft-boucadair-pcp-rtp-rtcp-05 (work in progress), October 2012.
- [I-D.ietf-behave-lsn-requirements] Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common requirements for Carrier Grade NATs (CGNs)", draft-ietf-behave-lsn-requirements-10 (work in progress), December 2012.
- [I-D.naito-nat-resource-optimizing-extension] Kengo, K. and A. Matsumoto, "NAT TIME_WAIT reduction", draft-naito-nat-resource-optimizing-extension-02 (work in progress), July 2012.
- [TCPWILD] Qian, F., Subhabrata, S., Spatscheck, O., Morley Mao, Z., and W. Willinger, "TCP Revisited: A Fresh Look at TCP in the Wild".

Authors' Addresses

Reinaldo Penno
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, California 95134
USA

Email: repenno@cisco.com

Simon Perreault
Viagenie
2875 boul. Laurier, suite D2-630
Quebec, QC G1V 2M2
Canada

Email: simon.perreault@viagenie.ca

Sarat Kamiset
Consultant
California

Phone:
Fax:

Mohamed Boucadair
France Telecom
Rennes, 35000
France

Email: mohamed.boucadair@orange.com

Kengo Naito
NTT
Tokyo
Japan

Email: kengo@lab.ntt.co.jp

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2013

S. Perreault
Viagenie
T. Tsou
Huawei Technologies (USA)
S. Sivakumar
Cisco Systems
July 9, 2012

Managed Objects for Carrier Grade NAT (CGN)
draft-perreault-sunset4-cgn-mib-00

Abstract

This memo defines a portion of the Management Information Base (MIB) that may be used for monitoring of a device capable of Carrier Grade NAT function.

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 10, 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. Terminology	3
3. Overview	3
4. Definitions	3
5. Security Considerations	9
6. IANA Considerations	9
7. Normative References	9
Authors' Addresses	9

1. Introduction

[I-D.ietf-behave-nat-mib] defines objects for managing network address translators (NATs). This document builds on top of it, defining objects specifically for Carrier Grade NATs (CGN).

2. Terminology

The "CGN" term is defined in [I-D.ietf-behave-lsn-requirements].

3. Overview

New features in this module are as follows:

Per-subscriber counters, limits, and notifications: Carrier-Grade NATs operate with a notion of "subscriber", to which are associated a set of counters, limits, and notifications. The subscriber identifier may not necessarily be an internal address, as in the case of DS-Lite, where the identifier is the IPv6 address of the tunnel endpoint and the internal addresses are the same for each subscriber.

4. Definitions

The following objects are added to the MIB module defined in [I-D.ietf-behave-nat-mib].

-- notifications

```
newNatNotifSubscriberMappings NOTIFICATION-TYPE
  OBJECTS { newNatSubscriberCntMappings }
  STATUS current
  DESCRIPTION
    "This notification is generated when newNatSubscriberCntMappings
     exceeds the value of newNatSubscriberMapNotifyThresh, unless
     newNatSubscriberMapNotifyThresh is zero.."
  ::= { newNatNotifications 5 }
```

-- limits

```
newNatLimitSubscribers OBJECT-TYPE
  SYNTAX Unsigned32
  MAX-ACCESS read-write
  STATUS current
```

```

DESCRIPTION
    "Global limit on the number of subscribers with active mappings.
    Zero means unlimited."
 ::= { newNatLimits 6 }

-- subscribers

newNatSubscribers OBJECT IDENTIFIER ::= { newNatObjects 5 }

newNatSubscribersTable OBJECT-TYPE
    SYNTAX SEQUENCE OF NewNatSubscribersTableEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "Table of CGN subscribers."
    ::= { newNatSubscribers 1 }

newNatSubscribersTableEntry OBJECT-TYPE
    SYNTAX NewNatSubscribersTableEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "Each entry describes a single CGN subscriber."
    INDEX { newNatSubscriberIdentifierType,
            newNatSubscriberIdentifier }
    ::= { newNatSubscribersTable 1 }

NewNatSubscribersTableEntry ::=
    SEQUENCE {
        newNatSubscriberIdentifierType    InetAddressType,
        newNatSubscriberIdentifier        InetAddress,
        newNatSubscriberIntPrefixType     InetAddressType,
        newNatSubscriberIntPrefix         InetAddress,
        newNatSubscriberIntPrefixLength   InetAddressPrefixLength,
        newNatSubscriberPool              NatPoolIndex,
        newNatSubscriberCntTranslates     Counter64,
        newNatSubscriberCntOOP            Counter64,
        newNatSubscriberCntResource       Counter64,
        newNatSubscriberCntStateMismatch Counter64,
        newNatSubscriberCntQuota          Counter64,
        newNatSubscriberCntMappings       Gauge32,
        newNatSubscriberCntMapCreations   Counter64,
        newNatSubscriberCntMapRemovals    Counter64,
        newNatSubscriberLimitMappings     Unsigned32,
        newNatSubscriberMapNotifyThresh   Unsigned32
    }

```

```
newNatSubscriberIdentifierType OBJECT-TYPE
    SYNTAX InetAddressType
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "Address type of the subscriber identifier."
    ::= { newNatSubscribersTableEntry 1 }

newNatSubscriberIdentifier OBJECT-TYPE
    SYNTAX InetAddress (SIZE (4|16))
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "Address used for uniquely identifying the subscriber.

        In traditional NAT, this is the internal address assigned to
        the CPE. In case an address range is assigned to a subscriber,
        the first address in the range is used as identifier. For
        tunnelled connectivity (e.g., DS-Lite [RFC6333]), the outer
        address is used as identifier (i.e., the IPv6 address in the
        case of DS-Lite)."
    ::= { newNatSubscribersTableEntry 2 }

newNatSubscriberIntPrefixType OBJECT-TYPE
    SYNTAX InetAddressType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Subscriber's internal prefix type."
    ::= { newNatSubscribersTableEntry 3 }

newNatSubscriberIntPrefix OBJECT-TYPE
    SYNTAX InetAddress
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Prefix assigned to a subscriber's CPE."
    ::= { newNatSubscribersTableEntry 4 }

newNatSubscriberIntPrefixLength OBJECT-TYPE
    SYNTAX InetAddressPrefixLength
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Length of the prefix assigned to a subscriber's CPE, in bits.
        In case a single address is assigned, this will be 32 for IPv4
        and 128 for IPv6."
    ::= { newNatSubscribersTableEntry 5 }
```

newNatSubscriberPool OBJECT-TYPE

SYNTAX NatPoolIndex

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"External address pool to which this subscriber belongs."

::= { newNatSubscribersTableEntry 6 }

newNatSubscriberCntTranslates OBJECT-TYPE

SYNTAX Counter64

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The number of packets received from or sent to this subscriber and to which NAT has been applied."

::= { newNatSubscribersTableEntry 7 }

newNatSubscriberCntOOP OBJECT-TYPE

SYNTAX Counter64

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The number of packets received from this subscriber to which NAT could not be applied because no external port was available, excluding quota limitations."

::= { newNatSubscribersTableEntry 8 }

newNatSubscriberCntResource OBJECT-TYPE

SYNTAX Counter64

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The number of packets received from this subscriber to which NAT could not be applied because of resource constraints (excluding out-of-ports condition)."

::= { newNatSubscribersTableEntry 9 }

newNatSubscriberCntStateMismatch OBJECT-TYPE

SYNTAX Counter64

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The number of packets received from or destined to this subscriber to which NAT could not be applied because of mapping state mismatch. For example, a TCP packet that matches an existing mapping but is dropped because its flags are incompatible with the current state of the mapping would cause this counter to be incremented."


```
 ::= { newNatSubscribersTableEntry 10 }

newNatSubscriberCntQuota OBJECT-TYPE
    SYNTAX Counter64
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The number of packets received from or destined to this
        subscriber to which NAT could not be applied because of quota
        limitations. Quotas include absolute limits as well as limits
        on the rate of allocation."
    ::= { newNatSubscribersTableEntry 11 }

newNatSubscriberCntMappings OBJECT-TYPE
    SYNTAX Gauge32
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Number of currently active mappings created by or for this
        subscriber.

        Equal to newNatSubscriberCntMapRemovals -
        newNatSubscriberCntMapCreations."
    ::= { newNatSubscribersTableEntry 12 }

newNatSubscriberCntMapCreations OBJECT-TYPE
    SYNTAX Counter64
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Number of mappings created by or for this subscriber."
    ::= { newNatSubscribersTableEntry 13 }

newNatSubscriberCntMapRemovals OBJECT-TYPE
    SYNTAX Counter64
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Number of mappings removed by or for this subscriber."
    ::= { newNatSubscribersTableEntry 14 }

newNatSubscriberLimitMappings OBJECT-TYPE
    SYNTAX Unsigned32
    MAX-ACCESS read-write
    STATUS current
    DESCRIPTION
        "Limit on the number of active mappings created by or for this
        subscriber. Zero means unlimited."
```

```
 ::= { newNatSubscribersTableEntry 15 }

newNatSubscriberMapNotifyThresh OBJECT-TYPE
    SYNTAX Unsigned32
    MAX-ACCESS read-write
    STATUS current
    DESCRIPTION
        "See newNatNotifSubscriberMappings."
    ::= { newNatSubscribersTableEntry 16 }

-- conformance groups

newNatGroupSubscriberObjects OBJECT-GROUP
    OBJECTS { newNatSubscriberIntPrefixType,
               newNatSubscriberIntPrefix,
               newNatSubscriberIntPrefixLength,
               newNatSubscriberPool,
               newNatSubscriberCntTranslates,
               newNatSubscriberCntOOP,
               newNatSubscriberCntResource,
               newNatSubscriberCntStateMismatch,
               newNatSubscriberCntQuota,
               newNatSubscriberCntMappings,
               newNatSubscriberCntMapCreations,
               newNatSubscriberCntMapRemovals,
               newNatSubscriberLimitMappings,
               newNatSubscriberMapNotifyThresh,
               newNatLimitSubscribers }
    STATUS current
    DESCRIPTION
        "Per-subscriber counters, limits, and thresholds."
    ::= { newNatGroups 4 }

-- compliance statements

newNatCGNCompliance MODULE-COMPLIANCE
    STATUS current
    DESCRIPTION
        "NATs that have 'Paired IP address pooling' and 'Receive
        Fragments Out of Order' behavior [RFC4787] and implement the
        objects in this group can claim this level of compliance.

        This level of compliance is to be expected of a CGN compliant
        with [I-D.ietf-behave-lsn-requirements]."
```

```
MODULE -- this module
    MANDATORY-GROUPS { newNatGroupBasicObjects,
```

```
newNatGroupBasicNotifications,  
newNatGroupAddrMapObjects,  
newNatGroupAddrMapNotifications,  
newNatGroupFragmentObjects,  
newNatGroupSubscriberObjects,  
newNatGroupSubscriberNotifs }  
 ::= { newNatCompliance 4 }
```

5. Security Considerations

TBD

6. IANA Considerations

TBD

7. Normative References

[I-D.ietf-behave-lsn-requirements]
Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A.,
and H. Ashida, "Common requirements for Carrier Grade NATs
(CGNs)", draft-ietf-behave-lsn-requirements-07 (work in
progress), June 2012.

[I-D.ietf-behave-nat-mib]
Perreault, S., Tsou, T., and S. Sivakumar, "Additional
Managed Objects for Network Address Translators (NAT)",
draft-ietf-behave-nat-mib-01 (work in progress),
June 2012.

Authors' Addresses

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: simon.perreault@viagenie.ca
URI: <http://viagenie.ca>

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Senthil Sivakumar
Cisco Systems
7100-8 Kit Creek Road
Research Triangle Park, North Carolina 27709
USA

Phone: +1 919 392 5158
Email: ssenthil@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2013

S. Perreault
Viagenie
W. George
Time Warner Cable
T. Tsou
Huawei Technologies (USA)
July 9, 2012

Turning off IPv4 Using DHCPv6
draft-perreault-sunset4-noipv4-00

Abstract

This memo defines a new DHCPv6 option for indicating to a dual-stack host or router that IPv4 is to be turned off.

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 10, 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. Terminology	3
3. The No-IPv4 Option	3
3.1. Wire Format	3
3.2. Semantics	4
3.3. Example	6
4. Open Issues	7
5. Security Considerations	7
6. IANA Considerations	7
7. References	7
7.1. Normative References	7
7.2. Informative References	8
Authors' Addresses	8

1. Introduction

When a dual-stack host makes a DHCPv4 request, it typically interprets the absence of a response as a failure condition. This makes it difficult to deploy such nodes in an IPv6-only network.

Take for example a home router that is dual-stack capable but provisioned with an IPv6-only WAN connection. When the router boots, it typically assigns an IPv4 address to its LAN interface, starts services on that interface, and starts handing out IPv4 addresses to clients on the LAN by answering DHCPv4 requests. This is done unconditionally, without taking the status of the IPv4 connectivity on the WAN interface into account. Hosts on the LAN, in turn, install a default route pointing to the router and start behaving as if IPv4 connectivity was available. IPv4 packets destined to the Internet get dropped at the router and timeouts happen. The end result is that IPv4 remains fully active on the LAN and on the router itself even when it is desired that it be turned off.

A new mechanism is needed to indicate the absence of IPv4 connectivity. Given that the goal is turning off IPv4, this new signaling mechanism shall be transported over IPv6. Therefore, we introduce a new DHCPv6 [RFC3315] option for the purpose of explicitly indicating to the DHCPv6 client that IPv4 connectivity is unavailable.

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].

The following terms are also used in this document:

Upstream Interface: An interface on which the No-IPv4 DHCPv6 option is received by a DHCPv6 client.

3. The No-IPv4 Option

3.1. Wire Format

The No-IPv4 DHCPv6 option is used to signal the unavailability of IPv4 connectivity. The format of the No-IPv4 option is:


```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               OPTION_NO_IPV4               | option-len |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   v4-level   |
+---+---+---+---+---+

```

option-code OPTION_NO_IPV4 (TBD).

option-len 1.

v4-level Level of IPv4 functionality.

The DHCPv6 client MUST place the OPTION_NO_IPV4 option code in the Option Request Option ([RFC3315] section 22.7). Servers MAY include the option in responses (if they have been so configured). Servers MAY also place the OPTION_NO_IPV4 option code in an Option Request Option contained in a Reconfigure message.

3.2. Semantics

The option applies to the link on which it is received by the DHCPv6 client. It is used to indicate to the client that it should disable some or all of its IPv4 functionality. What should be disabled depends on the value of v4-level.

v4-level can take the following values:

- 0 - IPv4 fully enabled: This is equivalent to the absence of the No-IPv4 option. It is included here so that a DHCPv6 server can explicitly re-enable IPv4 access by including it in a Reply message following a Reconfigure.
- 1 - No IPv4 upstream, local IPv4 permitted: Any kind of IPv4 connectivity is unavailable on the link on which the option is received. Therefore, any attempts to provision IPv4 by the host or to use IPv4 in any fashion, on that link, will be useless. IPv4 MAY be dropped, blocked, or otherwise ignored on that link.

Upon reception of the No-IPv4 option with value 1, the following IPv4 functionality MUST be disabled on the Upstream Interface:

- A. IPv4 addresses MUST NOT be assigned.
- B. Currently-assigned IPv4 addresses MUST be unassigned.

- C. Dynamic configuration of link-local IPv4 addresses [RFC3927] MUST be disabled.
- D. IPv4, ICMPv4, or ARP packets MUST NOT be sent.
- E. IPv4, ICMPv4, or ARP packets received MUST be ignored.
- F. DNS A queries MUST NOT be sent, even transported over IPv6.

If all DHCPv6-configured interfaces receive the No-IPv4 option with value 1 or 2, and no other interface provides IPv4 connectivity to the Internet, IPv4 is partially shut down, leaving only local connectivity active. On the Upstream Interface, IPv4 MUST be shut down as listed above. On other interfaces, IPv4 addresses MUST NOT be assigned except for the following:

- * Loopback (127.0.0.0/8)
- * Link Local (169.254.0.0/16) [RFC3927]
- * Private-Use (10.0.0.0/8, 172.16.0.0/12, 192.168.0.0/16) [RFC1918]

- 2 - No IPv4 at all: This is intended to be a stricter version of the above.

The host or router running the DHCPv6 client that receives this option MUST disable IPv4 functionality on the Upstream Interface in the same way as for value 1.

If all DHCPv6-configured interfaces received the No-IPv4 option with exclusively value 2, and no other interface provides IPv4 connectivity to the Internet, IPv4 is completely shut down. In particular:

- A. IPv4 address MUST NOT be assigned to any interface.
- B. Currently-assigned IPv4 addresses MUST be unassigned.
- C. Dynamic configuration of link-local IPv4 addresses [RFC3927] MUST be disabled.
- D. IPv4, ICMPv4, or ARP packets MUST NOT be sent on any interface.
- E. IPv4, ICMPv4, or ARP packets received on any interface MUST be ignored.

- F. In the above, "any interface" includes loopback interfaces. In particular, the 127.0.0.1 special address MUST be removed.
- G. Server programs listening on IPv4 addresses (e.g., a DHCPv4 server) MAY be shut down.
- H. DNS A queries MUST NOT be sent, even transported over IPv6.
- I. If the host or router also runs a DHCPv6 server, it SHOULD include the No-IPv4 option with value 2 in DHCPv6 responses it sends to clients that request it, unless prohibited by local policy. If it currently has active clients, it SHOULD send a Reconfigure to each of them with the OPTION_NO_IPV4 included in the Option Request Option.

The intent is to remove all traces of IPv4 activity. Once the No-IPv4 option with value 2 is activated, the network stack should behave as if IPv4 functionality had never been present. For example, a modular kernel implementation could accomplish the above by unloading the IPv4 kernel module at run time.

3.3. Example

A dual-stack home gateway is set up with a single WAN uplink and is configured to use DHCPv4 and DHCPv6 to automatically obtain IPv4 and IPv6 connectivity. On the LAN side, it has one link with multiple hosts.

When it boots, the router assigns 192.168.1.1/24 to its LAN interfaces and starts a DHCPv4 server listening on it. It hands out addresses 191.168.1.100-199 to clients. It also starts an IPv6 Router Advertisement daemon as well as a stateless DHCPv6 server, also listening on the LAN interfaces.

On the WAN side, it starts two provisioning procedures in parallel: one for IPv4 and one for IPv6.

At this point, the ISP does not know if the router supports IPv6-only operation. Therefore, by default, the ISP responds to DHCPv4 requests as usual.

As part of the IPv6 provisioning procedure, the router sends a DHCPv6 request containing OPTION_NO_IPV4 in an Option Request Option. The ISP's DHCPv6 server's reply includes the No-IPv4 option with value 2. When this procedure finishes, the ISP has determined that this customer will run in IPv6-only mode and starts dropping all IPv4 packets at the first hop. If an IPv4 address was assigned, it is reclaimed, and possibly reassigned to another subscriber.

The home router aborts the IPv4 provisioning procedure (if it is still running) and deactivates all IPv4 functionality. It shuts down its DHCPv4 server. It also configures its own stateless DHCPv6 server to send the No-IPv4 option to clients that request it.

As an optimization, the router could delay setting up IPv4 by a few seconds (10 seconds seems reasonable). If the IPv6 procedure completes with the No-IPv4 option during that time, IPv4 will never have been set up and the router will operate in pure IPv6-only mode from the start.

4. Open Issues

- o A legacy IPv4-only device connected to a network running in mode 2 (no IPv4 at all) will presumably keep retrying forever, e.g. sending DHCPDISCOVER messages endlessly. Do we want a way to signal to that host that IPv4 will never be available? But since that device was not updated for IPv6, it is doubtful that it would be updated to understand this new signaling. Could we reuse/overload some existing signaling that would have the same effect?

5. Security Considerations

One security concern is that an attacker could use the No-IPv4 option to deny IPv4 access to a victim. However, unprotected vanilla DHCP can already be exploited to cause such a denial of service ([RFC2131] section 7).

TO BE COMPLETED

6. IANA Considerations

IANA is requested to assign value TBD with description OPTION_NO_IPV4 in the "DHCP Option Codes" table which is part of the dhcpv6-parameters registry [1].

7. References

7.1. Normative References

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

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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.
- [RFC3927] Cheshire, S., Aboba, B., and E. Guttman, "Dynamic Configuration of IPv4 Link-Local Addresses", RFC 3927, May 2005.

7.2. Informative References

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

URIs

- [1] <<http://www.iana.org/assignments/dhcpv6-parameters>>

Authors' Addresses

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: simon.perreault@viagenie.ca
URI: <http://viagenie.ca>

Wes George
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA 20171
USA

Email: wesley.george@twcable.com

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 16, 2014

S. Perreault
Viagenie
W. George
Time Warner Cable
T. Tsou
Huawei Technologies (USA)
T. Yang
L. Li
China Mobile
July 15, 2013

Turning off IPv4 Using DHCPv6 or Router Advertisements
draft-perreault-sunset4-noipv4-03

Abstract

This memo defines a new DHCPv6 option and a new Router Advertisement option for indicating to a dual-stack host or router that IPv4 is to be turned off.

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 16, 2014.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. The Problems We're Trying to Fix	4
3.1. Load on DHCPv4 Server	4
3.2. Bandwidth Consumption	4
3.3. Power Inefficiency	4
3.4. IPv4 only Applications	4
4. Design Considerations	4
4.1. DHCPv6 vs DHCPv4	4
4.2. DHCPv6 vs RA	5
5. The No-IPv4 Option	6
5.1. DHCPv6 Wire Format	6
5.2. RA Wire Format	6
5.3. Semantics	7
5.4. Example	9
6. Security Considerations	10
7. IANA Considerations	10
8. Acknowledgements	10
9. References	10
9.1. Normative References	10
9.2. Informative References	11
Appendix A. Test Results of Terminals Behavior	11
Authors' Addresses	12

1. Introduction

When a dual-stack host makes a DHCPv4 request, it typically interprets the absence of a response as a failure condition. This makes it difficult to deploy such nodes in an IPv6-only network.

Take for example a home router that is dual-stack capable but provisioned with an IPv6-only WAN connection. When the router boots, it typically assigns an IPv4 address to its LAN interface, starts services on that interface, and starts handing out IPv4 addresses to clients on the LAN by answering DHCPv4 requests. This is done unconditionally, without taking the status of the IPv4 connectivity on the WAN interface into account. Hosts on the LAN, in turn, install a default route pointing to the router and start behaving as if IPv4 connectivity was available. IPv4 packets destined to the Internet get dropped at the router and timeouts happen. The end result is that IPv4 remains fully active on the LAN and on the router itself even when it is desired that it be turned off.

The other example is about DHCPv4 server. In Dual-Stack LAN/WLAN network or intranet, the core router or AC often plays the role of DHCP server, and the clients are server thousands PC or mobile phones. If the server is configured in IPv6-only, the dual-stack or IPv4-only clients will broadcast DHCPDISCOVER messages endlessly in the LAN or WLAN. The thousands clients will cause a DDOS-like attack to all the servers in the network. Since there are not specific descriptions in any RFCs for client's behavior when it does not receive the DHCPOFFER in response to its DHCPDISCOVER message, various OS deploy different backoff algorithms. We tested server popular OS(es), the test results is listed in the appendix.

A new mechanism is needed to indicate the absence of IPv4 connectivity or service that the goal is turning off IPv4, this new signaling mechanism shall be transported over IPv6. Therefore, we introduce a new DHCPv6 [RFC3315] option and a new Router Advertisement (RA) [RFC4861] option for the purpose of explicitly indicating to the host that IPv4 connectivity is unavailable.

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].

The following terms are also used in this document:

Upstream Interface: An interface on which the No-IPv4 option is received over either DHCPv6 or RA.

3. The Problems We're Trying to Fix

3.1. Load on DHCPv4 Server

When a DHCPv4 server is present but intentionally does not respond to a dual-stack node, the aggregated traffic generated by multiple such dual-stack nodes can represent a significant useless load. This scenario can be encountered for example with an ISP serving multiple types of subscribers where some will get IPv4 addresses and others not. It might not be feasible for operational reasons to block the useless requests before they reach the DHCPv4 server, e.g. if the DHCPv4 server itself is the one that has knowledge about which node should or should not get an IPv4 address.

3.2. Bandwidth Consumption

In addition to useless load on the DHCPv4 server, the above scenario could also consume a significant amount of bandwidth, particularly if the aggregated traffic from many clients goes through a low-bandwidth link.

3.3. Power Inefficiency

A dual-stack node that does not get a DHCPv4 response will usually continue retransmitting forever. Therefore, only providing IPv6 on a link will cause the node to needlessly wake up periodically and transmit a few packets. For example, the popular DHCPv4 client implementation by ISC wakes up every 5 minutes by default and tries to contact a DHCPv4 server for 60 seconds. With this configuration, a node will not be able to sleep 20% of the time.

3.4. IPv4 only Applications

In many cases, IPv4-only applications such as Skype use IPv4 LLA to bombard the LAN with IPv4 packets. In an IPv6-only environment, it can get quite annoying and waste a lot of bandwidth.

4. Design Considerations

4.1. DHCPv6 vs DHCPv4

NOTE: This section will be removed before publication as an RFC.

This document describes a new DHCPv6 option for turning off IPv4. An equivalent option could conceivably be created for DHCPv4. Here is a discussion of the pros and cons. Arguments with a + sign argue for a DHCPv4 option, arguments with a - sign argue against.

- + Devices that don't speak IPv6 won't be listening for a "turn off IPv4" code, and therefore won't stop trying to establish IPv4 connectivity.
- Devices that haven't been updated to speak IPv6 likely won't recognize a new DHCPv4 code telling them that IPv4 isn't supported.
 - + However, it's easier to implement something that turns off the IP stack than implement IPv6.
- Devices that don't speak IPv6 that are still active on the network mean that either IPv4 can't/shouldn't be turned off yet, or IPv4 local connectivity should be maintained to retain local services, even if global IPv4 connectivity is not necessary (think local LAN DLNA streaming, etc).
- When the goal is to turn off IPv4, having to maintain and operate an IPv4 infrastructure (routing, ACLs, etc.) just to be able to send negative responses to DHCPv4 requests is not productive. Having the option transported in IPv6 allows the ISP to focus on operating an IPv6-only network.
 - + However, a full IPv4 infrastructure would not be necessary in many cases. The local router could contain a very restricted DHCPv4 server function whose only purpose would be to reply with the No-IPv4 option. No IPv4 traffic would have to be carried to a distant DHCPv4 server. Note however that this may not be operationally feasible in some situations.
- Turning IPv4 off using an IPv4-transported signal means that there is no way to go back. Once the DHCPv4 option has been accepted by the DHCPv4 client, IPv4 can no longer be turned on remotely (rebooting the client still works). Configurations change, mistakes happen, and so it is necessary to have a way to turn IPv4 back on. With a DHCPv6 option, IPv4 can be turned back on as soon as the client makes a new DHCPv6 request, which can be the next scheduled one or can be triggered immediately with a Reconfigure message.

The authors conclude that a DHCPv6 option is clearly necessary, whereas it is not as clear for a DHCPv4 option. More feedback on this topic would be appreciated.

4.2. DHCPv6 vs RA

Both DHCPv6- and RA-based solutions are presented in this draft. It is expected that the working group will decide whether both solutions, only one, or none are desirable.

5. The No-IPv4 Option

The No-IPv4 DHCPv6 option is used to signal the unavailability of IPv4 connectivity.

5.1. DHCPv6 Wire Format

The format of the DHCPv6 No-IPv4 option is:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|               OPTION_NO_IPV4               | option-len |
+-----+-----+-----+-----+-----+-----+-----+-----+
|    v4-level    |
+-----+-----+-----+-----+-----+-----+-----+

```

option-code OPTION_NO_IPV4 (TBD).

option-len 1.

v4-level Level of IPv4 functionality.

The DHCPv6 client MUST place the OPTION_NO_IPV4 option code in the Option Request Option ([RFC3315] section 22.7). Servers MAY include the option in responses (if they have been so configured). Servers MAY also place the OPTION_NO_IPV4 option code in an Option Request Option contained in a Reconfigure message.

5.2. RA Wire Format

The format of the RA No-IPv4 option is:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|    Type    |    Length    |    v4-level    |    Reserved    |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Reserved               |
+-----+-----+-----+-----+-----+-----+-----+

```

Type TBD

Length	1.
v4-level	Level of IPv4 functionality.
Reserved	These fields are unused. They MUST be initialized to zero by the sender and MUST be ignored by the receiver.

5.3. Semantics

The option applies to the link on which it is received. It is used to indicate to the client that it should disable some or all of its IPv4 functionality. What should be disabled depends on the value of v4-level.

v4-level can take the following values:

- 0 - IPv4 fully enabled: This is equivalent to the absence of the No-IPv4 option. It is included here so that a DHCPv6 server can explicitly re-enable IPv4 access by including it in a Reply message following a Reconfigure, or similarly by a router in a spontaneous Router Advertisement.
- 1 - No IPv4 upstream: Any kind of IPv4 connectivity is unavailable on the link on which the option is received. Therefore, any attempts to provision IPv4 by the host or to use IPv4 in any fashion, on that link, will be useless. IPv4 MAY be dropped, blocked, or otherwise ignored on that link.

Upon reception of the No-IPv4 option with value 1, the following IPv4 functionality MUST be disabled on the Upstream Interface:

- a. IPv4 addresses MUST NOT be assigned.
 - b. Currently-assigned IPv4 addresses MUST be unassigned.
 - c. Dynamic configuration of link-local IPv4 addresses [RFC3927] MUST be disabled.
 - d. IPv4, ICMPv4, or ARP packets MUST NOT be sent.
 - e. IPv4, ICMPv4, or ARP packets received MUST be ignored.
 - f. DNS A queries MUST NOT be sent, even transported over IPv6.
- 2 - No IPv4 upstream, local IPv4 restricted: Same semantics as value 1, with the following additions:

If all DHCPv6- or RA-configured interfaces receive the No-IPv4 option with a mix of values 1, 2, and 3 (but not exclusively 3), and no other interface provides IPv4 connectivity to the Internet, IPv4 is partially shut down, leaving only local connectivity active. On the Upstream Interface, IPv4 MUST be shut down as listed above. On other interfaces, IPv4 addresses MUST NOT be assigned except for the following:

- * Loopback (127.0.0.0/8)
- * Link Local (169.254.0.0/16) [RFC3927]
- * Private-Use (10.0.0.0/8, 172.16.0.0/12, 192.168.0.0/16) [RFC1918]

- 3 - No IPv4 at all: This is intended to be a stricter version of the above.

The host or router receiving this option MUST disable IPv4 functionality on the Upstream Interface in the same way as for value 1 or 2.

If all DHCPv6- or RA-configured interfaces received the No-IPv4 option with exclusively value 3, and no other interface provides IPv4 connectivity to the Internet, IPv4 is completely shut down. In particular:

- a. IPv4 address MUST NOT be assigned to any interface.
- b. Currently-assigned IPv4 addresses MUST be unassigned.
- c. Dynamic configuration of link-local IPv4 addresses [RFC3927] MUST be disabled.
- d. IPv4, ICMPv4, or ARP packets MUST NOT be sent on any interface.
- e. IPv4, ICMPv4, or ARP packets received on any interface MUST be ignored.
- f. In the above, "any interface" includes loopback interfaces. In particular, the 127.0.0.1 special address MUST be removed.
- g. Server programs listening on IPv4 addresses (e.g., a DHCPv4 server) MAY be shut down.
- h. DNS A queries MUST NOT be sent, even transported over IPv6.

- i. If the host or router also runs a DHCPv6 server, it SHOULD include the No-IPv4 option with value 2 in DHCPv6 responses it sends to clients that request it, unless prohibited by local policy. If it currently has active clients, it SHOULD send a Reconfigure to each of them with the OPTION_NO_IPV4 included in the Option Request Option.
- j. If the router sends Router Advertisement, it SHOULD include the No-IPv4 option with value 2 in RA messages it sends, unless prohibited by local policy. It SHOULD also send RAs immediately so that the changes take effect for all current hosts.

The intent is to remove all traces of IPv4 activity. Once the No-IPv4 option with value 3 is activated, the network stack should behave as if IPv4 functionality had never been present. For example, a modular kernel implementation could accomplish the above by unloading the IPv4 kernel module at run time.

5.4. Example

A dual-stack home gateway is set up with a single WAN uplink and is configured to use DHCPv4 and DHCPv6 to automatically obtain IPv4 and IPv6 connectivity. On the LAN side, it has one link with multiple hosts.

When it boots, the router assigns 192.168.1.1/24 to its LAN interfaces and starts a DHCPv4 server listening on it. It hands out addresses 191.168.1.100-199 to clients. It also starts an IPv6 Router Advertisement daemon as well as a stateless DHCPv6 server, also listening on the LAN interfaces.

On the WAN side, it starts two provisioning procedures in parallel: one for IPv4 and one for IPv6.

At this point, the ISP does not know if the router supports IPv6-only operation. Therefore, by default, the ISP responds to DHCPv4 requests as usual.

As part of the IPv6 provisioning procedure, the router sends a DHCPv6 request containing OPTION_NO_IPV4 in an Option Request Option. The ISP's DHCPv6 server's reply includes the No-IPv4 option with value 3. When this procedure finishes, the ISP has determined that this customer will run in IPv6-only mode and starts dropping all IPv4 packets at the first hop. If an IPv4 address was assigned, it is reclaimed, and possibly reassigned to another subscriber.

The home router aborts the IPv4 provisioning procedure (if it is still running) and deactivates all IPv4 functionality. It shuts down its DHCPv4 server. It also configures its own stateless DHCPv6 server to send the No-IPv4 option to clients that request it.

As an optimization, the router could delay setting up IPv4 by a few seconds (10 seconds seems reasonable). If the IPv6 procedure completes with the No-IPv4 option during that time, IPv4 will never have been set up and the router will operate in pure IPv6-only mode from the start.

6. Security Considerations

One security concern is that an attacker could use the No-IPv4 option to deny IPv4 access to a victim. However, unprotected vanilla DHCP can already be exploited to cause such a denial of service ([RFC2131] section 7).

TO BE COMPLETED

7. IANA Considerations

IANA is requested to assign value TBD with description `OPTION_NO_IPV4` in the "DHCP Option Codes" table which is part of the `dhcpv6-parameters` registry [1].

IANA is requested to assign value TBD with description "No-IPv4 Option" in the IPv6 Neighbor Discovery Option Formats table which is part of the `icmpv6-parameters` registry.

8. Acknowledgements

Thanks in particular to Marc Blanchet who was the driving force behind this work.

Rajiv Asati contributed section Section 3.4.

9. References

9.1. Normative References

- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

- [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.
- [RFC3927] Cheshire, S., Aboba, B., and E. Guttman, "Dynamic Configuration of IPv4 Link-Local Addresses", RFC 3927, May 2005.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

9.2. Informative References

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

Appendix A. Test Results of Terminals Behavior

In RFC3315 [RFC3315, DHCPv6], SOL_MAX_RT is defined in DHCPv6 to prevent the frequently requesting of clients, which reduces the aggregated traffic. But in RFC2131 [RFC2131, DHCPv4], there are not corresponding IPv4 definitions or options for client's behavior if the server does not respond for the Discover messages.

In fact, most of the terminals creat backoff algorithms to help them retransmit DHCPDISCOVER message in different frequency according to their state machine. The same point of almost all the verious Operating Systems is that they could not stop DHCPDISCOVER requests to the server. And that will cause DDoS-Like attack to the server and bandwidth consumption in the link.

We test some of the most popular terminals' OS in WLAN, the results are illuminated as below.

DHCP Discovery Packages Time Table

No	Windows7		Windows XP		IOS_5.0.1		Android_2.3.7		Symbian_S60	
	Time	Time offset	Time	Time offset	Time	Time offset	Time	Time offset	Time	Time offset
1	0		0		0.1		7.8		0	
2	3.9	3.9	0.1	0.1	1.4	1.3	10.3	2.5	2	2
3	13.3	9.4	4.1	4	3.8	2.4	17.9	7.6	6	4
4	30.5	17.2	12.1	8	7.9	4.1	33.9	16	8	2
5	62.8	32.3	29.1	17	16.3	8.4	36.5	2.6	12	4

6	65.9	3.1	64.9	35.8	24.9	8.6	reconnect		14	2
7	74.9	9	68.9	4	33.4	8.5	56.6	20.1	18	4
8	92.1	17.2	77.9	9	42.2	8.8	60.2	3.6	20	2
9	395.2	303.1	93.9	16	50.8	8.6	68.4	8.2	24	4
10	399.1	3.9	433.9	340	59.1	8.3	84.8	16.4	26	2
11	407.1	8	438.9	5	127.3	68.2	86.7	1.9	30.1	4.1
12	423.4	16.3	447.9	9	128.9	1.6	reconnect		32.1	2
13	455.4	32	464.9	17	131.1	2.2	106.7	20	36.1	4
14	460.4	5	794.9	330	135.1	4	111.4	4.7	38.1	2
15	467.4	7	799.9	5	143.4	8.3	120.6	9.2	42.1	4
16	483.4	16	808.9	9	151.7	8.3	134.9	14.3	44.1	2
17	842.9	359.5	824.9	16	160.4	8.7	136.8	1.9	48.2	4.1
18	846.9	4	1141.9	317	168.8	8.4	reconnect		50.2	2

Figure:Terminals DHCPDISCOVER requests when Server's DHCPv4 module is down

In this figure:

For Windows7, it seems to initiate 8 times DHCPDISCOVER requests in about 300s interval.

For WindowsXP, firstly it launches 9 times DHCPDISCOVER messages, but after that it cannot get any response from the server, then it initiates 5 times requests in one cycle in around 330s intervals, and never stop.

For IOS5.0.1, it seems like WindowsXP. There are 10 times attempts in one cycle, and the interval is about 68s.

Symbian_S60 uses the simplest backoff method, it launches DISCOVER in every 2 or 4 seconds.

Android2.3.7 is the only Operating System which can stop DISCOVER request by disconnect its wireless connection. It reboot wireless and dhcp connection every 20 seconds.

Authors' Addresses

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: simon.perreault@viagenie.ca
URI: <http://viagenie.ca>

Wes George
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA 20171
USA

Email: wesley.george@twcable.com

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Tianle Yang
China Mobile
32, Xuanwumenxi Ave.
Xicheng District, Beijing 100053
China

Email: yangtianle@chinamobile.com

Li Lianyuan
China Mobile
32, Xuanwumenxi Ave.
Xicheng District, Beijing 100053
China

Email: lilianyuan@chinamobile.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 17, 2013

T. Tsou
Huawei Technologies (USA)
W. Liu
Huawei Technologies
S. Perreault
Viagenie
R. Penno
Cisco Systems, Inc.
July 16, 2012

Stateless IPv4 Network Address Translation
draft-tsou-stateless-nat44-01

Abstract

This memo describes a protocol for decentralizing IPv4 NAT to the customer-premises equipment (CPE) such that no state information is kept on the central NAT device. The CPE uses a restricted source port set that is encoded in its provisioned IPv4 WAN address. The NAT device performs only strictly stateless address (not port) translation.

Status of this Memo

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

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

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

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

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. Address Formats	4
4. CPE Provisioning	4
5. SLNAT44 Configuration	5
6. Port Set Computation	5
7. CPE Operation	7
7.1. ALG Handling	7
8. SLNAT44 Operation	7
8.1. Internal to External	7
8.2. External to Internal	7
8.3. Fragment Handling	8
9. Address Mapping Example	8
10. Security Considerations	9
11. Acknowledgements	9
12. References	9
12.1. Normative References	9
12.2. Informative References	9
Authors' Addresses	10

1. Introduction

IPv4 address exhaustion has become world-wide reality. NAT is one of the solutions to deal with the problem. The drawbacks of traditional NAT include statefulness and the need to track transport-layer sessions. This makes NAT complex, hard to scale up, and fragile.

This document describes a method of deploying stateless NAT as a backwards-compatible evolution of an IPv4-only network.

The assumed topology is illustrated in Figure 1.

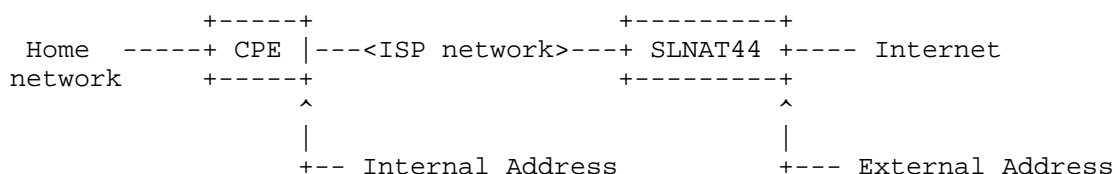


Figure 1: Stateless NAT44 topology

Note that SLNAT44 has no IPv6 component. Any deployment of IPv6 is unaffected by SLNAT44. Therefore, this document only describes IPv4 addresses and IPv4 packets. IPv6 is not discussed further.

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].

The following terms are used throughout this document:

Port set: Set of transport-layer ports that each CPE is assigned, to be used as source ports by packets sent by the CPE.

Port Set ID: A value from which a unique port set is algorithmically derived.

SLNAT44: Depending on the context, either the stateless NAT44 protocol or the stateless NAT44 device that translates between internal and external addresses. NAT44 in turn stands for "IPv4-to-IPv4 NAT".

Internal Address: The IPv4 address assigned to a CPE. It is used in the ISP network between the CPE and the SLNAT44.

External Address: The IPv4 address used on the Internet and routed to the SLNAT44.

Mapping rule: A set of parameters configured on the SLNAT44 (not on the CPE) describing the relationship between internal and external addresses.

3. Address Formats

Internal addresses have the format illustrated in Figure 2. The addresses are simply made of three parts concatenated together: the Internal Prefix, the External Suffix, and the Port Set ID.

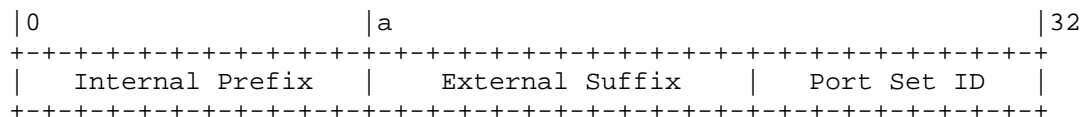


Figure 2: Internal Address format

External Addresses have the format illustrated in Figure 3. It is made of two parts: the External Prefix and the External Suffix.

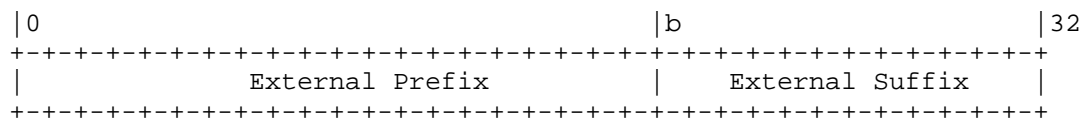


Figure 3: External Address format

The lengths of the Internal and External Prefixes, "a" and "b", are mandatory parameters of SLNAT44. They are determined by the ISP. They need not be communicated to the CPE. Other lengths can be computed from them as follows:

- o Length of External Suffix: 32 - b
- o Length of Port Set ID: b - a

4. CPE Provisioning

As part of its start up routine, the CPE is assigned an IPv4 address by the ISP using regular means (DHCP, PPP, etc.). This is the

Internal Address.

In addition, using new provisioning options, the CPE is assigned a Port Set ID.

Optionally, a Port Set Mask is also provisioned to the CPE. This mask is of the same length as the Port Set ID (i.e., b-a bits). Its purpose is to allow discontinuous port ranges. If no mask is provided, a mask of all ones is assumed by default, which implies a continuous port range. System ports (0-1023) should not to be assigned to any CPE.

In summary, the CPE is provisioned with the following elements:

- o IPv4 address (as usual)
- o Port Set ID
- o Port Set Mask (optional)

5. SLNAT44 Configuration

The SLNAT44 is configured with a set of mapping rules. Each rule contains:

- o Internal Prefix
- o External Prefix
- o Port Set Mask (optional)

Prefixes include their length. For simplicity, rule prefixes MUST NOT overlap with other rules.

If it is absent, the Port Set Mask is assumed to be all ones by default.

6. Port Set Computation

Given a Port Set ID and a Port Set Mask, both n bits in length, the set of allowed ports is defined as the set of port numbers for which the higher-order n bits of their binary expression whose corresponding mask bits are 1 are equal to corresponding bits from the Port Set ID.

```

|0           |5
+---+---+---+
|1 1 1 0 1|   Port Set ID = 29 (length n = 5 bits)
+---+---+---+
& & & & &
+---+---+---+
|1 1 1 1 1|   Port Set Mask
+---+---+---+
| | | | |
V V V V V
+---+---+---+
|1 1 1 0 1 x x x x x x x x x x x x|   Port Set = 59392-61439
+---+---+---+
|0                                     |16

```

Figure 4: Example Contiguous Port Set Computation

```

|0           |8
+---+---+---+
|0 0 1 0 1 1 1 1|   Port Set ID = 29 (length n = 8 bits)
+---+---+---+
& & & & & & &
+---+---+---+
|0 0 1 1 1 1 1 1|   Port Set Mask
+---+---+---+
| | | | | | | |
V V V V V V V V
+---+---+---+
|x x 1 0 1 1 1 1 x x x x x x x x x|   Port Set = 12032-12287, 28416-28671,
+---+---+---+                                     44800-45055, 61184-61439
|0                                     |16

```

Figure 5: Example Non-Contiguous Port Set Computation

It follows that the number of ports in the set is $2^{(16-x)}$, where x is the number of ones in the Port Set Mask.

This computation is performed by the CPE as part of its provisioning routine as well as by the SLNAT44 for dropping packets with ports outside the allowed range.

For the purposes of SLNAT44, a "source port" corresponds to either a TCP source port, a UDP source port, or an ICMPv4 identifier, while a "destination port" corresponds to either a TCP destination port, a UDP destination port, or an ICMPv4 identifier. Note that an ICMPv4 identifier plays the role of both source and destination port.

Transport protocols other than TCP and UDP, as well as ICMPv4 types without an identifier field are out of scope of this specification.

7. CPE Operation

Packets sent from the CPE MUST have the provisioned IPv4 address as source and MUST have a source port that is within the allowed set. This is usually accomplished by having the CPE run a NAT44 configured with the provisioned address and allowed port set and having it process all packets sent out the WAN interface.

Packets received by the CPE on its WAN interface with a destination port outside the allowed range MUST be dropped.

7.1. ALG Handling

If the CPE implements application level gateways (ALGs) such as FTP, RSTP or PPTP, it must ensure that ports present in the payload when translated fall within the allowed range.

8. SLNAT44 Operation

8.1. Internal to External

When it receives a packet on an internal interface, the SLNAT44 finds the rule whose Internal Prefix matches the packet's source address. It extracts the Port Set ID from the packet's source address. It then checks if the packet's source port is within the allowed set, using the rule's Port Set Mask. If it is not, the packet MUST be dropped.

If the packet's source port is within the allowed set, the SLNAT44 builds the External Address by concatenating the rule's External Prefix with the External Suffix extracted from the packet's source address. It then replaces the packet's source address with this External Address. The IPv4 and transport-layer checksums are updated as necessary. The packet is then forwarded as usual.

8.2. External to Internal

When it receives a packet on an external interface, the SLNAT44 finds the rule whose External Prefix matches the packet's destination address. It then builds the Internal Address by concatenating the rule's Internal Prefix, the External Suffix extracted from the packet's destination address, and the Port Set ID computed by applying the rule's Port Set Mask to the packet's destination port's

higher-order bits. It then replaces the packet's destination address with this Internal Address. The IPv4 and transport-layer checksums are updated as necessary. The packet is then forwarded as usual.

8.3. Fragment Handling

If the incoming IP packet contains a fragment, then more processing may be needed. This specification leaves open the exact details of how a SLNAT44 handles incoming IP packets containing fragments, and simply requires that the external behavior of the SLNAT44 be compliant with the following conditions.

The SLNAT44 MUST handle fragments. In particular, SLNAT44 MUST handle fragments arriving out of order, conditional on the following:

- o The SLNAT44 MUST limit the amount of resources devoted to the storage of fragmented packets in order to protect from DoS attacks.
- o As long as the SLNAT44 has available resources, the SLNAT44 MUST allow the fragments to arrive over a time interval. The time interval SHOULD be configurable and the default value MUST be of at least 2 seconds.
- o The SLNAT44 MAY require that the UDP, TCP, or ICMPv4 header be completely contained within the fragment that contains fragment offset equal to zero.

For incoming packets carrying TCP or UDP fragments with a non-zero checksum, SLNAT44 MAY elect to queue the fragments as they arrive and translate all fragments at the same time. In this case, the incoming tuple is determined as documented above to the un-fragmented packets. Alternatively, a SLNAT44 MAY translate the fragments as they arrive, by storing information that allows it to compute the necessary port number for fragments other than the first. In the latter case, subsequent fragments may arrive before the first, and the rules (in the bulleted list above) about how the SLNAT44 handles (out-of-order) fragments apply.

Implementers of SLNAT44 should be aware that there are a number of well-known attacks against IP fragmentation; see [RFC1858] and [RFC3128]. Implementers should also be aware of additional issues with reassembling packets at high rates, described in [RFC4963].

9. Address Mapping Example

An operator has two public ranges of size /18 and /19 called foo and

bar respectively. It plans to use 10/8 as its internal address prefix and PSID (port range) of length 5. Two prefixes of the internal network

The internal prefixes lengths are:

- o $32 - 18 - 5 = 13$ (derived from foo)
- o $32 - 19 - 5 = 14$ (derived from bar)

This will give the following possible mappings:

- o foo/18 <--> 10.0.0.0/13
- o bar/19 <--> 10.128.0.0/14

Author note: Discuss the where internal prefixes are overlapping

10. Security Considerations

The security considerations related to IP address sharing documented in RFC 6269 [RFC6269] and RFC 6056 [RFC6056] apply to SLNAT44.

11. Acknowledgements

Section 8.3 is adapted from [RFC6146].

12. References

12.1. Normative References

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

12.2. Informative References

- [RFC1858] Ziemba, G., Reed, D., and P. Traina, "Security Considerations for IP Fragment Filtering", RFC 1858, October 1995.
- [RFC3128] Miller, I., "Protection Against a Variant of the Tiny Fragment Attack (RFC 1858)", RFC 3128, June 2001.
- [RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, July 2007.

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

Authors' Addresses

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Will Liu
Huawei Technologies
Bantian, Longgang District
Shenzhen 518129
P.R. China

Phone:
Email: Liushucheng@huawei.com

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Email: simon.perreault@viagenie.ca

Reinaldo Penno
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, California 95134
USA

Phone:
Email: repenno@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

T. Tsou
Huawei Technologies (USA)
W. Liu
Huawei Technologies
S. Perreault
Viagenie
R. Penno
Cisco Systems, Inc.
M. Chen
FreeBit
October 22, 2012

Stateless IPv4 Network Address Translation
draft-tsou-stateless-nat44-02

Abstract

This memo describes a protocol for decentralizing IPv4 NAT to the customer-premises equipment (CPE) such that no state information is kept on the central NAT device. The CPE uses a restricted source port set that is encoded in its provisioned IPv4 WAN address. The NAT device performs only strictly stateless address (not port) translation.

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 25, 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. Terminology	3
3. Address Formats	4
4. Customer Provisioning	5
5. SLNAT44 Configuration	6
6. Port Set Computation	6
7. CPE Operation	7
7.1. ALG Handling	8
8. SLNAT44 Operation	8
8.1. Internal to External	8
8.2. External to Internal	8
8.3. Fragment Handling	9
9. Address Mapping Example	9
10. Security Considerations	10
11. Acknowledgements	10
12. References	10
12.1. Normative References	10
12.2. Informative References	10
Authors' Addresses	11

1. Introduction

IPv4 address exhaustion has become world-wide reality. NAT is one of the solutions to deal with the problem. The drawbacks of traditional NAT include statefulness and the need to track transport-layer sessions. This makes NAT complex, hard to scale up, and fragile.

This document describes a method of deploying stateless NAT as a backwards-compatible evolution of an IPv4-only network.

The assumed topology is illustrated in Figure 1.

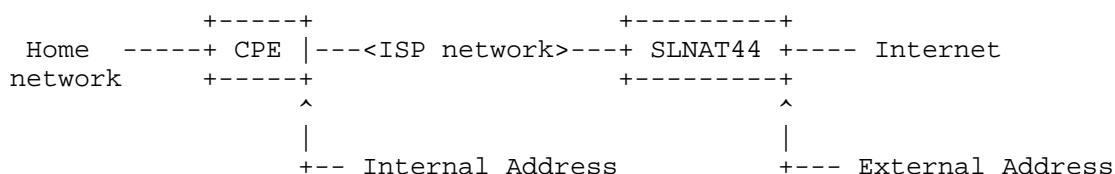


Figure 1: Stateless NAT44 topology

When CPE is configured working as a transparent bridge, internal addresses are directly assigned to the end hosts in the home network, as is shown in Figure 2.

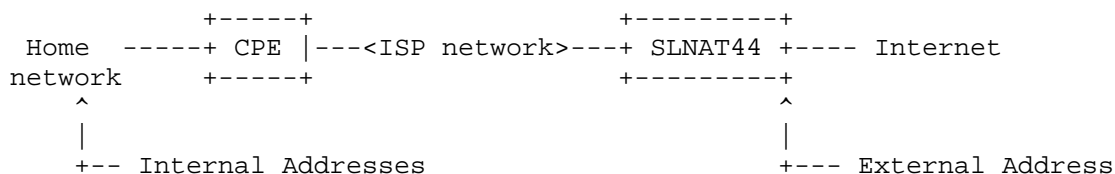


Figure 2: Stateless NAT44 topology: CPE as bridge

Note that SLNAT44 has no IPv6 component. Any deployment of IPv6 is unaffected by SLNAT44. Therefore, this document only describes IPv4 addresses and IPv4 packets. IPv6 is not discussed further.

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].

The following terms are used throughout this document:

Port set: Set of transport-layer ports that each CPE is assigned, to be used as source ports by packets sent by the CPE.

Port Set ID: A value from which a unique port set is algorithmically derived.

SLNAT44: Depending on the context, either the stateless NAT44 protocol or the stateless NAT44 device that translates between internal and external addresses. NAT44 in turn stands for "IPv4-to-IPv4 NAT".

Internal Address: The IPv4 address assigned to a CPE. It is used in the ISP network between the CPE and the SLNAT44.

External Address: The IPv4 address used on the Internet and routed to the SLNAT44.

Mapping rule: A set of parameters configured on the SLNAT44 (not on the CPE) describing the relationship between internal and external addresses.

3. Address Formats

Internal addresses have the format illustrated in Figure 3. The addresses are simply made of three parts concatenated together: the Internal Prefix, the External Suffix, and the Port Set ID.

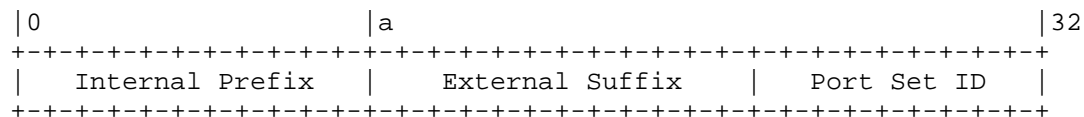


Figure 3: Internal Address format

External Addresses have the format illustrated in Figure 4. It is made of two parts: the External Prefix and the External Suffix.

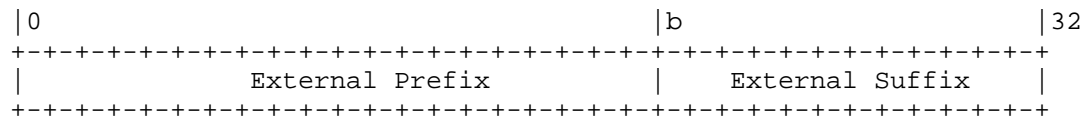


Figure 4: External Address format

The lengths of the Internal and External Prefixes, "a" and "b", are mandatory parameters of SLNAT44. They are determined by the ISP. They need not be communicated to the CPE. Other lengths can be

computed from them as follows:

- o Length of External Suffix: $32 - b$
- o Length of Port Set ID: $b - a$

4. Customer Provisioning

Customer Provisioning is applied to the CPE when the CPE serves a gateway.

As part of its start up routine, the CPE is assigned an IPv4 address by the ISP using regular means (DHCP, PPP, etc.). This is the Internal Address.

In addition, using new provisioning options, the CPE is assigned a Port Set ID.

Optionally, a Port Set Mask is also provisioned to the CPE. This mask is of the same length as the Port Set ID (i.e., $b-a$ bits). Its purpose is to allow discontinuous port ranges. If no mask is provided, a mask of all ones is assumed by default, which implies a continuous port range. System ports (0-1023) should not to be assigned to any CPE.

In summary, the CPE is provisioned with the following elements:

- o IPv4 address (as usual)
- o Port Set ID
- o Port Set Mask (optional)

When the CPE is configured in the bridge mode, all the above features are provisioned directly to the end host behind the CPE.

Note: no matter in which mode the CPE is running, the customer provisioning could be either dynamic or static. Static provisioning implies an address planning for the private IPv4 address (i.e., RFC1918 addresses) inside in the domain. Static provisioning enables servers (passive daemons) at the home network being accessible within the domain. CPE running as bridge makes this feature easy to deploy while running as L3 gateway requires port redirection if an in-domain server at a host is demanded.

5. SLNAT44 Configuration

The SLNAT44 is configured with a set of mapping rules. Each rule contains:

- o Internal Prefix
- o External Prefix
- o Port Set Mask (optional)

Prefixes include their length. For simplicity, rule prefixes MUST NOT overlap with other rules.

If it is absent, the Port Set Mask is assumed to be all ones by default.

6. Port Set Computation

Given a Port Set ID and a Port Set Mask, both n bits in length, the set of allowed ports is defined as the set of port numbers for which the higher-order n bits of their binary expression whose corresponding mask bits are 1 are equal to corresponding bits from the Port Set ID.

```

|0           |5
+---+---+---+
|1 1 1 0 1|   Port Set ID = 29 (length n = 5 bits)
+---+---+---+
& & & & &
+---+---+---+
|1 1 1 1 1|   Port Set Mask
+---+---+---+
| | | | |
V V V V V
+---+---+---+---+---+---+---+---+---+---+---+---+
|1 1 1 0 1 x x x x x x x x x x x x|   Port Set = 59392-61439
+---+---+---+---+---+---+---+---+---+---+---+---+
|0                                     |16

```

Figure 5: Example Contiguous Port Set Computation

```

|0                                     |8
+---+---+---+---+---+---+
|0 0 1 0 1 1 1 1| Port Set ID = 29 (length n = 8 bits)
+---+---+---+---+---+---+
& & & & & & &
+---+---+---+---+---+---+
|0 0 1 1 1 1 1 1| Port Set Mask
+---+---+---+---+---+---+
| | | | | | | |
V V V V V V V V
+---+---+---+---+---+---+
|x x 1 0 1 1 1 1 x x x x x x x x| Port Set = 12032-12287, 28416-28671,
+---+---+---+---+---+---+
|0                                     |16

```

Figure 6: Example Non-Contiguous Port Set Computation

It follows that the number of ports in the set is $2^{(16-x)}$, where x is the number of ones in the Port Set Mask.

This computation is performed by the CPE as part of its provisioning routine as well as by the SLNAT44 for dropping packets with ports outside the allowed range.

For the purposes of SLNAT44, a "source port" corresponds to either a TCP source port, a UDP source port, or an ICMPv4 identifier, while a "destination port" corresponds to either a TCP destination port, a UDP destination port, or an ICMPv4 identifier. Note that an ICMPv4 identifier plays the role of both source and destination port.

Transport protocols other than TCP and UDP, as well as ICMPv4 types without an identifier field, are not supported.

7. CPE Operation

CPE can be configured as either a gateway or transparent bridge.

In the gateway mode, packets sent from the CPE MUST have the provisioned IPv4 address as source and MUST have a source port that is within the allowed set. This is usually accomplished by having the CPE run a NAT44 configured with the provisioned address and allowed port set and having it process all packets sent out the WAN interface.

Packets received by the CPE on its WAN interface with a destination port outside the allowed range MUST be dropped.

In the bridge mode, however, CPE only transfers packets and therefore the service of stateless NAT44 is performed by the SLNAT44 directly towards end hosts that possibly running as in-domain servers.

Regardless of any mode of the CPE, the operation involves injecting private addresses (or prefixes) into the intra-domain backbone routing infrastructure. It is necessary to operationally ensure that there are no private addresses/prefixes are leaking into the backbone route tables unless they are assigned by the SLNAT44 to CPEs or directly to hosts.

7.1. ALG Handling

If the CPE implements application level gateways (ALGs) such as FTP, RSTP or PPTP, it must ensure that ports present in the payload when translated fall within the allowed range.

8. SLNAT44 Operation

8.1. Internal to External

When it receives a packet on an internal interface, the SLNAT44 finds the rule whose Internal Prefix matches the packet's source address. It extracts the Port Set ID from the packet's source address. It then checks if the packet's source port is within the allowed set, using the rule's Port Set Mask. If it is not, the packet MUST be dropped.

If the packet's source port is within the allowed set, the SLNAT44 builds the External Address by concatenating the rule's External Prefix with the External Suffix extracted from the packet's source address. It then replaces the packet's source address with this External Address. The IPv4 and transport-layer checksums are updated as necessary. The packet is then forwarded as usual.

8.2. External to Internal

When it receives a packet on an external interface, the SLNAT44 finds the rule whose External Prefix matches the packet's destination address. It then builds the Internal Address by concatenating the rule's Internal Prefix, the External Suffix extracted from the packet's destination address, and the Port Set ID computed by applying the rule's Port Set Mask to the packet's destination port's higher-order bits. It then replaces the packet's destination address with this Internal Address. The IPv4 and transport-layer checksums are updated as necessary. The packet is then forwarded as usual.

8.3. Fragment Handling

If the incoming IP packet contains a fragment, then more processing may be needed. This specification leaves open the exact details of how a SLNAT44 handles incoming IP packets containing fragments, and simply requires that the external behavior of the SLNAT44 be compliant with the following conditions.

The SLNAT44 **MUST** handle fragments. In particular, SLNAT44 **MUST** handle fragments arriving out of order, conditional on the following:

- o The SLNAT44 **MUST** limit the amount of resources devoted to the storage of fragmented packets in order to protect from DoS attacks.
- o As long as the SLNAT44 has available resources, the SLNAT44 **MUST** allow the fragments to arrive over a time interval. The time interval **SHOULD** be configurable and the default value **MUST** be of at least 2 seconds.
- o The SLNAT44 **MAY** require that the UDP, TCP, or ICMPv4 header be completely contained within the fragment that contains fragment offset equal to zero.

For incoming packets carrying TCP or UDP fragments with a non-zero checksum, SLNAT44 **MAY** elect to queue the fragments as they arrive and translate all fragments at the same time. In this case, the incoming tuple is determined as documented above to the un-fragmented packets. Alternatively, a SLNAT44 **MAY** translate the fragments as they arrive, by storing information that allows it to compute the necessary port number for fragments other than the first. In the latter case, subsequent fragments may arrive before the first, and the rules (in the bulleted list above) about how the SLNAT44 handles (out-of-order) fragments apply.

Implementers of SLNAT44 should be aware that there are a number of well-known attacks against IP fragmentation; see [RFC1858] and [RFC3128]. Implementers should also be aware of additional issues with reassembling packets at high rates, described in [RFC4963].

9. Address Mapping Example

An operator has two public ranges of size /18 and /19 called foo and bar respectively. It plans to use 10/8 as its internal address prefix and PSID (port range) of length 5. Two prefixes of the internal network

The internal prefixes lengths are:

- o 32 - 18 - 5 = 13 (derived from foo)
- o 32 - 19 - 5 = 14 (derived from bar)

This will give the following possible mappings:

- o foo/18 <--> 10.0.0.0/13
- o bar/19 <--> 10.128.0.0/14

Author note: Discuss the where internal prefixes are overlapping

10. Security Considerations

The security considerations related to IP address sharing documented in RFC 6269 [RFC6269] and RFC 6056 [RFC6056] apply to SLNAT44.

11. Acknowledgements

Section 8.3 is adapted from [RFC6146].

12. References

12.1. Normative References

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

12.2. Informative References

- [RFC1858] Ziemba, G., Reed, D., and P. Traina, "Security Considerations for IP Fragment Filtering", RFC 1858, October 1995.
- [RFC3128] Miller, I., "Protection Against a Variant of the Tiny Fragment Attack (RFC 1858)", RFC 3128, June 2001.
- [RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, July 2007.
- [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.
- [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.

Authors' Addresses

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Will Liu
Huawei Technologies
Bantian, Longgang District
Shenzhen 518129
P.R. China

Email: liushucheng@huawei.com

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Email: simon.perreault@viagenie.ca

Reinaldo Penno
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, California 95134
USA

Email: repenno@cisco.com

Maoke Chen
FreeBit
3-6 Maruyama-cho
Shibuya-ku, Tokyo 150-0044
Japan

Email: fibrib@gmail.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 10, 2013

C. Zhou
Huawei Technologies
T. Tsou
Huawei Technologies (USA)
C. Grundemann
Cablelabs
July 16, 2012

Scenarios of IPv4 sunsetting
draft-zhou-sunset4-scenarios-01

Abstract

This document describes scenarios at subscriber, carrier and enterprise sites during IPv4 sunsetting. In each site, there may be different requirements and issues. The aim of this document is to put forward some issues in these scenarios and to identify whether further specifications are needed to solve these issues to facilitate IPv4 sunsetting.

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 10, 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. Conventions	3
3. Subscriber Site Scenario	3
4. Carrier Site Scenario	3
4.1. Traceback	3
4.2. Stateless CGN	4
4.3. High Availability	4
4.4. ALG	4
5. Enterprise Site Scenario	5
6. IANA Considerations	5
7. Security Considerations	5
8. Normative References	7

1. Introduction

There are already a set of documents in IETF which to some extent facilitate IPv6 transition. For example, [I-D.ietf-behave-lsn-requirements] describes the common requirements of CGN (NAT44). For devices which implement NAT, MIB module is introduced in [I-D.ietf-behave-nat-mib]. However, there are many scenarios and issues encountered at subscriber, carrier and enterprise sites, e.g., source trace, high availability, and ALG issues at carrier site scenario. In this document, these scenarios will be proposed in detail and some issues in these scenarios will be discussed.

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.

3. Subscriber Site Scenario

Some subscribers have the need to run some servers at home, for example, web server, webcam, FTP server, etc. Sometimes when a subscriber equipment reboots it may be assigned a new IP address which is different from the previous one. To accomodate this IP address change, DDNS is used. If NAT is used in subscribe premise, static port-forwarding can be configured for a specific service so that DDNS can continue to work. But if CGN is deployed in the operator's network, one CGN will serve a lot of users, static port-forwarding configuration will require a lot of operational work, and there will be IP address and port conflict if multiple subscribers require a same public IP address and / or port. A traditional solution is to assign public IP address to subscribers who needs to run a server at home, but this will also require extra operational work, make the network more complicated. In such a case, a possible solution is that DDNS system works together with some dynamic NAT traversal technologies, e.g. UPnP/PCP, or the CGN provide DDNS proxy.

4. Carrier Site Scenario

For carrier site case, we provide some scenarios and issues as below for the working group discussion.

4.1. Traceback

Before CGN is introduced, the servers use the source IPv4 address as an identifier to treat incoming packets differently. When the

address sharing scheme is proposed, the server could not identify which host sends the packet because the packets are from the same source address. [I-D.boucadair-intarea-nat-reveal-analysis] proposed solutions to identify each host sharing the same IP address with a unique host identifier. But there are at least two issues existing in the traceback solutions: logging architecture and port allocation algorithm.

As described in section 4 of [I-D.ietf-behave-lsn-requirements], the destination addresses or ports should not be logged in CGN in order to reduce the logs in CGN. [RFC6302] provides recommendations for Internet-facing servers logging incoming connections. But it does not provide any recommendations about logging on carrier-grade NAT. So, a logging architecture in CGN to maintain records of the relation between a customer's identity and IP/port resources is needed.

[RFC6431] provides port set options for port range allocation: contiguous, non-contiguous and random. In the random-based solution, the algorithm should be reversible in order to trace the host. But this may bring some security problems.

4.2. Stateless CGN

Carrier-grade NAT44 is one of the solutions to deal with the IPv4 address shortage problem. But the current NAT44 CGN (Carrier-grade NAT) is stateful and TCP/UDP session based, which makes the CGN complex. There have been a number of efforts at IETF moving the NAT function from a stateful carrier grade NAT to the CPEs by allocating port sets to each customer, e.g., MAP/4RD-U, LAFT6, and etc. There is also a requirement for NAT44 CGN to become completely stateless.

4.3. High Availability

In most ISP networks, one CGN device may serve large number of customers. For stateful NAT, if there is a single point of failure in the CGN, the service may be interrupted or degraded. Therefore, redundancy capabilities (including hot and cold standby) of the CGN devices are strongly needed to deliver highly available services to customers. [I-D.xu-behave-stateful-nat-standby] may be a possible way to solve this problem. In addition, pre-configuring a pool of public IPv4 addresses to the CGN device when it is in failure may also be a candidate solution.

4.4. ALG

Carrier-grade NAT44 performs NAT-44 and inherits the limitations of NAT. Some protocols require ALGs in the CGN to traverse through the NAT, e.g., FTP, RTP. However, in most ISP's network, CGN is a shared

network device which needs to support a large number of sessions. It is a huge work load for CGN to implement every ALGs, which will obviously bring bad performance for CGN. How to make CGN more efficiency under the pressure of ALG becomes an issue. One possible solution is to let the CPE or host implement ALG instead of CGN, or a flexible way to make ALG at either CPE or CGN is needed.

5. Enterprise Site Scenario

NAT is a basic feature of enterprise network. The firewall/NAT device is deployed at the entrance of the enterprise network, following by the web server and the terminal. Part of the web servers are required to open publically to provide one domain name and corresponding IP address (Two ways: the enterprise has its own DNS server; the enterprise has no DNS server and needs to publicize one public address). NAT device is required to support this specific case. In addition, the terminal or the web server following NAT device need to access Internet. There are requirements for the enterprise users to record the NAT translation information.

Some basic requirements of NAT device are also valid in enterprise scenarios, e.g., NAT traceback, port range allocation and NAT standby. NAT device needs to record the NAT translation log in traceback solutions. NAT server is required to support port range allocation. Two NAT devices should store the information of each other to guarantee normal operation when one device is in failure in enterprise scenarios.

6. IANA Considerations

No request to IANA.

7. Security Considerations

TBD

Authors' Addresses

Cathy Zhou
Huawei Technologies
Bantian, Longgang District
Shenzhen 518129
P.R. China

Phone:
EMail: cathy.zhou@huawei.com

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
EMail: tina.tsou.zouting@huawei.com

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

Phone: 303.661.3779
Email: c.grundemann@cablelabs.com

8. Normative References

- [I-D.ietf-behave-lsn-requirements] Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common requirements for Carrier Grade NATs (CGNs)", May 2012.
- [I-D.ietf-behave-nat-mib] Perreault, S., Tsou, I., and S. Sivakumar, "Additional Definitions of Managed Objects for Network Address Translators (NAT)", April 2012.
- [I-D.boucadair-intarea-nat-reveal-analysis] Boucadair, M., Touch, J., Levis, P., and R. Penno, "Analysis of Solution Candidates to Reveal a Host Identifier in Shared Address Deployments", September 2011.
- [I-D.xu-behave-stateful-nat-standby] Xu, X., Boucadair, M., Lee, Y., and G. Chen, "Redundancy Requirements and Framework for Stateful Network Address Translators (NAT)", October 2010.