Abstract

This memo defines a simple stateless and deterministic mode of operating a carrier-grade NAT in both a NAT444 and DS-Lite environment.

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 27, 2012.

Copyright Notice

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

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

1. Introduction ................................................. 3
2. SD-CPE operation ............................................. 4
   2.1. SD-CPE NAT function .................................. 4
   2.2. SD-CPE port translation function ..................... 4
3. SD-CGN and SD-AFTR operation ............................... 5
   3.1. Anycast IPv6 address for SD-AFTR ..................... 5
   3.2. SD-CGN and SD-AFTR operation ......................... 5
   3.3. Mapping function ................................. 5
   3.4. Example mapping function .................. 6
   3.5. Port exceeded message .......................... 7
   3.6. Address compression ratio ....................... 7
   3.7. Anycast IPv4 pool .............................. 7
   3.8. Stateless operation .............................. 7
   3.9. SD-AFTR stateless domain ....................... 7
   3.10. Premium customers .............................. 8
4. SD-NAT 444 example ........................................... 8
   4.1. SD-NAT444 architecture ......................... 9
   4.2. SD-CPE3 NAT bindings in NAT444 mode ............... 10
   4.3. SD-CPE7 NAT bindings in NAT444 mode ............... 10
   4.4. SD-CGN NAT bindings in NAT444 mode ............... 10
5. SD-DS-Lite example ........................................... 10
   5.1. SD-DS-Lite architecture ....................... 11
   5.2. SD-CPE3 NAT bindings in DS-Lite mode ............. 12
   5.3. SD-CPE7 NAT bindings in DS-Lite mode ............. 12
   5.4. SD-AFTR NAT bindings in DS-Lite mode ............. 12
6. Comparison with other techniques .................... 12
   6.1. Comparison with DS-Lite and NAT444 CGN .......... 13
   6.2. Comparison with 4rd/4via6/A+P/DIVI ............... 13
7. IANA Considerations ........................................ 13
8. Security Considerations ..................................... 13
9. References ............................................... 14
   9.1. Normative references .......................... 14
   9.2. Informative references ...................... 14
Authors’ Addresses ........................................... 14
1. Introduction

NAT444 and DS-Lite [RFC6333], are two solutions to deal with the IPv4 exhaustion problem. NAT444 solves it by introducing a second layer of NAT in the ISP network. DS-Lite solves it by decoupling the deployment of IPv6 in the access network from the deployment of IPv6 in the applications. DS-Lite is based on a combination of IPv4 over IPv6 encapsulation and a carrier grade NAT (CGN).

There have been a number of efforts at IETF to evolve the DS-Lite model by moving the NAT function from a centralized, stateful carrier grade NAT to the CPEs by allocating a fixed number of ports to each customer. The provider equipment would then only do the decapsulation of the IPv4 traffic, providing a stateless operation model, where a number of those relay devices could be operated independently in the ISP network. One drawback of such design, is that the service provider looses some flexibility on how ports are allocated. In particular, The IPv6 and global IPv4 address spaces are mapped to each over, making it difficult for an ISP to add or remove addresses from the NAT pool. Another drawback is that the number of IPv4 port allocated per subscriber is encoded in the IPv6 address. This results in a renumbering event when the ISP want to change the IPv4 oversubscription ratio.

Another direction has been to define a deterministic operation model for NAT444 CGNs, where ports are statically distributed to users on that CGN. A deterministic mapping function is defined to convert an internal address to an external address and port range. This mapping function is reversible, eliminating the need to keep logs for abuse/LEA purpose. However, such a deterministic NAT still need to maintain per connection state and as such is not stateless.

This proposal enhance the deterministic NAT model. By offering a fully programmatic mapping of the complete NAT binding, it enables a carrier-grade NAT/AFTR to become completely stateless and deterministic. This model of operation applies similarly to DS-Lite and NAT444 environments.

By leveraging this stateless and deterministic mode of operation, an ISP can deploy any number of SD-CGNs/SD-AFTRs to provide redundancy and scalability at low cost. Because the NAT mappings on the CGNs or AFTRs are fully stateless and deterministic, routers can implement the functionality in hardware and perform it at very high speed with very low overhead in terms of packet delay.

The SD-NAT CPE remains a simple NAT router. Its WAN interface can be configured either with IPv6 and a B4 element to do IPv4 over IPv6 encapsulation or with a private unique IPv4 address. The only change
required is on the algorithm used to select an outgoing port in the NAT function.

Another important aspect of SD-NAT is that CPEs do not need to learn any special configuration. This allows the ISP to modify at will its NAT environment, adding or removing resources from the NAT pool or changing the number of ports per user without having to reconfigure or renumber the CPEs.

2. SD-CPE operation

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].

An SD-CPE operates like a regular CPE with some modification on the way the NAT function is performed.

2.1. SD-CPE NAT function

The SD-CPE performs IPv4 NAPT from the internal RFC1918 addresses to the IPv4 address configured on the WAN interface. In a DS-Lite environment, the B4 address 192.0.0.2 is used in lieu of the IPv4 WAN interface.

2.2. SD-CPE port translation function

Internal IPv4 source ports are translated into the range [1024..65635]. The outgoing ports MUST be allocated sequentially by the SD-CPE.

When the SD-CPE needs to allocate a new external port, it MUST start from 1024 and increment until it finds an un-allocated port.

The first key design element here is that the SD-CGN or SD-AFTR can make the assumption that all datagrams coming from the SD-CPE will have source ports formatted in a well-known range. This will make the stateless deterministic NAT function on the SD-CGN or SD-AFTR easy to implement.

The second key design element is that the SD-CPE does not need to know how many ports are reserved for its usage on the SD-CGN or SD-AFTR. If the SD-CPE allocates a port number that is too high and does not fit into the range allocated on the SD-CGN or SD-AFTR, the outgoing packet will be dropped by the SD-CGN or SD-AFTR and an ICMP [RFC0792] message type 13 (Communication administratively prohibited) MAY be returned to the SD-CPE. That design element allows for a
great flexibility on the ISP side, the maximum number of ports per subscriber can be changed at any time without impacting the subscriber.

A third key design element is that the SD-CPE does not need to know which external IPv4 address it will be mapped to. That design element allows for a great flexibility on the ISP side, the pool of global IPv4 addresses on the SD-CGNs or SD-AFTRs can be changed at any time with little impact on the subscriber. The subscriber configuration will not have to change, the SD-CPE will not have to renumber or reboot, however the current connections may or may not survive depending on the pool update mechanism implemented on the SD-CGNs / SD-AFTRs.

Note: if the SD-CPE is running its own applications sourcing datagrams on its WAN interface (or B4 element 192.0.0.2), it must select a source port using the same algorithm as the NAPT function.

3. SD-CGN and SD-AFTR operation

3.1. Anycast IPv6 address for SD-AFTR

All SD-AFTRs associated to a domain (or group of users) will be configured with the same IPv6 address on the interface facing IPv6 customers. A route for that IPv6 address will be anycasted within the access network. Anycast IPv4 address for SD-CGN Similarly, in the NAT444 environment, the IPv4 traffic will be anycasted to a number of SD-CGNs.

3.2. SD-CGN and SD-AFTR operation

All SD-CGNs or SD-AFTRs associated to a domain (or group of users) will be configured with the same pool of global IPv4 addresses. They will be configured also with the same deterministic address and port mapping function. This function will map the customer ID, derived from either its private IPv4 address in the NAT444 case or its IPv6 address in the DS-Lite case and the source port used by the incoming connection into a global IPv4 address and port.

3.3. Mapping function

Various mapping functions can be defined. This is an implementation issue on the SD-CGNs or SD-AFTRs, however that function MUST be the same on all the SD-CGNs or SD-AFTRs within a domain.

The input parameters of that mapping function are the number of addresses in the IPv4 global pool, the number of customers and the
maximum number of ports per customers.

Note: external ports in the range of [0..1023] MAY be excluded in the mapping function, however this is not a requirement.

3.4. Example mapping function

Lets say that all external ports above 1023 are used and a 1000 ports are allocated per subscriber, and only one IPv4 address is configured in the NAT pool. Customer #1 will have external ports 1024 to 2023, customer #2 will have external ports 2024 to 3023, etc...

Here is an algorithm that produces this result:

Input variables:

- \( i \): subscriber index
- \( \text{maxPort} \): maximum number of ports per subscriber
- \( \text{baseCPE} \): CPE address base, determined by Service Provider, e.g. 172.16.0.1
- \( \text{baseCGN} \): CGN address base, determined by Service Provider e.g. 1.2.3.1.

Each SD-CPE implements a stateful NAT44 function with the following mappings:

- \( \text{host_src_ip} \) is mapped to: CPE WAN IPv4.
- \( \text{host_src_port} \) is mapped to: port, Where port is the first available port starting at 1024.

Each SD-CGN/SD-AFTR implements a deterministic stateless NAT44 function with following mappings:

\[
\begin{align*}
\text{cpe_src_ip} & \text{ is mapped to: } \text{baseCGN} + \text{floor} \left( \frac{i}{P} \right) \\
\text{cpe_src_port} & \text{ is mapped to: } \text{cpe_src_port} + \text{maxPort} \times \left( i \mod P \right)
\end{align*}
\]

Where \( i = \text{cpe_src_ip} - \text{baseCPE} \) and \( P = \text{floor} \left( \frac{(65536-1024)}{\text{maxPort}} \right) \).
3.5. Port exceeded message

If the SD-CGN or SD-AFTR receives an incoming datagram with a source port number higher than 1023 + the maximum number of allocated ports per customer, the datagram MUST be dropped and an ICMP [RFC0792] error message type 13 (Communication administratively prohibited) MAY be returned to the originating SD-CPE.

3.6. Address compression ratio

The maximum number of subscriber per IPv4 address is equal to (65536 minus 1024) divided by the number of ports per user, as configured on the SD-CGN or SD-AFTR.

3.7. Anycast IPv4 pool

Routes to the pool of global IPv4 addresses configured on the SD-CGN or SD-AFTR will be anycasted by all relevant SD-CGN or SD-AFTRs within the ISP routing domain.

3.8. Stateless operation

That way, return IPv4 traffic can go back to any SD-CGN or SD-AFTR associated with that pool of address. Because the mapping function is deterministic (shared on all AFTRs) and because there is no dynamic state associated with any particular NAT mappings on any SD-CGN or SD-AFTR, the returning IPv4 traffic can be translated back, re-encapsulated in IPv6 in the case of DS-Lite, and send back to the customer by any SD-CGN or SD-AFTRs within the domain.

No need to do port mapping garbage collection on SD-CGN/SD-AFTR

Because those bindings are completely stateless and deterministic, an SD-CGN or SD-AFTR does not need to maintain state in its port-mapping table, and as such, does not need to perform any garbage collection on idle connections. From a subscriber application perspective, that mean that the application only need to negotiate ports with the local SD-CPE, for example using the PCP protocol.

3.9. SD-AFTR stateless domain

Using the DHCPv6 DS-Lite tunnel-end-point option, groups of subscribers and can be associated to a different SD-AFTR domain. That can allow for differentiated level of services, e.g. number of ports per customer device, QoS, bandwidth, value added services,...
3.10. Premium customers

Premium customers can be allocated a full global IPv4 address, using the exact same mechanism as described above. The maximum number of ports per user is then simply set to 65536 minus 1024.

An alternative architecture is to allocate a global IPv4 address by DHCPv4, either natively over IPv4 or over the IPv6 softwire, to the SD-CPE. That SD-CPE then acts as a regular home NAT. In the DS-Lite case, it sends the translated IPv4 packets, encapsulated over IPv6, to the SD-AFTR.

4. SD-NAT 444 example

This examples assume two SD-CGNs or two SD-AFTRs configured with the same IPv4 global address in their NAT pool. The operator has configured a limit of 1000 ports per subscriber. Subscriber #3 has 2 hosts, each running one connection and subscriber #7 has one host running one connection. Subscriber #3 SD-CPE WAN IPv4 address is 172.16.0.3 in case of NAT444 or WAN IPv6 address 2001:db8::3 in case of DS-LITE. Subscriber #7 WAN IPv4 address is 172.16.0.7 in case of NAT444 or WANT IPv6 address 2001:db8::7 in case of DS-LITE. Host 1 and 2 in subscriber #3 network are using IPv4 address 192.168.1.2 and 192.168.1.3, and host 1 in subscriber #7 network is using IPv4 address 192.168.1.2.
4.1. SD-NAT444 architecture

Figure 1: SD-NAT NAT444 CGN example
4.2. SD-CPE3 NAT bindings in NAT444 mode

| Connection subscriber3-1:                      |
| SRC IPv4 192.168.1.2 mapped to SRC IPv4 172.16.0.3 |
| SRC port 3786 mapped to SRC port 1024            |

| Connection subscriber3-2:                       |
| SRC IPv4 192.168.1.3 mapped to SRC IPv4 172.16.0.3 |
| SRC port 60345 mapped to SRC port 1025           |

Figure 2: SD-CPE3-NAT444

4.3. SD-CPE7 NAT bindings in NAT444 mode

| Connection subscriber7-1:                      |
| SRC IPv4 192.168.1.2 mapped to SRC IPv4 172.16.0.7 |
| SRC port 3786 mapped to SRC port 1024            |

Figure 3: SD-CPE7-NAT444

4.4. SD-CGN NAT bindings in NAT444 mode

| Connection subscriber3-1:                      |
| SRC IPv4 172.16.0.3 mapped to SRC IPv4 1.2.3.4  |
| SRC port 1024 mapped to SRC port 3024           |

| Connection subscriber3-2                        |
| SRC IPv4 172.16.0.3 mapped to SRC IPv4 1.2.3.4  |
| SRC port 1025 mapped to SRC port 1025           |

| Connection subscriber7-1:                      |
| SRC IPv4 172.16.0.7 mapped to SRC IPv4 1.2.3.4  |
| SRC port 1024 mapped to SRC port 7024           |

Figure 4: SD-CGN-NAT444

5. SD-DS-Lite example

This example assumes two SD-AFTRs configured with the same IPv4 global address in their NAT pool. The operator has configured a limit of 1000 ports per subscriber. Subscriber #3 has 2 hosts, each
running one connection and subscriber #7 has one host running one connection. Subscriber #3 SD-CPE WAN IPv6 address is 2001:db8::3. Subscriber #7 WAN IPv6 address is 2001:db8::7. Both SD-CPE use the same B4 address 192.0.0.2, as specified in DS-Lite. Host 1 and 2 in subscriber #3 network are using IPv4 address 192.168.1.2 and 192.168.1.3, and host 1 in subscriber #7 network is using IPv4 address 192.168.1.2.

5.1. SD-DS-Lite architecture

```
+-------+     +-------+     +-------+
|Host 1 |     |Host 2 |     |Host 1 |
+-------+     +-------+     +-------+
192.168.1.2\       /192.168.1.3    |192.168.1.2
\ /                |
\ /                 |
\ /                 |
+-------+     +-------+     +-------+
|SD-CPE #3|     |SD-CPE #7|     |
+-------+     +-------+     +-------+
2001:db8::3|    //2001:db8::7
192.0.0.2|    //192.0.0.2
|
-------------
\ISP core network /
------------- IPv6-------------
|| Anycast     \     Anycast   |
|  AFTR       \     AFTR       |
|IPv6 route   \     IPv6 route |
+-------------+     +-------------+
|SD-AFTR1  |     |SD-AFTR2  |
+-------------+     +-------------+
1.2.3.4\       /1.2.3.4
Anycast IPv4 route\ / Anycast IPv4 route
for 1.2.3.4 \ / for 1.2.3.4
-------------R-R-------------
\                     /
|  Internet          |
\                     |
------------- IPv4-------------
```

Figure 5: SD-NAT DS-Lite AFTR example
5.2. SD-CPE3 NAT bindings in DS-Lite mode

Connection subscriber3-1:
SRC IPv4 192.168.1.2 mapped to SRC IPv4 192.0.0.2
SRC port 3786 mapped to SRC port 1024

Connection subscriber3-2:
SRC IPv4 192.168.1.3 mapped to SRC IPv4 192.0.0.2
SRC port 60345 mapped to SRC port 1025

Figure 6: SD-CPE3-DS-Lite

5.3. SD-CPE7 NAT bindings in DS-Lite mode

Connection subscriber7-1:
SRC IPv4 192.168.1.2 mapped to SRC IPv4 192.0.0.2
SRC port 3786 mapped to SRC port 1024

Figure 7: SD-CPE7-DS-Lite

5.4. SD-AFTR NAT bindings in DS-Lite mode

Connection subscriber3-1 (2001:db8::3):
SRC IPv4 192.0.0.2 mapped to SRC IPv4 1.2.3.4
SRC port 1024 mapped to SRC port 3024

Connection subscriber3-2 (2001:db8::3):
SRC IPv4 192.0.0.2 mapped to SRC IPv4 1.2.3.4
SRC port 1025 mapped to SRC port 1025

Connection subscriber7-1 (2001:db8::7):
SRC IPv4 192.0.0.2 mapped to SRC IPv4 1.2.3.4
SRC port 1024 mapped to SRC port 7024

Figure 8: SD-AFTR-DS-Lite

6. Comparison with other techniques
6.1. Comparison with DS-Lite and NAT444 CGN

The trade-off of this stateless deterministic model of operation is efficiency in IPv4 address sharing. A typical DS-Lite or NAT444 environment can share an IPv4 address among over 1000 customers, if the ratio of maximum number of port used by any given subscriber to the maximum average number of ports used by all subscribers is high. A typical SD-DS-Lite or SD-NAT444 environment will share the same address among only about 100 subscribers, using a maximum number of ports per subscribers of 600.

6.2. Comparison with 4rd/4via6/A+P/DIVI

The main difference with 4rd/A+P/4via6/DIVI techniques is that the CPE is unaware of the exact port distribution algorithm. That way, the ISP maintains the flexibility to change the numbers of ports allocated per subscriber, or add/delete/modify the pool of available global IPv4 addresses at any time without creating a customer outage.

SD-DS-Lite is based on encapsulation, 4via6 and DIVI are based on double translation IPv4 to IPv6 and back to IPv4. Service providers have deployed encapsulation techniques for many years, double translation is new and there is less operational experience with it.

An SD-CPE sends packets through a set of identified devices, the SD-CGs or SD-AFTRs. This enables simple enforcement points to do IPv4 subscriber management (e.g. counting IPv4 traffic). A consequence of that architecture in the SD-DS-Lite model is that is does not allow for 4rd style shortcuts, where traffic between customers within the same domain does not go through the relay.

7. IANA Considerations

None.

8. Security Considerations

As described in [RFC6269], with any fixed size address sharing techniques, port randomization is achieve with a smaller entropy.

Recommendations listed in [RFC6302] applies.

9. References
9.1. Normative references


9.2. Informative references


Authors’ Addresses

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

Email: adurand@juniper.net

Reinaldo Penno
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089-1206
USA

Email: rpenno@juniper.net