Internet Engineering Task Force	R.P. Penno
Internet-Draft	A.D. Durand
Intended status: Standards Track	Juniper Networks
Expires: May 04, 2012	L.H. Hoffmann
	Bouygues Telecom
	November 01, 2011

Stateless Deterministic NAT
draft-penno-softwire-sdnat-01

Abstract

This memo define a simple stateless and deterministic mode of operating a carrier-grade NAT in both a NAT44 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 May 04, 2012.

Copyright Notice

Table of Contents

1. Introduction

NAT44 and DS-Lite [RFC6333],are two solutions to deal with the IPv4 exhaustion problem. NAT44 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 NAT44 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 enhances 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 NAT44 environments.

The SD-NAT architecture has the following characteristics:

*Zero log: Because the SD-CGN/SD-AFTR NAT bindings are deterministic and reversible, the ISP does not need to keep any NAT binding logs.

*Zero state on SD-CGN/SD-AFTR: There is no operational per-session state on the SD-CGNs/SD-AFTRs. 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.

*Low barrier of adoption: SD-CPEs can remain IPv4-only devices. Only minor code change or configuration is required to modify the port mapping algorithms, the rest of the CPE implementation remains the same. The ISP access network can also remain IPv4-only, enabling the ISP to deploy IPv6 independently to this IPv4-exhaustion solution.

*Flexibility of operation: The ISP can add or remove addresses from the NAT pool without having to renumber the access network.

*Compatible with IPv6: This SD-NAT mechanism can be deployed similarly by IPv4-only ISPs, Dual-Stack ISPs or IPv6-only ISPs. In the IPv4 case, the SD-CPE WAN Ipv4 address is used as customer index, in the IPv6 case, IPv4 packets are tunneled over IPv6 just like in the DS-Lite case and the SD-CPE WAN IPv6 address is used as customer index.

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.
Internal IPv4 source ports are translated into the range [1024..maxport]. Different SD-CPEs can implement different methods to determine the value of maxport, some of which are defined below.

A 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. Another 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: the SD-CPE can learn the external address it is mapped to using the PCP protocol defined in [I-D.ietf-pcp-base].

2.2. SD-CPE maxport configuration

The simplest way to build an SD-CPE is to configure it with a maxport value. This could be done, for example, using either manual configuration, a TR69 parameter to be defined or a DHCP option, to be defined.
A quick look at the open source OpenWRT project show that the functionality to restrict the source port space is already present, so no new code is necessary.

2.3. SD-CPE with incremental port allocation

In this model, an SD-CPE would dynamically learn the value of maxport instead of being configured with it. This removes a configuration step of the CPE. The trade-off is that some new code implementing the following algorithm must be implemented on the SD-CPE.
When the incremental SD-CPE needs to allocate a new external port, it MUST start from 1024 and increment until it finds an un-allocated port. 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) will be returned to the SD-CPE.
The trade-off to the simplicity of this algorithm are a) the restart at 1024 may use extra cycles and b) port randomization is impacted. See alternative scramble techniques on SD-CGN/SD-AFTR to mitigate this issue.

2.4. SD-CPE with dynamic maxport determination

In this model the SD-CPE will try to determine the value of maxport. The easiest way of doing this is to send dummy packets at boot time starting with source port 1024 and incrementing the source port number until the ICMP message type 13 is received. This is, however, quite

inefficient. A better way is to use a dichotomy algorithm. Start with a high port of 65535 and a low port of 1024 and then divide the space in two depending on the reception (or non-reception of the ICMP message type 13. The advantage over incremental port allocation is that source port randomization is preserved.
Note 1: this dichotomy algorithm can be apply to real traffic, there is no need to send dummy packets. If the ICMP message type 13 is received, then the packet is retransmitted with a lower source port number.
Note 2: if an ICMP message type 13 does not come back within (TBD) seconds, the packets is considered successfully delivered through the SD-CGN/SD-NAT.

3. Host based SD-NAT, Wireless application

If the SD-CPE is running it's own applications sourcing datagrams on its WAN interface (or B4 element 192.0.0.2), or if a host is directly connected to the ISP access network, it must select a source port using the same algorithm as the NAPT function.
In particular, this makes SD-NAT well suited for wireless environments where the UE can be easily programmed to restrict its outgoing source port space.

4. SD-CGN and SD-AFTR operation

4.1. 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.

4.2. Multiple default IPv4 routes leading to SD-CGN

Similarly, in the NAT44 environment, the IPv4 traffic will be send to a number of SD-CGNs using a number of default routes or equivalent techniques.

4.3. 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 NAT44 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.

4.4. 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.

4.5. Example mapping function

Lets say that all external ports above 1023 are used and a 1024 ports are allocated per subscriber, and only one IPv4 address is configured in the NAT pool. Customer #1 will have external ports 1024 to 2047, customer #2 will have external ports 2048 to 3071, etc...
Here is an algorithm that produces this result:
Input variables:

*i: subscriber index

*maxPort: maximum number of ports per subscriber

*baseCPE:CPE address base, determined by Service Provider, e.g. 172.16.0.1

*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:

*host_src_ip is mapped to: CPE WAN IPv4.

*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:

*cpe_src_ip is mapped to: baseCGN + floor (i / P)

*cpe_src_port is mapped to: cpe_src_port + maxPort x (i mod P))

*Where i = cpe_src_ip - baseCPE and P = floor ((65536-1024) / maxPort)

*Note that in this simple algorithm, (65536-1024) must be multiple of maxPort. However, with minor modification to the algorithm, you can avoid this limitation.

4.6. Port exceeded message

If the SD-CGN or SD-AFTR receives an incoming datagram with a source port number higher than1023 + the maximum number of allocated ports per customer, the datagram MUST be dropped and an ICMP [RFC0792] error message type 13 (Communication administratively prohibited) MUST be returned to the originating SD-CPE.

4.7. 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.

4.8. Anycast global 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.

4.9. Stateless operation

By using anycast IPv4 pool, 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.

4.10. 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,...

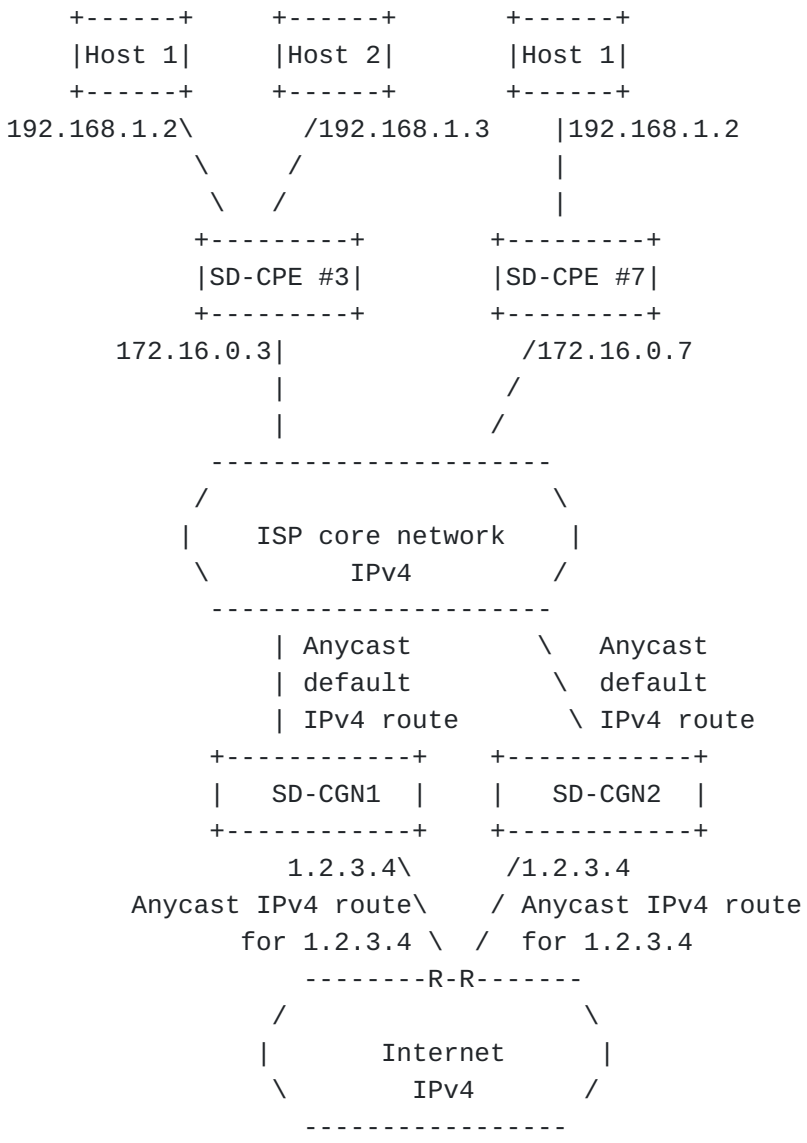
4.11. SD-CGN/SD-AFTR post mapping source port scrambling

In order to improve source port randomization, an SD-CGN or SD-AFTR can do post-mapping source port scrambling. The details of this optional feature are out of scope of this document. The scrambling function MUST be the same on all SD-CGN/SD-AFTR responsible for the same global IPv4 pool, and configured with the same parameters on all those devices. That scrambling function MUST be reversible.

5. 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 1024 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 NAT44 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 NAT44 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.

5.1. SD-NAT44 architecture



5.2. SD-CPE3 NAT bindings in NAT44 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     |
+-------------------------------------------------+
```

5.3. SD-CPE7 NAT bindings in NAT44 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     |
+-------------------------------------------------+
```

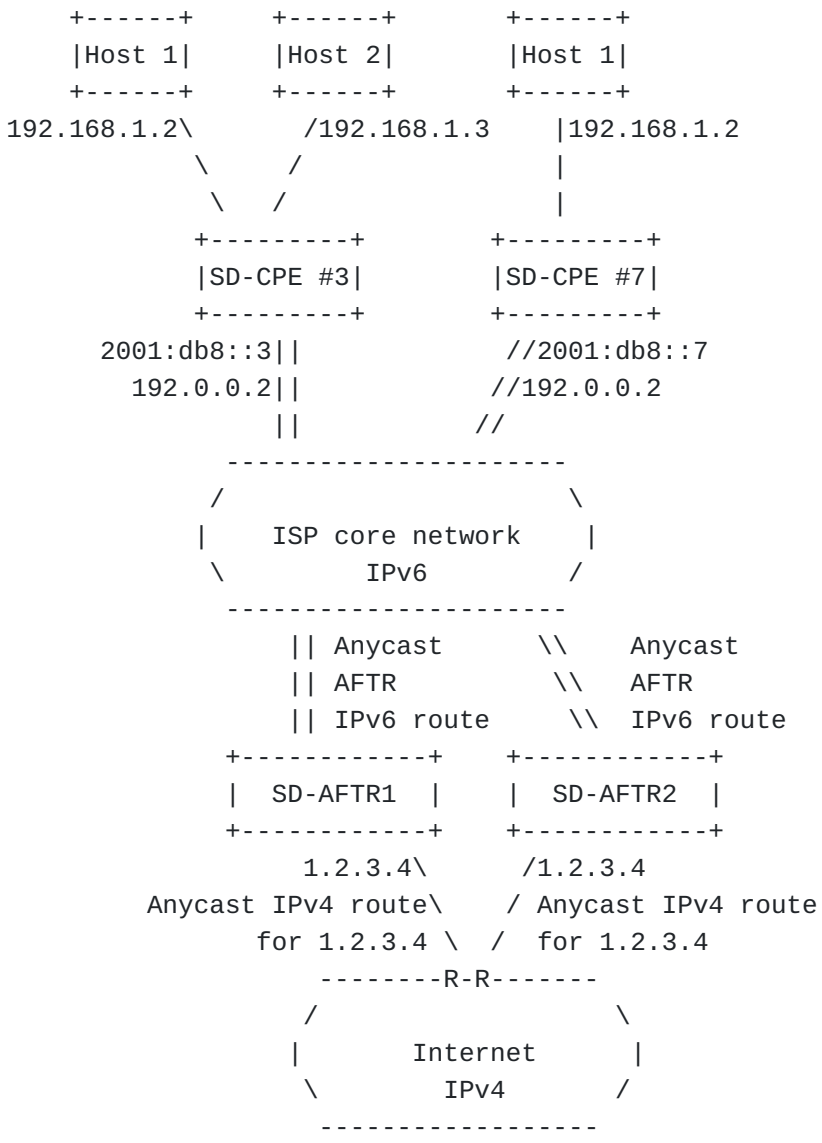
5.4. SD-CGN NAT bindings in NAT44 mode

```
+-------------------------------------------------+
|Connection subscriber3-1:                        |
|SRC IPv4 172.16.0.3 mapped to SRC IPv4 1.2.3.4   |
|SRC port 1024        mapped to SRC port 3072     |
+-------------------------------------------------+
|Connection subscriber3-2                         |
|SRC IPv4 172.16.0.3 mapped to SRC IPv4 1.2.3.4   |
|SRC port 1025       mapped to SRC port 3073      |
+-------------------------------------------------+
|Connection subscriber7-1:                        |
|SRC IPv4 172.16.0.7 mapped to SRC IPv4 1.2.3.4   |
|SRC port 1024        mapped to SRC port 7168     |
+-------------------------------------------------+
```

6. 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 1024 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.

6.1. SD-DS-Lite architecture



6.2. 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      |
+-------------------------------------------------+
```

6.3. 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     |
+-------------------------------------------------+
```

6.4. 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 3072       |
+-------------------------------------------------+
|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 3073       |
+-------------------------------------------------+
|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 7168       |
+-------------------------------------------------+
```

7. Comparison with other techniques

7.1. Comparison with DS-Lite and NAT44 CGN

The trade-off of this stateless deterministic model of operation is efficiency in IPv4 address sharing. A typical DS-Lite or NAT44 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-NAT44 environment will share the same address among only about 100 subscribers, using a maximum number of ports per subscribers of 600.

7.2. 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-CGNs 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.

8. IANA Considerations

None.

9. Security Considerations

As described in [RFC6269], with any fixed size address sharing techniques, port randomization is achieve with a smaller entropy. Post mapping source port scrambling (see section XXX) can mitigate this issue.
Recommendations listed in [RFC6302] applies.

10. References

10.1. Normative references

[RFC0792]	Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.
[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
[RFC6333]	Durand, A., Droms, R., Woodyatt, J. and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.

10.2. Informative references

[RFC6269]	Ford, M., Boucadair, M., Durand, A., Levis, P. and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.
[RFC6302]	Durand, A., Gashinsky, I., Lee, D. and S. Sheppard, "Logging Recommendations for Internet-Facing Servers", BCP 162, RFC 6302, June 2011.
[I-D.ietf-pcp-base]	Wing, D, Cheshire, S, Boucadair, M, Penno, R and P Selkirk, "Port Control Protocol (PCP)", Internet-Draft draft-ietf-pcp-base-17, October 2011.

Authors' Addresses

Reinaldo Penno Penno Juniper Networks 1194 North Mathilda Avenue Sunnyvale, CA 94089-1206 USA EMail: rpenno@juniper.net

Alain Durand Durand Juniper Networks 1194 North Mathilda Avenue Sunnyvale, CA 94089-1206 USA EMail: adurand@juniper.net

Lionel Hoffmann Hoffmann Bouygues Telecom TECHNOPOLE 13/15 Avenue du Marechal Juin Meudon, 92360 France EMail: lhoffman@bouyguestelecom.fr