Internet Engineering Task Force	S. Perreault, Ed.
Internet-Draft	Viagénie
Intended status: Best Current Practice	I. Yamagata
Expires: February 19, 2012	S. Miyakawa
	NTT Communications
	A. Nakagawa
	Japan Internet Exchange (JPIX)
	H. Ashida
	IS Consulting G.K.
	August 18, 2011

Common requirements for Carrier Grade NAT (CGN)

draft-ietf-behave-lsn-requirements-03

Abstract

This document defines common requirements for Carrier-Grade NAT (CGN).

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

With the shortage of IPv4 addresses, it is expected that more ISPs may want to provide a service where a public IPv4 address would be shared by many subscribers. Each subscriber is assigned a private address, and a NAT situated in the ISP's network translates between private and public addresses. This is known as NAT444 [I-D.shirasaki-nat444-isp-shared-addr] when the CPE includes a NAT function.
This is not to be considered a solution to the shortage of IPv4 addresses. It is a service that can conceivably be offered alongside others, such as IPv6 services or regular, un-NATed IPv4 service. Some ISPs started offering such a service long before there was a shortage

of IPv4 addresses, showing that there are driving forces other than the shortage of IPv4 addresses.

This document describes behavioral requirements that are to be expected of those ISP-controlled NAT. Meeting this set of requirements will greatly increase the likelihood that subscribers' applications will function properly.

Readers should be aware of potential issues that may arise when sharing a public address between many subscribers. See [I-D.ford-shared-addressing-issues] for details.

This document builds upon previous works describing requirements for generic NATs [RFC4787][RFC5382][RFC5508]. These documents still apply in this context. What follows are additional requirements, to be satisfied on top of previous ones.

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].

Readers are expected to be familiar with [RFC4787] and the terms defined there. The following additional term is used in this document:

Carrier-Grade NAT (CGN): A NAT-based [RFC2663] functional element operated by an administrative entity (e.g. operator) to share the same address among several subscribers. A CGN is managed by the administrative entity, not the subscribers.

- Note that the term "carrier-grade" has nothing to do with the quality of the NAT; that is left to discretion of implementers. Rather, it is to be understood as a topological qualifier: the NAT is placed in an ISP's network and translates the traffic of potentially many subscribers. Subscribers have limited or no control over the CGN, whereas they typically have full control over a NAT placed on their premises.

Figure 1 summarizes a common network topology in which a CGN operates.

```
                .
                :
                |       Internet
............... | ...................
                |       ISP network
                |
                |
            ++------++  External realm
........... |  CGN   |...............
            ++------++  Internal realm
              |    |
              |    |
              |    |    ISP network
............. | .. | ................
              |    |  Customer premises
      ++------++  ++------++
      |  CPE1  |  |  CPE2  |  etc.
      ++------++  ++------++
```

Another possible topology is one for hotspots, where there is no customer premise or CPE, but where a CGN serves a bunch of customers who don't trust each other and hence fairness is an issue. One important difference with the previous topology is the absence of NAT444. This, however, has no impact on CGN requirements since they are driven by fairness and robustness in the service provided to customers, which applies in both cases.

3. Requirements for CGNs

What follows is a list of requirements for CGNs. They are in addition to those found in other documents such as [RFC4787], [RFC5382], and [RFC5508].

REQ-13 : A CGN MUST support at least the following transport protocols: TCP (MUST support [RFC5382]), UDP (MUST support [RFC4787]), and ICMP (MUST support [RFC5508]). Support for additional transport protocols is OPTIONAL.

Justification: These protocols are the ones that NATs traditionally support. The IETF has documented the best current practices for them.

REQ-14 : A CGN MUST have a default "IP address pooling" behavior of "Paired". The CGN administrator MAY change this behavior on an application protocol basis.

*When multiple overlapping internal address ranges share the same external address pool (e.g. DS-Lite [I-D.ietf-softwire-dual-stack-lite]), external addresses are paired with subscribers rather than internal addresses.

Justification: This stronger form of REQ-2 from [RFC4787] is justified by the stronger need for not breaking applications that depend on the external address remaining constant.

Note that this requirement applies regardless of the transport protocol. In other words, a CGN must use the same external IP address mapping for all sessions associated with the same internal IP address, be they TCP, UDP, ICMP, something else, or a mix of different protocols.

The justification for allowing other behaviors is to allow the administrator to save external addresses and ports for application protocols that are known to work fine with other behaviors in practice. However, the default behavior MUST be "Paired".

REQ-15 : A CGN SHOULD limit the number of external ports (or, equivalently, "identifiers" for ICMP) that are assigned per subscriber.

a. Limits SHOULD be configurable by the CGN administrator.

b. Limits MAY be configured and applied independently per transport protocol.

c. Additionally, it is RECOMMENDED that the CGN include administrator-adjustable thresholds to prevent a single subscriber from consuming excessive CPU resources from the CGN (e.g. rate limit the subscriber's creation of new mappings).

Justification: A CGN can be considered a network resource that is shared by competing subscribers. Limiting the number of external ports assigned to each subscriber mitigates the DoS attack that a subscriber could launch against other subscribers through the CGN in order to get a larger share of the resource. It ensures fairness among subscribers. Limiting the rate of allocation mitigates a similar attack where the CPU is the resource being targeted instead of port numbers.

REQ-16 : A CGN SHOULD limit the amount of state memory allocated per mapping and per subscriber. This may include limiting the number of TCP sessions, the number of filters, etc., depending on the NAT implementation.

a. Limits SHOULD be configurable by the CGN administrator.

b. Additionally, it SHOULD be possible to limit the rate at which memory-consuming state elements are allocated.

Justification: A NAT needs to keep track of TCP sessions associated to each mapping. This state consumes resources for which, in the case of a CGN, subscribers may compete. It is necessary to ensure that each subscriber has access to a fair share of the CGN's resources. Limiting TCP sessions per subscriber and per time unit is an effective mitigation against inter-subscriber DoS attacks. Limiting the rate of allocation is intended to prevent against CPU resource exhaustion.

REQ-17 : It SHOULD be possible to administratively turn off translation for specific destination addresses and/or ports.

Justification: It is common for a CGN administrator to provide access for subscribers to servers installed in the ISP's network, in the external realm. When such a server is able to reach the internal realm via normal routing (which is entirely controlled by the ISP), translation is unneeded. In that case, the CGN may forward packets without modification, thus acting like a plain router. This may represent an important efficiency gain.

Figure 2 illustrates this use-case.

```
X1:x1            X1':x1'            X2:x2
+---+from X1:x1  +---+from X1:x1    +---+
|   |  to X2:x2  |   |  to X2:x2    | S |
| C |>>>>>>>>>>>>| C |>>>>>>>>>>>>>>| e |
| P |            | G |              | r |
| E |<<<<<<<<<<<<| N |<<<<<<<<<<<<<<| v |
|   |from X2:x2  |   |from X2:x2    | e |
|   |  to X1:x1  |   |  to X1:x1    | r |
+---+            +---+              +---+
```

REQ-18 : It is RECOMMENDED that a CGN have an "Endpoint-Independent Filtering" behavior.

Justification: This is a stronger form of REQ-8 from [RFC4787]. An "Address-Dependent Filtering" behavior is NOT RECOMMENDED. This is based on the observation that some games and peer-to-peer applications require EIF for the NAT traversal to work. In the context of a CGN it is important to minimise application breakage.

REQ-19 : When a CGN loses state (due to a crash, reboot, failover to a cold standby, etc.), it MUST NOT reuse the same external address+port pairs for new dynamic mappings for at least 120 seconds, except for the following cases:

a. If the CGN tracks TCP sessions (e.g. with a state machine, as in [RFC6146] section 3.5.2.2), TCP ports MAY be reused immediately.

b. If the allocated external ports used address-dependent or address-and-port-dependent filtering before state loss, they MAY be reused immediately.

Justification: This is necessary in order to prevent collisions between old and new mappings and sessions. It ensures that all established sessions are broken instead of redirected to a different peer.

The exceptions are for cases where reusing a port immediately does not create a possibility that packets would be redirected to the wrong peer.

The 120 seconds value corresponds to the Maximum Segment Lifetime (MSL) from [RFC0793].

One way that this requirement could be satisfied would be have two distinct address pools: one dormant and one active. When rebooting, the CGN would swap the dormant pool with the active pool. Another way would be simply to wait 120 seconds before resuming NAT activity.

REQ-20 : Once an external port is deallocated, it SHOULD NOT be reallocated to a new mapping until at least 120 seconds have passed. The length of time and the maximum number of ports in this state SHOULD be configurable by the CGN administrator. The following exceptions apply:

a. If the CGN tracks TCP sessions (e.g. with a state machine, as in [RFC6146] section 3.5.2.2), TCP ports MAY be reused immediately.

b. If the allocated external ports used address-dependent or address-and-port-dependent filtering before state loss, they MAY be reused immediately.

Justification: This is to prevent users from receiving unwanted traffic. It also helps prevent against clock skew when mappings are logged.

The exceptions are for cases where reusing a port immediately does not create a possibility that packets would be redirected to the wrong peer.

The 120 seconds value corresponds to the Maximum Segment Lifetime (MSL) from [RFC0793].

REQ-21 : A CGN SHOULD include a Port Control Protocol server [I-D.ietf-pcp-base].

Justification: Allowing subscribers to manipulate the NAT state table with PCP greatly increases the likelihood that applications will function properly.

REQ-22 : A CGN SHOULD support [RFC4008].

Justification: It is anticipated that CGNs will be primarily deployed in ISP networks where the need for management is critical.

Note also that there are efforts within the IETF toward creating a MIB specifically for CGNs [I-D.jpdionne-behave-cgn-mib].

REQ-23 : When packets pass from one side to the other, the DSCP values MUST be preserved. If the CGN also includes diffserv classifier and marker functionality it MAY change the DSCP values.

Justification: See [RFC2983], in particular section 6.

REQ-24 : When a CGN is unable to create a mapping due to resource constraints or administrative restrictions (i.e. quotas)...

a. it MUST drop the original packet;

b. it SHOULD send an ICMP Destination Unreachable message with code 3 (Port Unreachable) to the session initiator;

c. it SHOULD send a notification (e.g. SNMP trap) towards a management system (if configured to do so);

d. and it SHOULD NOT delete existing mappings in order to "make room" for the new one.

Justification: This is a slightly different form of REQ-8 from [RFC5508]. Code 3 is preferred to code 13 because it is listed as a "soft error" in [RFC5461], which is important because we don't want

TCP stacks to abort the connection attempt in this case. Sending an ICMP error may be rate-limited for security reasons, which is why requirement B is a SHOULD, not a MUST.

Applications generally handle connection establishment failure better than established connection failure. This is why dropping the packet initiating the new connection is to preferred to deleting existing mappings. See also the rationale in [RFC5508] section 6.

4. Logging

It may be necessary for CGN administrators to be able to identify a subscriber based on external IPv4 address, port, and timestamp in order to deal with abuse and lawful intercept requests. When multiple subscribers share a single external address, the source address and port that are visible at the destination host have been translated from the ones originated by the subscriber.
In order to be able to do this, the CGN would need to log the following information for each mapping created:

* subscriber identifier (e.g. internal source address or tunnel endpoint identifier)

* external source address

* external source port

* destination address (but see below)

* destination port (but see below)

* timestamp

By "subscriber identifier" we mean information that uniquely identifies a subscriber. For example, in a traditional NAT scenario, the internal source address would be sufficient. In the case of DS-Lite, many subscribers share the same internal address and the subscriber identifier is the tunnel endpoint identifier (i.e. the B4's IPv6 address).
A disadvantage of logging mappings is that CGNs under heavy usage may produce large amounts of logs, which may require large storage volume. Readers should be aware of logging recommendations for Internet-facing servers [I-D.ietf-intarea-server-logging-recommendations]. With compliant servers, the destination address and port do not need to be logged by the CGN. This can help reduce the amount of logging.

5. Bulk Port Allocation

So far we have assumed that a CGN allocates one external port for every outgoing connection. In this section, the impacts of allocating multiple external ports at a time are discussed.
There is a range of things a CGN can do:

Traditional: For every outgoing connection, allocate one external port.

Scattered port set: For an outgoing connection, create a set of several non-consecutive external ports. Subsequent outgoing connections will use ports from the set. When the set is exhausted, a new connection causes a new set to be created. A set is smaller or equal to the user's maximum port limit.

Consecutive port set: Same as the scattered port set, but the ports allocated to a set are consecutive.

Note that this list is not exhaustive. There is a continuum of behavior that a CGN may choose to implement. For example, a CGN could use scattered port sets of consecutive port sets.
The impacts of bulk port allocation are as follows.

Port Utilization: The mechanisms at the top of the list are very efficient in their port utilization. In that sense, they have good scaling properties (nothing is wasted). The mechanisms at the bottom of the list will waste ports. The number of wasted ports is proportional to size of the "bin".

Logging: Traditional allocation creates a lot of log entries as compared to allocation by port sets create much fewer entries. Scattered and consecutive port sets generate the same number of log entries. In the case of consecutive port sets, entries can be expressed very compactly by indicating a range (e.g. "12000-12009"). Some scattered port set allocation schemes can also generate small log entries containing the parameters and algorithm used for the port set generation (see e.g. [I-D.boucadair-pppext-portrange-option]).

With large set sizes, the logging frequency for scattered and consecutive port sets can approach that of DHCP servers.

Logging destination addresses and ports can only be done on a per-session basis. This means that destination logging for a CGN implementing bulk port allocation would create one log entry per session containing the destination address and port. Other information could still be logged in one entry per port set.

Security: Traditional and scattered port sets provide very good security in that ports numbers are not easily guessed. Easily guessed port numbers put subscribers at risk of the attacks described in [RFC6056]. Consecutive port sets provides poor security to subscribers, especially if the set size is small.

6. Deployment Considerations

Several issues are encountered when CGNs are used [I-D.ietf-intarea-shared-addressing-issues]. There is current work in the IETF toward alleviating some of these issues. For example, see [I-D.boucadair-intarea-nat-reveal-analysis].
The address sharing ratio is the ratio between the number of external addresses and the number of internal addresses that a CGN is configured to handle. See [I-D.ietf-intarea-shared-addressing-issues] section 26.2 for guidance on picking an appropriate ratio.

7. IANA Considerations

There are no IANA considerations.

8. Security Considerations

If a malicious subscriber can spoof another subscriber's CPE, it may cause a DoS to that subscriber by creating mappings up to the allowed limit. Therefore, the CGN administrator SHOULD ensure that spoofing is impossible. This can be accomplished with ingress filtering, as described in [RFC2827].

9. Acknowledgements

Thanks for the input and review by Arifumi Matsumoto, Benson Schliesser, Dai Kuwabara, Dan Wing, Dave Thaler, Francis Dupont, Joe Touch, Lars Eggert, Kousuke Shishikura, Mohamed Boucadair, Nejc Skoberne, Reinaldo Penno, Senthil Sivakumar, Takanori Mizuguchi, Takeshi Tomochika, Tomohiro Fujisaki, Tomohiro Nishitani, Tomoya Yoshida, and Yasuhiro Shirasaki. Dan Wing also contributed much of section 5.

10. References

10.1. Normative References

[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
[RFC4008]	Rohit, R., Srisuresh, P., Raghunarayan, R., Pai, N. and C. Wang, "Definitions of Managed Objects for

	Network Address Translators (NAT)", RFC 4008, March 2005.
[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.
[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.

10.2. Informative Reference

[RFC0793]	Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.
[RFC2663]	Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.
[RFC2827]	Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.
[RFC2983]	Black, D., "Differentiated Services and Tunnels", RFC 2983, October 2000.
[RFC5461]	Gont, F., "TCP's Reaction to Soft Errors", RFC 5461, February 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.
[I-D.ietf-intarea-server-logging-recommendations]	Durand, A, Gashinsky, I, Lee, D and S Sheppard, "Logging recommendations for Internet facing servers", Internet-Draft draft-ietf-intarea-server-logging-recommendations-04, April 2011.
[I-D.ietf-intarea-shared-addressing-issues]	Ford, M, Boucadair, M, Durand, A, Levis, P and P Roberts, "Issues with IP Address Sharing", Internet-Draft draft-ietf-intarea-shared-addressing-issues-05, March 2011.
[I-D.ietf-softwire-dual-stack-lite]	Durand, A, Droms, R, Woodyatt, J and Y Lee, "Dual-Stack Lite Broadband Deployments

	Following IPv4 Exhaustion", Internet-Draft draft-ietf-softwire-dual-stack-lite-11, May 2011.
[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", Internet-Draft draft-boucadair-intarea-nat-reveal-analysis-04, September 2011.
[I-D.boucadair-pppext-portrange-option]	Boucadair, M, Levis, P, Bajko, G, Savolainen, T and T Tsou, "Huawei Port Range Configuration Options for PPP IPCP", Internet-Draft draft-boucadair-pppext-portrange-option-09, September 2011.
[I-D.ford-shared-addressing-issues]	Ford, M, Boucadair, M, Durand, A, Levis, P and P Roberts, "Issues with IP Address Sharing", Internet-Draft draft-ford-shared-addressing-issues-02, March 2010.
[I-D.jpdionne-behave-cgn-mib]	Dionne, J and M Blanchet, "CGN Management Information Base (MIB)", Internet-Draft draft-jpdionne-behave-cgn-mib-00, July 2011.
[I-D.shirasaki-nat444-isp-shared-addr]	Yamaguchi, J, Shirasaki, Y, Miyakawa, S, Nakagawa, A and H Ashida, "NAT444 addressing models", Internet-Draft draft-shirasaki-nat444-isp-shared-addr-06, July 2011.

Appendix A. Change Log (to be removed by RFC Editor prior to publication)

Appendix A.1. Changed in -03

* Added exceptions for which it is not necessary to wait 120 seconds before reusing a port.

* Renamed "random port set" to "scattered port set", which is more accurate.

* Log "subscriber identifier" instead of internal address+port to allow for overlapping internal address ranges (DS-Lite).

* Adjusted logging text and added reference to I-D.boucadair-pppext-portrange-option.

* Adjusted destination logging text for bulk port allocation schemes.

* Removed requirement for I-D.ietf-intarea-ipv4-id-update.

* Made PCP support a SHOULD-level requirement.

* Lowered the level of requirement for not dropping existing mappings in order to "make room" to SHOULD level, and added rationale.

Appendix A.2. Changed in -02

* CGNs MUST support at least TCP, UDP, and ICMP.

* Add requirement from I-D.ietf-intarea-ipv4-id-update.

* Add informative reference to [I-D.ietf-intarea-shared-addressing-issues].

* Add requirement (SHOULD level) for a port forwarding protocol.

* Allow any pooling behavior on a per-application protocol basis.

* Adjust wording for external port allocation rate limiting.

* Add requirement for RFC4008 support (SHOULD level).

* Adjust wording for swapping address pools when rebooting.

* Add DSCP requirement (stolen from draft-jennings-behave-nat6).

* Add informative reference to draft-boucadair-intarea-nat-reveal-analysis.

* Add requirement for hold-down pool.

* Change definition of CGN.

* Avoid usage of "device" loaded word throughout the document.

* Add requirement about resource exhaustion.

* Change title.

* Describe additional CGN topology where there is no NAT444.

* Better justification for "Paired" pool behavior.

* Make it clear that rate limiting allocation is for preserving CPU resources

* Generalize the requirement for limiting the number of TCP sessions per mapping so that it applies to all memory-consuming state elements.

* Change CPE to subscriber where it applies throughout the text.

* Better terminology for bulk port allocation mechanisms.

* Explain how external address pairing works with DS-Lite.

Appendix A.3. Changed in -01

* Terminology: LSN is now CGN.

* Imported all requirements from RFCs 4787, 5382, and 5508. This allowed us to eliminate some duplication.

* Added references to draft-ietf-intarea-server-logging-recommendations and draft-ford-shared-addressing-issues.

* Incorporated a requirement from draft-xu-behave-stateful-nat-standby-06.

Authors' Addresses

Simon Perreault editor Perreault Viagénie 2875 boul. Laurier, suite D2-630 Québec, QC G1V 2M2 Canada Phone: +1 418 656 9254 EMail: simon.perreault@viagenie.ca URI: http://www.viagenie.ca

Ikuhei Yamagata Yamagata NTT Communications Corporation Gran Park Tower 17F, 3-4-1 Shibaura, Minato-ku Tokyo, 108-8118 Japan Phone: +81 50 3812 4704 EMail: ikuhei@nttv6.jp

Shin Miyakawa Miyakawa NTT Communications Corporation Gran Park Tower 17F, 3-4-1 Shibaura, Minato-ku Tokyo, 108-8118 Japan Phone: +81 50 3812 4695 EMail: miyakawa@nttv6.jp

Akira Nakagawa Nakagawa Japan Internet Exchange Co., Ltd. (JPIX) Otemachi Building 21F, 1-8-1 Otemachi, Chiyoda-ku Tokyo, 100-0004 Japan Phone: +81 90 9242 2717 EMail: a-nakagawa@jpix.ad.jp

Hiroyuki Ashida Ashida IS Consulting G.K. 12-17 Odenma-cho Nihonbashi Chuo-ku Tokyo, 103-0011 Japan EMail: assie@hir.jp