Carrier-Grade-NAT (CGN) Deployment Considerations.
draft-nishizuka-cgn-deployment-considerations-00

Abstract

   This document provides deployment considerations for Carrier-Grade-
   NAT (CGN) based on the verification result include the investigation
   of the number of sessions of applications. The verification was
   conducted in StarBED which is one of the largest scale network
   experiment environment in Japan. A million of subscribers was
   emulated and it revealed the realistic behavior of CGN.

described in the Simplified BSD License.

Table of Contents

1. Introduction

IP address sharing is tentative technic to deal with the shortage of
IPv4 addresses. As described in [RFC6269], IP address sharing causes
many issues such as application failures and security
vulnerabilities. A part of these issues is based on the assigned
number of sessions per user and port allocation method of CGN. This
document lists the number of port consumption of major application
and web sites.

The efficiency of CGN is based on the number of sessions allocated to
each user and the installation location. This document also
describes the deployment considerations of CGN to specify the optimum
place according to CGN performance. CGN performance was
experimentally-verified with realistic traffic generated by amount of
emulated users.

The growth of IPv6 is continual solution of the shortage of IPv4
addresses and frees these issues. By adopting the combination of the
IPv4 shared address and native IPv6, the duty of CGN will decrease
and as the result, the bad effect on applications which are caused by
the limitation of available ports and address translation itself and
security vulnerability will be resolved. The most effective way of
deploying CGN is examined in this document. Further discussion about
the integration of CGN into the existing network is studied in
[I-D.ietf-opsawg-lsn-deployment].

2. Conventions used in this document

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

3. Motivation

With a progressive exhaustion of IPv4 addresses, the demands for
sharing IPv4 addresses with multiple customers are rapidly rising,
thus many proposals are getting much attention include Carrier Grade
NAT (CGN, or LSN for Large Scale NAT)
[I-D.ietf-behave-lsn-requirements], Dual-Stack Lite [RFC6333], NAT64
[RFC6146], Address+Port (A+P) [RFC6346], 464XLAT
[I-D.ietf-v6ops-464xlat] and MAP [I-D.ietf-softwire-map]. The
practical configuration of these method is based on the same
considerations as follows:

     - Stateful or Stateless
     - Centralized or Distributed
     - Dynamic port assignment or Static port assignment
     - Log reduction strategy
     - Security considerations

   The best practice about these considerations should be derived from
   realistic experiment because there are pros and cons. Though we
   tested them in NAT444 environment, the result is applicable for other
   approaches. The investigation of number of sessions is described in
   this document and it can be also helpful for all of them.

4. **The number of sessions of applications**

   The number of concurrent sessions of applications is important factor
   of designing of CGN because there is trade-off between the efficiency
   of IPv4 address saving and the availability of those applications.
   In addition, for security and fairness, we should limit the number of
   sessions per user. As described in [RFC6269], infected devices could
   rapidly exhaust the available ports of global pool addresses, hence
   all the rest of users could not through the CGN anymore. So to place
   the CGN to existing network, we should know how many sessions are
   sufficient for every user. Here is a list of applications and their
   average sessions. We selected and tested 50 sites from the list of
   top sites and remarkable applications.

Application	Total sessions	TCP port80	TCP port443	UDP port53
Web mail	65	35	30	20
Video	83	77	6	20
Portal site	47	47	0	13
EC site	45	43	2	11
blog	61	59	2	17
Search Engine	8	8	0	4
Online Banking	20	2	18	4
Cloud Service	29	23	6	6
iTunes	20	1	19	7
Twitter	33	1	32	12
Twitter(mobile)	14	2	11	3
facebook	51	40	11	18
facebook(mobile)	18	11	7	10
Game	95	86	9	19

Figure 1: The number of sessions of applications.

Figure 1

The number of sessions of these applications are up to 100 sessions.
There are no longer high-consumption applications. This observation
implies that modern applications such as facebook have changed to use
multiplexed requests. Previously, to achieve high-performance web,
it abused HTTP sessions. Now, current cutting edge technologies such
as WebSocket, SPDY and HTTP2.0 avoid such an abusing. Basically,
multiple requests are multiplexed into one TCP connection. However,
a kind of game applications still consume many sessions.

The last factor of the estimation of number of sessions is how many

applications are used simultaneously within a single CPE (Customer
Premises Equipment) which includes non-PC devices like gaming
devices. Our investigation shows that the average number of session
of active subscriber is 400. We daresay the limitation of 1000
sessions per user would not affect the most of users while preventing
the severe abuse from certain users.

5. Feasibility of port assignment methods

   Basing on the investigation of the number of sessions of
   applications, the realistic parameter of each port assignment method
   was estimated by the verification.

5.1. Port assignment methods

   The efficiency of IPv4 saving by CGN is highly depending on how to
   allocate the ports of pool addresses to each users. There are 2
   major methods: dynamic assignment and static assignment
   [I-D.chen-sunset4-cgn-port-allocation]. There are combined problem
   involving efficiency of address saving and logging information
   reduction. Typical IP Network Address Translator (NAT)
   [RFC2663][RFC2993] implementation uses dynamic assignment, so NAT444,
   NAT64, DS-lite and 464XLAT are originally dynamic assignment
   approach. To avoid the huge amount of information needed to be
   recorded, those approaches have variations of static assignment
   [I-D.donley-behave-deterministic-cgn] and MAP is inherently static
   assignment approach. For taking advantage of both methods, the
   hybrid method that is dynamic assignment of port ranges has been
   implemented in some CGN. The merits of the port block assignment
   have been referred in [RFC6346],
   [I-D.donley-behave-deterministic-cgn] and
   [I-D.chen-sunset4-cgn-port-allocation].

5.2. Efficiency of address saving

   In the dynamic assignment, the ports of pool address are allocated
   randomly for active users. This method can use pool addresses and
   ports most effectively. The average number of port consumption (N)
   per active subscriber is the key value for dynamic assignment. In
   the verification, the average number of port consumption (N) was
   estimated to be 400. At the same time, user-quota of 1000 sessions
   was set to avoid the abuse. The percentage of the active subscribers
   (a) was estimated to be 25% at the value during the busy hour of
   traffic (21:00 pm to 1:00am). In this time, "active" subscriber
   means who create a new session in certain period of time. Then, when
   a CGN adopt the dynamic assignment, the required number of the pool
   address is as follows:

# of pool address (P) = # of Subscriber (S) * a * N / (65536 - R)

Here, (R) is reserved TCP/UDP port list referred in
[I-D.donley-behave-deterministic-cgn]. CGN should eliminate the
wellknown ports (0-1023 for TCP and UDP) to avoid the bad
interpretation from destination servers. It is natural to translate
source port of outgoing packet to ephemeral ports. The ports after
32768 is considered to be used without any problems because ephemeral
ports are 32768-61000 for Linux and 49152-65535 in the IANA
recommendations. As the result, 3,051 pool addresses are sufficient
for 1,000,000 subscribers. The feasibility of configuration was
confirmed in the verification.

On the other hand, in static assignment, the ports are allocated a
priori for every users. The pool addresses and ports are reserved to
every users, so most of them could be a dead stock because there are
light users and heavy users in aspect of port consumption. The max
number of port consumption in all subscribers is the key value for
static assignment. The true peak number of the session by a heavy
user could be over 10,000 sessions. However it is assumed that such
a severe consumption of ports to be an abuse, so the number of
statically assigned port (M) is controllable parameter by each
providers. In the static assignment, the required number of the pool
address is as follows:

# of pool address (P) = # of Subscriber (S) * M / (65536 - R)

Taking account into the investigation of number of sessions of
applications, the desirable value of (M) is over 1,000. As the
result, no less than 30,517 pool addresses are needed for 1,000,000
subscribers. The compression ratio is one tenth of the case of
dynamic assignment.

5.3. Logging design

5.3.1. Amount of the NAT log

In the aspect of the efficiency, the dynamic assignment is preferable
than the static assignment. However the size of the log is the main
consideration. There is a case that providers must identify a user
to respond abuse or public safety requests. Conventionally, source
IP address and a timestamp are needed. It was possible to identify a
user by comparing IP address with authentication logs of the exact
time. However, when IP address is shared by the CGN, it is necessary
to compare the translated address and port information which are
given by the destination host with the NAT log to identify the
untranslated IP address. According to the
[I-D.ietf-behave-lsn-requirements], following information is

recommended to log (for NAT444):

- Transport Protocol
- Source IP address:port
- Source IP address:port after translation
- Timestamp

In addition, the indicator of the allocation and deallocation are
needed because it assures that the identified subscriber certainly
had been using the translated IP address and port. Plus, including
the index of the CGN host, the average size of NAT log is about 120
byte in ASCII format. Every active subscriber generate 400 sessions
in average for a certain amount of time. It is assumed that the
event happens every 5 minute in the most severe condition. The size
of the log (L) for time frame (T) can be estimated as follows:

The size of log (L) = # of Subscriber (S) * a * N * 120byte * 2 * (
Time frame(T) / 5 min.)

It should be noted that the log is generated at the timing of NAT
table creation and freeing. As the result, for 1,000,000 users, the
size of log is piled up to 6.4 terabytes per day. The verification
result confirm the existing estimation referred in
[I-D.donley-behave-deterministic-cgn].

The size of the log can be reduced without loss of the amount of
information. Compact format is the technique of reducing the amount
of log by using a notational change (hexadecimal number). It was
confirmed by verification that the compact format can reduce amount
of log to about 80% as compared with ASCII format. Though it was not
tested, theoretically binary format is the smallest notation and
amount of log can be reduced to 30%.

In static allocation, the amount of log is dramatically reduced even
to zero because the untranslated IP address and the translated IP
address / port range are mapped a priori.

5.3.2. Necessity for destination information

In [RFC6269], it is pointed out that only providing information about
the external address to a service provider is no longer sufficient to
unambiguously identify customers. One of the solutions is the method
of recording the source port information (and exact time stamp)
additionally by the destination server or FW, which is demanded in
[RFC6302]. The other solution is the method of recording destination
IP address and port information by CGN of service provider. The both
solutions are imperfect. In [I-D.tsou-behave-natx4-log-reduction],
it is noted that source port recording is not supported by every

application. Thus, to increase the certainty, additional logging of
destination address and port is effective measure to deal with the
legal request from servers which are not compliant with [RFC6302].
In dynamic assignment, to log destination address is additional. It
is confirmed by the verification that by logging destination address,
only 4% of amount of log is increased in ASCII format. On the other
hand, in static assignment, logging of every session is newly
required and it has the same amount of log as the dynamic assignment.
It completely breaks the merit of the static assignment.

6. Scalability of CGN

The estimation of efficiency of address saving and the logging design
are depending on the number of subscribers accommodated with a CGN.
The scalability of the current CGN was verified by the measurement of
the performance.

6.1. Performance of CGN

According to the experimental results, there are three base
capacities to indicate CGN performance as follows:

- Through put
- MCS: Max Concurrent Sessions
- CPS: Connections per Sec

These capacities are not independent of each other, but become mixed
load for CGN. Each load will be combined in real network traffic,
thus using subscriber emulated traffic is important for measuring the
performance in realistic way.

Through put is forwarding performance of CGN. Currently CGN
equipments with an IF of 1GigabitEthernet and 10GigabitEthernet are
flagship models of the manufactures, but CGN has an upper limit
internally because the performance depends on internal devices such
as CPUs. By ON / OFF of ALGs (Application Level Gateway), the
forwarding performance will be affected because the traffic process
is possibly changed to the path through CPU.

MCS shows an upper limit of numbers of records kept in NAT table.
Number of holding sessions depends on retention time of NAT table.
That is because, even after the end of data transmission, the NAT
table is held in a certain period of time to guarantee the behavior
of an application. As described in
[I-D.ietf-behave-lsn-requirements] REQ-8, if the CGN tracks TCP
sessions, NAT tables may be released when RST or FIN of TCP has been
observed. In case of TCP session where RST or FIN session has not

been observed, and UDP and ICMP communication, NAT table should
retain a certain amount of time. Also, in case of Full Cone NAT, a
table of Full Cone NAT also should retain a certain time to await
communication from outside for a certain period of time. It is
effective to shorten the time-out value in order to suppress the
overflowing NAT table, but it is needed to be careful not to inhibit
the behavior of the application. It is desirable that retention time
of NAT table is configurable as time-out value. In the experiment,
the time-out values are as follows:

```
   +------------------+---------+--------+--------+--------+--------+
   |    Protocols     |  TCP    |  TCP   |  UDP   |  DNS   |  ICMP  |
   |                  |         |  SYN   |        |        |        |
   +------------------+---------+--------+--------+--------+--------+
   | Time-out Value   |  300    |   60   |  300   |   3    |   2    |
   +------------------+---------+--------+--------+--------+--------+
```
   Figure 2: The time-out values (sec) in the experiment.

                              Figure 2

These settings didn't break the behavior of applications.

It is very difficult to estimate maximum number of concurrent
sessions in the network where traffic is already flowing. Maximum
number of concurrent sessions was estimated to be 1M sessions per
10,000 users by our assumption as follows:

Max Concurrent Sessions (MCS) = # of Subscriber (S) * a * N

As the result, it is verified that tested current CGN is able to have
16M sessions for 160,000 subscribers with the capability of the
dynamic assignment and logging. It means that introducing CGN up to
about 15G traffic section is capable, which implies that CGN can be
placed to more centralized position of the network. In summary, the
settings and the performance result are as follows:

```
    +-----------------------------------------------+
    |                                |   Assumed    |
    |                                |   Values     |
    +--------------------------------+--------------+
    |    average # of sessions(N)    |     400      |
    +--------------------------------+--------------+
    | % of the active subscribers (a)|      25      |
    +--------------------------------+--------------+
    |                                |   Verified   |
    |                                |   Values     |
    +--------------------------------+--------------+
    |    # of Subscriber (S)         |   160,000    |
    +--------------------------------+--------------+
    |    Max Concurrent Sessions(MCS)| 16,000,000   |
    +--------------------------------+--------------+
    |    Connection Per Sec(CPS)     |    30,000    |
    +--------------------------------+--------------+
    |    # of pool address (P)       |    4,000     |
    +--------------------------------+--------------+
    |    size of log (L) (in 10min)  |    7.0GB     |
    +--------------------------------+--------------+
```

          Figure 3: The performance results of tested CGN.

                              Figure 3

   In the verification, session arrival rate by emulated subscribers was
   not so high because the load of concurrent sessions is noticeable in
   the equipment used in the experiment. There were no problems in weak
   load of about 30,000 CPS. In case that traffic flows suddenly change
   to standby equipment in redundant network, CPS performance becomes
   rate-limiting, so CPS performance is also important factor to
   minimize the effect of failures.

6.2. Redundancy features of CGN

   It is often referred that introduction of CGN could create Single
   Point of Failure(SPOF) (ex. in [RFC6269]). CGN is stateful, in
   contrast to stateless BR of MAP, so the redundant configuration must
   be achieved by the synchronization of the NAT table between redundant
   equipments. Moreover, introduction of CGN creates layer 3 boundary
   to NATed traffic, so the redundancy features may work with around
   routers via dynamic routing. Nevertheless, it is verified that
   current CGN can be configured and introduced to service providers
   network with the redundancy features. In the verification, CGN can
   be switched to another CGN with sub-sec loss of traffic even in the
   situation that they holds 16M concurrent sessions.

6.3. DNS query traffic considerations

   How to deal with the DNS query traffic is unignorable concern for
   deploiment of CGN. In the test scenario, control experiment was
   conducted to reveal the impact of the huge amount of DNS queries.


```
                   Access Node           CGN
          Emulated                                   +-----------+
          Subscribers  +-------+         +-------+    |           |
            +----+     |       |         |       |    |           |
            | --+-----+   R    +---------+ CGN   +------+          |
            +----+     |       |         |       |    | Core      |
                       +---+---+         +-------+    | Network   |
                           |                         |           |
                           |                         |           |
                           |             +-------+   | +-------+ |
                           |             |       |   | |       | |
                           +-------------+  DNS  +------+ | DNS   | |
                                         |       |   | |       | |
                                         +-------+   | +-------+ |
                                                     +-----------+

                              DNS Forwarder
```
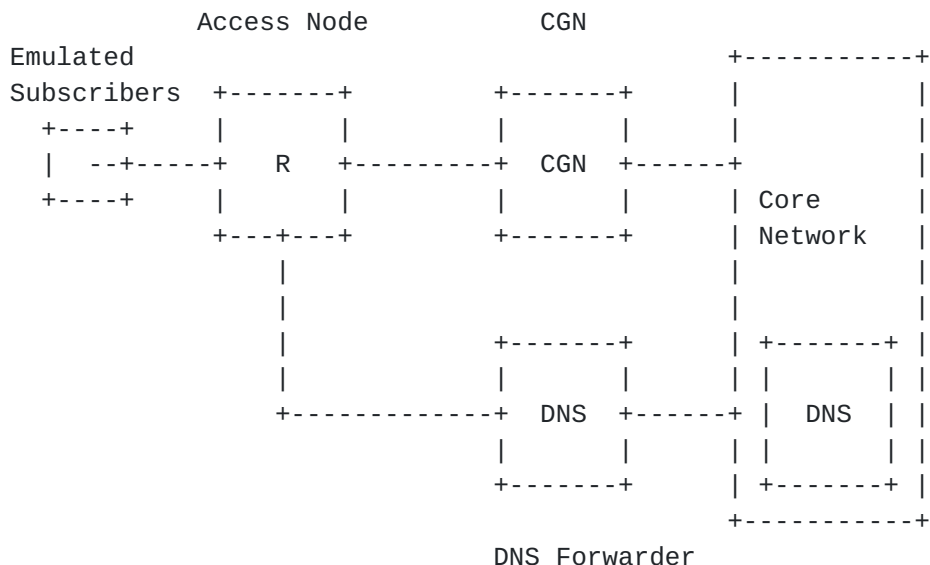
      Figure 4: Bypassing of DNS queries using DNS forwarder.

   In the first case, the original DNS server IP address in the service
   provider network is distributed to the subscribers. The emulated
   subscribers use the DNS server to get host IP address by name, so all
   query packets go through the CGN. The generated DNS query is 12M at
   the speed of 10k query per sec. In the second case, IP address of a
   DNS server placed in the bypassing position of CGN is used instead.
   The second DNS server works as a forwarder, so all queries are
   forwarded to the first DNS server. Therefore, all DNS queries are
   bypassed from the CGN while data traffic is still going through the
   CGN.

   As the result, it was shown that DNS query almost does not affect the
   performance of the CGN. The max concurrent sessions of DNS packet
   was only 40k. NAT table of DNS timeouts in 3 seconds, thus It saves
   the consumption of NAT tables. However NAT log was generated for
   every query and it doubled the total amount of the log. It would be
   rare that the NAT log of DNS is needed to react to a legal request.
   The impact of the DNS query traffic is relatively small if DNS
   timeout is adjusted.

6.4. Separation of traffic

   In the existing network, IPv4 communication and IPv6 communication
   may already be mixed in the dual stack. In this case, by introducing
   CGN which can route IPv6 and existing IPv4 aside from NAT function,
   the influence for the network architecture could be suppressed and so
   a flexible design is possible. However, though current CGN is
   scalable enough to be deployed in core of the service providers
   network, the feature of routing is insufficient to replace the
   exsisting routers. Such a CGN is desirable, otherwise the design
   which makes IPv6 traffic and traditional IPv4 traffic bypass from CGN
   is effective choice for providers. In dividing NAT flows and non-NAT
   flows routers around the CGN, VRF (Virtual Routing and Forwarding)
   and PBR (policy based routing) are needed in a router under CGN. In
   that case it is indispensable to configure routers so that the
   hairpining communication between the NAT user and non-NAT user to be
   possible. The considerations about the separation of traffic and
   effective deployment configuration are discussed in detail in
   [I-D.ietf-opsawg-lsn-deployment].

7. Tested web sites and applications (Excerpts)

   - Web Mail
      gmail
      yahoo mail
      hot mail
   - Video
      ustream
      youtube
      nicovideos
      Hulu
      dailymotion
      daum
      qq
      fc2
      xvideos
   - Portal&EC site
      yahoo
      rakuten
      amazon
      apple
   - Blog
      livedoor blog

         ameba blog
      - Search Engine
         google
      - Online Banking
         mizuho bank
         DC card
      - Cloud Service
         drop box
         Evernote
      - InstantMessenger & VoIP
         skype
         Line
      - facebook
      - twitter
      - google map
      - Online PC Game
         aeria games
         ameba pigg
         nexon
         hangame
      - Consumer Game
         Armored Core V (Play Station3)
         Dark Souls 2 (Play Station3)
         Gundam Extreme VS. (Play Station3)
         Kinect adventure (XBox)
         Persona 4 the ultimate in mayonaka arena (XBox)
         Mingol 4 (WiiU)
         Monster Hunter 3G (DS-lite)
         Keri-hime sweets (iOS)
         PuzzDra (iOS)

[8](). IANA Considerations

   This document makes no request of IANA.

[9](). Security Considerations

   TBD

[10](). Acknowledgments

   This research and experiment are conducted under the great support of
   Ministry of Internal Affairs and Communications of Japan. Many
   thanks to MIC, JAIST members and Shin Miyakawa for their ideas and
   feedback in documentation.

11. References

11.1. Normative References

[I-D.donley-behave-deterministic-cgn]
          Donley, C., Grundemann, C., Sarawat, V., Sundaresan, K.,
          and O. Vautrin, "Deterministic Address Mapping to Reduce
          Logging in Carrier Grade NAT Deployments",
          draft-donley-behave-deterministic-cgn-05 (work in
          progress), January 2013.

[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.ietf-opsawg-lsn-deployment]
          Kuarsingh, V. and J. Cianfarani, "CGN Deployment with BGP/
          MPLS IP VPNs", draft-ietf-opsawg-lsn-deployment-02 (work
          in progress), February 2013.

[RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address
          Translator (NAT) Terminology and Considerations",
          RFC 2663, August 1999.

[RFC2993] Hain, T., "Architectural Implications of NAT", RFC 2993,
          November 2000.

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

11.2. Informative References

[I-D.chen-sunset4-cgn-port-allocation]
          Chen, G., "Analysis of NAT64 Port Allocation Method",
          draft-chen-sunset4-cgn-port-allocation-01 (work in
          progress), February 2013.

[I-D.ietf-softwire-map]
          Troan, O., Dec, W., Li, X., Bao, C., Matsushima, S., and
          T. Murakami, "Mapping of Address and Port with
          Encapsulation (MAP)", draft-ietf-softwire-map-04 (work in
          progress), February 2013.

[I-D.ietf-v6ops-464xlat]
          Mawatari, M., Kawashima, M., and C. Byrne, "464XLAT:

                      Combination of Stateful and Stateless Translation",
                      draft-ietf-v6ops-464xlat-10 (work in progress),
                      February 2013.

   [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.

   [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.

   [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.

   [RFC6346]   Bush, R., "The Address plus Port (A+P) Approach to the
               IPv4 Address Shortage", RFC 6346, August 2011.

Author's Address

   Kaname Nishizuka
   NTT Communications Corporation
   Granpark Tower
   3-4-1 Shibaura, Minato-ku
   Tokyo 108-8118
   Japan

   Email: kaname@nttv6.jp