Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 23, 2011

X. Xu
Huawei Technologies Co.,Ltd
M. Boucadair
France Telecom
Y. Lee
Comcast
G. Chen
China Mobile

October 20, 2010

Redundancy Requirements and Framework for Stateful Network Address Translators (NAT)
draft-xu-behave-stateful-nat-standby-06

Abstract

This document defines a set of requirements and a framework for ensuring redundancy for stateful Network Address Translators (NAT), including NAT44, NAT64 and NAT46.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 23, 2011.

Copyright Notice

Table of Contents

1. Introduction 3
2. Terminology and Acronyms 3
2.1. Acronyms 3
2.2. Terminology 4
3. Reference Architecture 5
4. Dynamic and Static States 6
5. Overview of Redundancy Mechanisms 6
6. Cold Standby Mode 8
6.1. Internal Realm 8
6.2. External Realm 8
6.3. NAT Reachability Announcement 9
7. Hot Standby Mode 10
7.1. Internal Realm 10
7.2. External Realm 10
7.3. NAT Reachability Announcement 10
8. IANA Considerations 11
9. Security Considerations 11
10. Acknowledgements 11
11. References 11
11.1. Normative References 11
11.2. Informative References 11
Appendix A. State Synchronization Protocol Considerations 12
Appendix B. Election Protocol Considerations 13
Authors' Addresses 14

1. Introduction

Network Address Translation (NAT) has been used as an efficient way to share the same IPv4 address among several hosts. Recently, due to IPv4 address shortage, several proposals have been elaborated to rely on Carrier Grade NAT (CGN, a.k.a.- LSN for Large Scale NAT) (e.g., [I-D.shirasaki-nat444-isp-shared-addr], [I-D.ietf-softwire-dual-stack-lite] and [I-D.ietf-behave-v6v4-xlate-stateful]). In such models, CGN function (which may be embedded in a router or be deployed in standalone devices) is deployed within large-scale networks, such as ISP networks or enterprise ones, where a large number of customers are located. These customers within a network which is served by a single CGN device may experience service degradation due to the presence of the single point of failure or loss or state information. Therefore, redundancy capabilities of the CGN devices are strongly desired in order to deliver highly available services to customers. Failure detection and repair time must be therefore shortened.

This document describes a framework of redundancy for stateful NAT including: NAT44 including DS-Lite, NAT64 and NAT46.

The main purpose of this document is to analyze means to ensure high availability in environments where carrier grade NAT44, NAT64 and NAT46 are deployed. Some engineering recommendations are provided for the selection of the IPv6 prefix to build IPv4-Embedded IPv6 addresses [I-D.ietf-behave-address-format] and the routing configuration.

Except dealing with the exceptional failures (e.g., power outage, OS crash-down or link failure, etc.), the redundancy mechanism described in this document can also be used for planned maintenance operations (i.e., graceful shutdown of the Primary NAT due to maintenance needs).

Unless otherwise mentioned, NAT and CGN terms throughout this document, pertain to stateful NAT and stateful CGN. Stateless NAT is out of the scope of this document.

2. Terminology and Acronyms

2.1. Acronyms

CGN	Carrier Grade NAT
LSN	Large Scale NAT
DS-Lite	Dual Stack Lite
AFTR	Address Family Transition Router
NAT	Network Address Translation
ISP	Internet Service Provider

2.2. Terminology

This document makes use of the terms defined in [RFC2663]. Below are provided terms specific to this document:

- CGN (Carrier Grade NAT) or LSN (Large Scale NAT): a NAT device placed within a large-scale network (e.g., ISP network, enterprise network, or campus network). CGN may be placed at the boundary between the large-scale private network and the public Internet, between a private network and a large-scale public network or between two heterogeneous IP realms (i.e., IPv4 and IPv6).

- CGN internal address realm (internal realm for short): a realm internal to the CGN.

  For NAT44, the internal realm refers to the private networks.

  For NAT64, the internal realm means IPv6 network or IPv6 Internet.

  For NAT46, the internal realm refers to IPv4 network or IPv4 Internet.

  For DS-Lite, the internal address realm is assumed to be private IPv4 addresses even if the transport mode used to convey exchanged traffic is IPv6. A DS-Lite CGN device (a.k.a., Address Family Transition Router) is a special NAT44 function which uses the IPv6 address as a means to de-multiplex users sharing the same IPv4 address [I-D.ietf-softwire-dual-stack-lite].

- The hosts located in the internal realm are called internal hosts, and the addresses used in the internal realm are called internal addresses.

- CGN external address realm (external realm for short): a realm external to the CGN.

  For NAT44, the external realm refers to the IPv4 Internet.

For NAT64, the external realm means the IPv4 Internet or IPv4 network.

For NAT46, the external realm refers to the IPv6 Internet or IPv6 network.

- The hosts located in the external realm are called external hosts, and the addresses used in the external realm are called external addresses.

- Internal address pool: an address pool used for assigning internal addresses to represent the external hosts in the internal realm. This address pool is specific to NAT46 and NAT64.

  For NAT46, the CGN will allocate one internal address (which is an IPv4 address) from the pool to an external IPv6 host and map the external IPv6 host's IPv6 address to this IPv4 address.

  For NAT64, the CGN internal address pool is the Prefix64 [I-D.ietf-behave-address-format]. Prefix64 is used for synthesizing internal IPv6 addresses to represent external IPv4 hosts in the internal realm.

- External address pool: an address pool used by the CGN for assigning external addresses to the internal hosts.

  For NAT44 and NAT64, the external address pool contains a set of public IPv4 addresses.

  For NAT46, the external IPv6 address pool is the Prefix64. Prefix64 is used by the CGN for synthesizing the external IPv6 addresses to represent internal IPv4 hosts in the external realm.

- CPE (Customer Premises Equipment): a device which is used to interconnect the customer premise with the service provider's network.

3. Reference Architecture

In a typical operational scenario, as illustrated in Figure 1, two NAT devices are deployed for redundancy purposes. This is the reference architecture for the mechanisms we describe in this memo. Note that these mechanisms are also suitable in scenarios where more than two NAT devices are used.

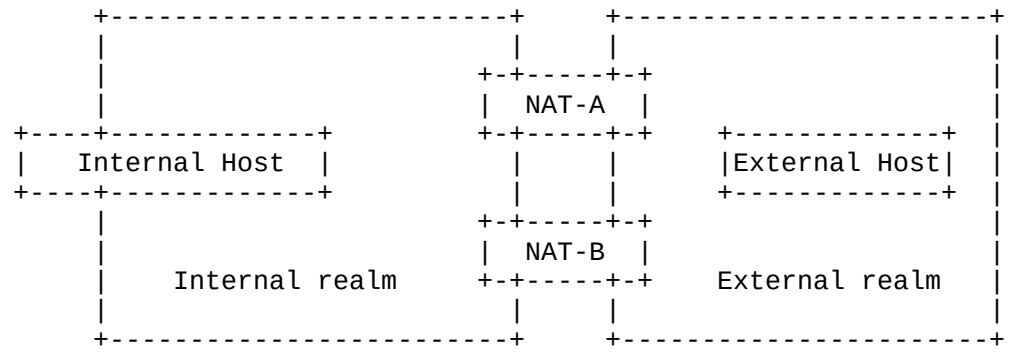

Figure 1: General Scenario of Dual NAT Routers

The redundancy mechanisms for NAT44, NAT46 and NAT64 are almost identical. In all cases, the NAT device or the immediate router of the NAT device announces the reachability of the NAT device to the external realm. The slight difference is the NAT reachability information. For example, NAT64 announces an IPv6 route for the Prefix64; NAT44 announces an IPv4 default route; DS-Lite AFTR announces an IPv6 route pointing to itself; and NAT46 announces a route for its internal address pool. This difference does not affect the general redundancy mechanisms, so the mechanisms described in this memo can be applied to NAT44, NAT64 and NAT46 devices.

4. Dynamic and Static States

The NAT states mentioned in this document only mean those NAT states which are created dynamically by outgoing packets, rather than those static NAT states which are configured manually or with automatic means such as UPnP or PCP. For those static NAT states (a.k.a., port forwarding entries), they are essentially part of the configuration data.

Port forwarding entries SHOULD be stored in permanent storage whatever the deployed redundancy mode.

5. Overview of Redundancy Mechanisms

The fundamental principle of NAT redundancy is to make two or more NAT devices function as a redundancy group, and select one as the Primary NAT and the other(s) as the Backup NAT through a dedicated election procedure or manual configuration.

In the nominal regime, traffic exchanged between one host in the

internal realm and another host in the external realm is handled by the Primary NAT. Once the Primary NAT is out of service, the Backup NAT with the highest priority (if several Backup NAT devices are deployed) takes over and provides the NAT services to the internal hosts. This Backup NAT is then identified as new Primary NAT. Once the former Primary NAT became operational, it could either preempt the role of Primary NAT or stay as a candidate in the redundancy group. This is part of administrative policies and out of scope of this memo.

In order to implement the aforementioned procedure, means to detect and to notify the failure of the Primary NAT to the redundancy group SHOULD be activated.

To ensure a coherent behavior when NAT device fails, this document assumes that both Primary and Backup NAT devices are managed by the same administrative domain. Thus, consistent configuration policies SHOULD be enforced in all devices. Note that the election process MUST be deterministic and does not lead to ambiguous situation where two or more NAT devices become Primary NAT. Moreover, the failover SHOULD be quick to ensure service continuity and keep end-users from perceiving service unavailability.

Three redundancy modes are described hereafter: the cold standby, the hot standby and the partial hot standby modes:

1. The cold standby mode is simple. The NAT states are not replicated from the Primary NAT to the Backup NAT. When the Primary NAT fails, all the existing established sessions will be flushed out. The internal hosts are required to re-establish sessions to the external hosts;

2. The hot standby mode keeps established sessions while failover happens. NAT states are replicated from the Primary NAT to the Backup NAT. When the Primary NAT fails, the Backup NAT will take over all the existing established sessions. The internal hosts are not required to re-establish sessions to the external hosts.

3. The partial hot standby mode is a flavor of the hot standby mode described above. It is used to avoid replicating NAT states of trivial sessions (e.g., short lifetime sessions) while achieving hot standby for significant sessions (e.g., critical protocols or applications, long lifetime sessions etc.). Criteria for sessions to be replicated on backup NATs SHOULD be explicitly configured on the NAT devices of a redundancy group.

The following sections provide more information about the cold standby and the hot standby modes.

6. Cold Standby Mode

6.1. Internal Realm

The internal addresses used to represent the external hosts in the internal realm SHOULD be retained after the NAT failover. The following assesses how this requirement is met in each NAT flavor:

- For NAT44 and DS-Lite, the external hosts' internal addresses (i.e., the addresses used to represent the external hosts in the internal realm) are unchanged (i.e., not NAT-ed). Therefore, the above requirement is met without additional work.

- For NAT64, the NAT devices belonging to a redundancy group SHOULD be configured with an identical Prefix64. Since the NAT64 uses stateless address translation for the external hosts, using the same Prefix64 in the Backup NAT can guarantee the internal hosts to see the same internal addresses for the external hosts.

- For NAT46, NAT devices in a redundancy group SHOULD be configured with an identical IPv4 address pool. A subset of translation state information SHOULD be synchronized among these NAT devices through a dedicated state synchronization protocol such as [I-D.xu-behave-nat-state-sync]. This translation state ensures that the Backup NAT, once taking over as a Primary NAT, will assign the same IPv4 addresses to the external IPv6 hosts for the internal IPv4 hosts.

6.2. External Realm

Each NAT device in a NAT redundancy group is configured with a different external address pool. A route to that external pool is then announced into the external realm by the NAT device or the NAT immediate router.

- For NAT44, DS-Lite and NAT64: NAT devices SHOULD be configured with different external IPv4 address pools. These address pools are not overlapped. Otherwise, when the Primary NAT fails and the Backup NAT takes over the Primary NAT, a NAT collision may happen. For example, assuming a Primary NAT NAT-ed internal host Host-A to IPv4-A. IPv4-A is an address which belongs to the external address pool. If the Backup NAT after taking over the primary NAT was configured with the same pool, the Backup NAT MAY assign the same IPv4-A to another internal host Host-B. So, Host-B may receive datagrams originally targeted for Host-A. This might cause confusion to Host-B. In addition, by using different external address pools on each NAT device, incoming datagrams of a given session from the external hosts are ensured to always

traverse through the Backup NAT device after the Primary NAT failover happens.

- For NAT46, the issue occurred in NAT44 and NAT64 cases will not happen. NAT46 relies on stateless address translation for the internal hosts. The Primary and Backup NAT SHOULD use the same external Prefix64, hence the external hosts can use the Backup NAT46 to reach the internal hosts. In Cold Standby mechanism, the Primary and Backup NAT MAY use different Prefix64s. In contrast, the Primary and Backup NAT in Hot Standby mechanism MUST use an identical Prefix64.

6.3. NAT Reachability Announcement

In order to force the IP datagrams from the internal realm to always traverse through the Primary NAT to the external realm, the Primary NAT SHOULD announce into the internal realm a route towards the external realm.

- For NAT44, the Primary NAT announces an IPv4 default route into the internal realm.

- For DS-Lite, the Primary NAT announces a host route into the internal realm.

- For NAT64, the Primary NAT announces a route for the Prefix64 into the internal realm.

- For NAT46, the Primary NAT announces a route for the internal address pool into the internal realm (If the internal address pool can be aggregated to one prefix).

The Primary NAT SHOULD attempt to withdraw its previously announced routes when it ceases the Primary role due to pre-configured conditions, e.g.- it loses the IP connectivity to the external realm.

When the Primary NAT fails and the Backup NAT takes over, datagrams from the internal hosts destined for the external realm SHOULD pass through the Backup NAT. Hence, when the Backup NAT is manually configured to switch over to become the Primary NAT, the Backup NAT (or associated router) SHOULD announce the same route into the internal realm, but the routing cost of this route MUST be set to a higher value than the route announced by the Primary NAT.

Alternatively, the Primary NAT announces several more specific routes into the internal realm while the Backup NAT announces an aggregate route. Taken the NAT46 as an example, assuming the internal address pool is 10.0.0.0/8, the Primary NAT announces two more specific

routes to 10.0.0.0/9 and 10.128.0.0/9 respectively while the Backup NAT announces an aggregate route to 10.0.0.0/8. In case the Primary NAT and the Backup NAT are automatically elected through a dedicated election process, the Backup NAT would be elected as a new Primary NAT once the old Primary one fails, so it is not necessary for the Backup NAT to make the above route announcements until it is elected as a new Primary NAT.

In order for the external hosts to traverse through the NAT to reach the internal hosts, the Primary and Backup NAT SHOULD announce a route of its own external address pool into the external realm.

7. Hot Standby Mode

7.1. Internal Realm

The procedure is identical to Section 6.1.

7.2. External Realm

To preserve the established sessions during the failover and to keep the internal addresses unchanged for the external hosts, the external addresses for the internal hosts MUST also be preserved. To preserve the external address of the internal host after NAT-ed, the NAT devices in a redundancy group MUST use an identical external address pool. In addition, they MUST assign the same external address (or address/port pair) to a given internal host.

- For NAT46, the Primary NAT and Backup NAT MUST use an identical Prefix64.

- For NAT44, DS-Lite and NAT64, the NAT devices in a redundancy group MUST use the same external address pool and the translation states on the Primary NAT device MUST be synchronized to the Backup NAT(s) in a timely fashion.

7.3. NAT Reachability Announcement

In order to force IP datagrams between the internal realm and the external realm always traverse through the Primary NAT, the Primary NAT (or its associated router) SHOULD announce into the internal realm a route towards the external realm and announce into the external realm a route towards the external address pool.

Once the connectivity to either the external realm or the internal realm is lost, the Primary NAT device itself or a third party SHOULD attempt to withdraw the above routes. If the Primary NAT and the

Backup NAT are specified manually, the Backup NAT device (or its associated router) SHOULD also announce those routes, but with higher enough cost or larger granularity, so as to prepare for the failover.

When the Primary NAT fails, the datagrams towards the external realm will pass through the Backup NAT device. In case the Primary NAT and the Backup are automatically elected through a dedicated election procedure, the Backup NAT would be elected as a new Primary NAT when the old Primary NAT device fails. Consequently, it is not necessary for the Backup NAT to make the above route announcement until it is elected as a new Primary NAT.

8. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

9. Security Considerations

TBD.

10. Acknowledgements

The author would like to thank Dan Wing and Dave Thaler for their insightful comments and reviews, and thank Dacheng Zhang and Xuewei Wang for their valuable editorial reviews.

11. References

11.1. Normative References

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

11.2. Informative References

[I-D.ietf-behave-address-format]
Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", draft-ietf-behave-address-format-10 (work in progress), August 2010.

[I-D.ietf-behave-v6v4-xlate-stateful]
Bagnulo, M., Matthews, P., and I. Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", draft-ietf-behave-v6v4-xlate-stateful-12 (work in progress), July 2010.

[I-D.ietf-softwire-dual-stack-lite]
Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", draft-ietf-softwire-dual-stack-lite-06 (work in progress), August 2010.

[I-D.shirasaki-nat444-isp-shared-addr]
Shirasaki, Y., Miyakawa, S., Nakagawa, A., Yamaguchi, J., and H. Ashida, "NAT444 addressing models", draft-shirasaki-nat444-isp-shared-addr-04 (work in progress), July 2010.

[I-D.xu-behave-nat-state-sync]
Cheng, D. and X. Xu, "NAT State Synchronization Using SCSP", draft-xu-behave-nat-state-sync-02 (work in progress), August 2010.

[RFC2334] Luciani, J., Armitage, G., Halpern, J., and N. Doraswamy, "Server Cache Synchronization Protocol (SCSP)", RFC 2334, April 1998.

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

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC5798] Nadas, S., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, March 2010.

Appendix A. State Synchronization Protocol Considerations

[I-D.xu-behave-nat-state-sync] defines a candidate solution to NAT state synchronization by using Server Cache Synchronization Protocol

(SCSP) [RFC2334]. For more information about the proposed solution, refer to [I-D.xu-behave-nat-state-sync].

[[Note: What to do with this section?]]

Appendix B. Election Protocol Considerations

[[Note: What to do with this section?]]

An election process and associated protocol(s) can be used to automatically elect one NAT device among a NAT redundancy group as the Primary NAT and the others as Backup NATs. Once the Primary NAT fails, the Backup NAT with the highest priority SHOULD take over the Primary NAT role after a short delay. The election protocol is also used to track the connectivity to the external realm and the internal realm. Once connections to the external realm or the internal realm lost, the NAT device is not qualified to be the Primary NAT and it will withdraw the route towards the external realm announced previously. In the case of hot standby, it SHOULD also withdraw the route towards the external address pool.

As an implementation example, VRRP [RFC5798] can be used as the automatic election protocol. In addition, an interface tracking mechanism can also be used to adjust the priority to influence the election results.

If two NAT devices are directly connected via an Ethernet network, VRRP can run directly on the Ethernet interfaces. Otherwise, some extra configuration or protocol changes need to be implemented. One option is to create conditions for VRRP to run among these devices. For example, to create a VPLS [RFC4761][RFC4762] instance and enable IP functions and run VRRP on those VLAN interfaces which are bound to that VPLS instance. If enabling IP on those interfaces is not supported, the following trick to realize the same goal, but at a cost of consuming two physical interfaces on each NAT router: create a VPLS instance among a set of NAT devices, and on each of them one Ethernet interface is bound to that VPLS instance, and another IP-enabled Ethernet interface is locally connected with that interface. Then VRRP can run on those IP enabled Ethernet interfaces which are all connected to that VPLS instance. Another option is to extend VRRP so that VRRP neighbors can be specified manually and VRRP messages can be exchanged directly among VRRP neighbors in unicast.

VRRP is only an implementation example of the election process. Other protocols MAY be used to manage the roles of Primary and Backup.

Authors' Addresses

Xiaohu Xu
Huawei Technologies Co.,Ltd
KuiKe Building, No.9 Xinxi Rd.,
Hai-Dian District, Beijing 100085
P.R. China

Email: xuxh@huawei.com

Mohamed Boucadair
France Telecom
Rennes
France

Email: mohamed.boucadair@orange-ftgroup.com

Yiu L. Lee
Comcast

Email: yiu_lee@cable.comcast.com
URI: http://www.comcast.com

Gang Chen
China Mobile
53A,Xibianmennei Ave.
Beijing, 100053
P.R.China

Email: phdgang@gmail.com