Network Working Group
Internet-Draft
Intended status: Experimental
Expires: April 29, 2010

G. Chen
H. Deng
B. Zhou
China Mobile
M. Xu
L. Song
Y. Cui
Tsinghua University

October 26, 2009

Reliable and Scalable NAT mechanism (RS-NAT) based on BGP for IPv4/IPv6 Transition

draft-chen-behave-rsnat-02

Status of this Memo

Copyright Notice

Abstract

For the rapid exhaustion of IPv4 address pool against the slow development of IPv6, IPv4/IPv6 co-existence/transition would be a long period. In the IPv4/IPv6 transition process, there are many NAT- like technologies active in the internet. However, the NAT boxes such as IPv4 NAT, IPv4/IPv6 NAT are so poor in their reliability and scalability, which put a severe threat on the development of IPv4/IPv6 transition. This document defines a reliable and scalable NAT (RS- NAT) mechanism to solve the problem.

Table of Contents

1. Introduction 3
2. RS-NAT Overview 4
3. RS-NAT Box 5
4. Load Balancing Mechanisms 6
   4.1. IPv6-IPv4 scenario 6
   4.2. IPv4-IPv6 scenario 7
5. Redundancy Mechanisms 9
   5.1. Address mapping Attribute 9
   5.2. Performance consideration 10
6. Security Considerations 11
7. IANA Considerations 12
8. References 13
   8.1. Normative References 13
   8.2. Informative References 13
Authors' Addresses 14

1. Introduction

For the rapid exhaustion of IPv4 address pool against the slow development of IPv6, IPv4/IPv6 co-existence/transition would be a long period. In order to facilitate the connectivity between IPv4 and IPv6 network, a NAT functionality should be deployed on the edge of different IP family network.

However most of the NAT-like functions are stateful, which maintain the state of address mapping for network translation or ALG function. The stateful boxes in the network will bring high risks on reliability and scalability when the network becomes huge. For example the box will be a single point of failure in a large-scale network. Although some advices are proposed such as NAT64 using multi-box, the static configuration and localized mapping information in each box are not able to accommodate the dynamic internet environment.

In this document, we proposed a Reliable and Scalable NAT (RS-NAT) mechanism to overcome the stateful NAT problem mentioned above, which include IPv4 NAT and IPv4/IPv6 NAT.

2. RS-NAT Overview

In the topology shown in Figure 1, the network can be divided into two parts: the User Network and Service Network. User Network is the realm where the users initiate a communication with servers. The Service Network is the realm where the remote destination (e.g., server) is attached. In addition there are some RS-NAT boxes which act as bridges between these networks.

```
             ____________               _______________
            /. .. .. .. .\  +---------+  /. .. .. .. .. .\
            |            |--|RS-NAT A|--|               |
            |            |  +---------+  |               |
            |User Network|      |       |Service Network|
            |            |  +---------+  |               |
            |. .. .. .. .|--|RS-NAT B|--|. .. .. .. .. .|
            \____________/  +---------+  \_______________/
```

Figure 1: General Topology of RS-NAT framework

The User Network and Service Network could be IPv4,IPv6 or Dual-stack. As a result, there are several communication scenarios could be deduced from the general topology using the form of IPvx-IPvy, which means users with IPvx protocol initiate connections to servers reachable with IPvy protocol. These communication scenarios are (private)IPv4-IPv4, IPv4-IPv6, IPv6-IPv4, and IPv6-IPv6. VRRP[RFC3768] is suitable for IPv4-IPv4 scenarios and there is no need to use NAT for IPv6-IPv6. So in this document we mainly focus on IPv4-IPv6/IPv6-IPv4 interconnection scenarios.

The User Network and Service Network are logical concepts, which may be composed of several ASes. For example, the User Network shown in Figure 1 may consist of several IPv4 networks belong to diffrent network providers. The User Network may connect to Service Network separately through RS-NAT boxes on which BGP [RFC4271] is performed .

Note that the User Network and Service Network are exchangable because an end user can be regarded as both initiator and server from different views.

3. RS-NAT Box

The following Sections will discuss the requirements of RS-NAT Box and its basic embedded functions.

In order to achieve the role of bridge between the two networks in the studied scenarios, which are depicted in Sections 3, the RS-NAT box is capable of dual-stack and forwording traffic based on IPv4/IPv6 protocol translator.

In the IPv6-IPv4 scenario RS-NAT router advertises different Prefix64s [I-D.miyata-behave-prefix64] routing information to User Network, and advertise the prefix info of static IPv4 address pool to Service Network. In this scenario, DNS64[I-D.bagnulo-behave-dns64] is employed to assign different Prefix64 for each DNS request randomly, which will be discussed in Section 5. In the IPv4-IPv6 scenario RS-NAT router advertises its own IPv6 prefix routing information to Service Network, and sends the prefix info of static IPv4 address pool to User Network. In this scenario, DNS ALG mentioned in NAT-PT[RFC2766] will be modified to support the separation of the data plane and control plane, which will be discussed in Section 5.

The address mapping modules in RS-NAT is useful not only for the IP head translation, but essential for some application that embed network-layer addresses as well, such as FTP, SIP etc.

4. Load Balancing Mechanisms

This Section will show how the RS-NAT run and balance the traffic among these RS-NAT boxes.

4.1. IPv6-IPv4 scenario

Figure 2 illustrates the connection setup in the IPv6-IPv4 scenario. The connection set-up follows two steps:

1) User sends DNS query to DNS64 and gets the DNS reply with an IPv4-embedded IPv6 addresses in the form of Prefix64::IPv4 address;

2) User sends the packet to the IPv4-embeded IPv6 addresses. The different IPv6 prefix will lead the packet to different RS-NAT routers, which is achieved by the RS-NAT routing function.

```
                    The Control Plane
                  ....................
                        +-----+
               .........|DNS64|
              .         +-----+
       +----+.          +------+         +------+
       |User| --------- |RS-NAT|---------|server|
       +----+\          +------+        /+------+
              \         +------+       /
               ---------|RS-NAT|-------
                        +------+
                   --------------------
                    The Data Plane
```

Figure 2: IPv6-IPv4 Connection Setup

As mentioned previously RS-NAT routers run BGP and keep BGP neighbor information with each other. Each RS-NAT router will maintain the IPv6 prefix which is identical with the prefix DNS64 stores. RS-NAT will performe a Prefix-Assignment Algorithm to decide individually which part of Prefix64 they are in charge of. The Prefix-Assignment Algorism follows the new idea that the IPv6 prefix is equally divided into several portions. And, each of them is assigned to RS-NAT routers.

For example, there is 2^8 2001::/24 in Prefix64 pool of 2001::/16 and 2 RS-NAT routers. The Assignment plan is that prefixes from 2001:0000::/24 to 2001:7f00::/24 will be assigned to the router with larger IPv4 address, and the prefixes from 2001:8000::/24 to 2001:ff00::/24 is in the charge of the other router. If there are more

RS-NAT routers, these prefixes can also easy assigned to them according to the IP address sorting.

In order to balance the traffic among these RS-NAT routers, each router should advertise the route of its aggregated Prefix64 to User Network. Note that for the redundancy consideration each router could advertise overlapped Prefix64 with low priority in case other RS-NAT routers are failed.

Note that once RS-NAT routers are failed or new RS-NAT routers are configured to join in, the routing for load balance can be automatically configured by RS-NAT routers by themselves. Prefix-Assignment Algorithm will be triggered in each RS-NAT router to re-compute the router prefix. BGP KEEPALVE and OPEN messages are used to achieve that trigger.

4.2. IPv4-IPv6 scenario

The load balancing mechanism in IPv4-IPv6 interconnection scenario is similar to the one in IPv6-IPv4 in that the IPv4 address pool should shared by RS-NAT routers and each RS-NAT router is responsible to advertise the route of their IPv4 address pool, which is similar to the routing procedure of RS-NAT routers in IPv6-IPv4. The IPv4-IPv6 connection set-up is shown in Figure 3.

```
                  The Control Plane
                 .....................
           +-----+    +--------+      +-----+
           |DNS4|....|DNS ALG|......|DNS6|
           +-----+    +---|---+      +-----+
    +-----+.           +--|---+         +-------+
    |User| ---------- |RS-NAT| --------|server|
    +-----+\           +--|---+        /+-------+
           \          +--|---+       /
            --------- |RS-NAT| ------
                       +-------+
                 --------------------
                   The Date plane
```

Figure 3: IPv4-IPv6 connection Set-up

Figure 3 illustrates the connection setup in IPv4-IPv6 scenario. The connection setup follows three steps:

1) User sends DNS query to DNS4 and the query will be redirected to DNS6 through a DNS-ALG box. Once the DNS reply reachs the DNS-ALG box, the box will pick a IPv4 address from the IPv4 address pool and form a mapping with the IPv6 address form the answer of the DNS reply. A new DNS relpy will be generated and sent to DNS4 and User.

2) Because the packet translation will be done in the RS-NAT router, the DNS ALG box should send the mapping info to RS-NAT routers using new BGP attribute which will be defined in Section 5

3) User sends the packet to the IPv4 address got from the answer of DNS reply. The different IPv4 addresses will lead the packet to different RS-NAT routers, which is achieved by the RS-NAT routing function.

Note that in step 1 the DNS-ALG box acts just as DNS-ALG functions module in NAT-PT box. The difference between the two box is that DNS-ALG box in our plan is only responsible for the control plan without packet translation. In addition DNS-ALG box should in charge of the mapping distribution among those RS-NAT routers

The differences between the two scenarios include two parts:

- o The control plane: In IPv6-IPv4 scenario, it is the DNS64 that synchronizes the IPv6 and IPv4 address for IPv4 hosts, while in IPv4-IPv6 scenario, a DNS ALG server monitors the DNS requests and replies and forms the mapping of IPv4/IPv6 address.

- o Address mapping advertisement: For the load balancing reason,DNS ALG server is not designed for traffic translation and forwarding, which are in the charge of RS-NAT routers. As a result the DNS ALG server should send the mapping info to RS-NAT routers using new BGP attribute which is defined in Section 5.

5. Redundancy Mechanisms

If there exits multi-boxes between the two edge of network, problems will arise when some boxes are not stable or failed. The problems are mainly in two aspects. The first problem is in routing aspect: when one box fails, there is no other valid routes to the destination. The second is in address mapping aspect: when one box is failed, the address mapping information in the box is lost. Furthermore, it will cause the flows broke up and reconnection.

The first problem is solved in Section 4 in which the routing mechanism makes sure that the taffic will find a way out through another RS-NAT router by setting the different route cost or preference. In this Section we will define a BGP attribute that one RS-NAT can advertise the local address mapping to other neighbors which guarantees the redundancy of mapping info. With that redundancy address mapping information RS-NAT routers are able to translate the new traffic

5.1. Address mapping Attribute

Address mapping attribute is an optional transitive attribute that is composed of a set of TLVs. The type code of the attribute is to be assigned by IANA. Each TLV contains information corresponding to a particular mapping information. The TLV is structured as follows:

```
 0                   1                   2                   3

 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|    mapping Type (2 Octets)    |        Length (2 Octets)      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                             Value                             |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

apping Type (2 octets): It identifies the type of the mapping information being transmitted. This document defines the following types:

- IPv4-IPv4: mapping Type = 1

- IPv4-IPv6/IPv6-IPv4: mapping Type = 2

- IPv6-IPv6: mapping Type=3

Unknown types are to be ignored and skipped upon receipt.

Length (2 octets): the total number of octets of the Value field.

Value (variable): The value is composed of the address mapping information. If mapping type is 2, the value contains an IPv4/IPv6 address mapping just simply structured as follows:

```
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |  IPv4 address  (4 Octets)      |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |  IPv4 prot nubmer (2 Octets)   |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |                                |
        |  IPv6 address   (16 Octets)    |
        |                                |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
        |  IPv6 prot nubmer (2 Octets)   |
        +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

5.2. Performance consideration

As the mapping information is tremendous and dynamic. The performance of RS-NAT is an important issue. BGP reflector can be utilized to reduce the BGP update massage. If reflector is deployed, new mechanism should guarantee each RS-NAT routers knowing the number of routers. In addition some optimization of RS-NAT and possible modifications of BGP will be explored in the next version of this document. For example the mapping information should be advertised in a certain refresh-time interval

Note that RS-NAT routers are located on the edge of network and they may not connect directly. BGP has its nature advantage to do signaling among edge routers over some intra-domain protocol.

6. Security Considerations

It needs to be further identified.

7. IANA Considerations

This memo includes no request to IANA.

8. References

8.1. Normative References

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC2766] Tsirtsis, G. and P. Srisuresh, "Network Address Translation - Protocol Translation (NAT-PT)", RFC 2766, February 2000.

[RFC3768] Hinden, R., "Virtual Router Redundancy Protocol (VRRP)", RFC 3768, April 2004.

[RFC4966] Aoun, C. and E. Davies, "Reasons to Move the Network Address Translator - Protocol Translator (NAT-PT) to Historic Status", RFC 4966, July 2007.

8.2. Informative References

[I-D.bagnulo-behave-nat64] Bagnulo, M., Matthews, P., and I. Beijnum, "NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", draft-bagnulo-behave-nat64-03 (work in progress), March 2009.

[I-D.miyata-behave-prefix64] Miyata, H. and M. Bagnulo, "PREFIX64 Comparison", draft-miyata-behave-prefix64-02 (work in progress), March 2009.

[I-D.bagnulo-behave-dns64] Bagnulo, M., Sullivan, A., Matthews, P., Beijnum, I., and M. Endo, "DNS64: DNS extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", draft-bagnulo-behave-dns64-02 (work in progress), March 2009.

Authors' Addresses

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

Phone: +86-13910710674
Email: phdgang@gmail.com

Hui Deng
China Mobile
53A,Xibianmennei Ave.
Beijing 100053
P.R.China

Phone: +86-13910750201
Email: denghui02@gmail.com

Bo Zhou
China Mobile
53A,Xibianmennei Ave.
Beijing 100053
P.R.China

Phone: +86-13811948723
Email: zhouboyj@chinamobile.com

Mingwei Xu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5822
Email: xmw@csnet1.cs.tsinghua.edu.cn

Linjian Song
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5822
Email: songlinjian@csnet1.cs.tsinghua.edu.cn

Yong Cui
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5822
Email: cuiyong@tsinghua.edu.cn