v6ops
Internet-Draft
Intended status: Informational
Expires: September 13, 2012

C. Xie
China Telecom
X. Li
Tsinghua University
J. Qin
Consultant
M. Chen
FreeBit

A. Durand

Juniper Networks

March 12, 2012

Practice of IPv4/IPv6 transition system for data center
draft-sunq-v6ops-contents-transition-03

Abstract

This document describes deployment practice of IPv4/IPv6 translation technologies for data center transition, aiming at rapidly increasing the amount of IPv6 accessible contents for users from IPv6 Internet while preserving the continuity of IPv4 service delivery. System based on this design has been deployed in production network to provide transition service for several ICP websites.

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 September 13, 2012.

Table of Contents

1. Introduction

Facing the pressure of IPv4 address shortage, the operators may like to provide services through IPv6 by upgrade their IP infrastructure to support IPv6. As part of the Infrastructure, Data center (in short, IDC) is the main faculty to house service system that provides services and contents. It is obvious that data center also plays an important role in IPv6 transition in accordance with the transition of IP network. Dual-stack is the basic transition strategy for most data centers, as well as IP transport network. However, in our practices, we found that dual-stack alone is not enough to meet the transition demand of ICPs(in short, ICP) in data centers. The reason behind this is that providing IPv6 services requires the service software of ICP, i.e., website system, database system, supporting system, etc., should be IPv6-aware and can deal with IPv6-related information. Upgrading the service system to support IPv6 is technological-complicated and financially costly, especially for some small and medium-sized ICPs, which is the main reason that the IPv6 transition on the ICP sides moves even more slowly than the readiness of operators' IP network. The lack of IPv6-reachable contents becomes one of the main obstacles. On the other hand, some progressive ICPs who are willing to setup an IPv6-only system also would like to offer IPv4 continuity for end-users.

Under such circumstances, we propose to deploy IDC transition system in data center, aiming at aiding CP/SP to provide IPv6 services rapidly and smoothly. Another purpose of our approach is to increase the amount of IPv6 accessible contents for users from IPv6 Internet. It can also keep the IPv4 continuity for IPv6-only contents.

This document describes our current experiences on two deployment models for the transition of data center based on the approaches specified by IETF (e.g., NAT64 [RFC6146], Dual-Stack [RFC4213], IVI[RFC6219], etc.), targeting different use cases or conditions. Based on these models, an IDC transition system was designed and developed by China Telecom to provide transition services to ICPs in data centers. Some issues and considerations were also identified from the actual deployment.

2. 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 [RFC2119].

3. Motivations

As mentioned above, IDC's transition is closely related to the IPv6 service provisioning of ICPs. There have been statements from several popular ICPs that they have turned on IPv6 (no matter by which means), which do have a beneficial effect on encouraging end users' transition to IPv6. However, given the operational cost, it is still difficult for most ICPs (especially the great many ones of small-to- medium size) to immediately make their publically-facing services accessible through both IPv4 and IPv6 natively. It will involve a lot of workload for upgrading numerous application systems and the supporting systems in ICPs. On the other hand, from the users' perspective, the IPv6 reachability of resources required for their daily lives is one of the foremost concerns when making the decision on whether or not to access Internet using IPv6. It is a chicken or egg dilemma, but the two perspectives are interdependent. If the transition of one side passes the point of inflexion, the other side will be speeded up after. So, more efforts are needed to encourage the IPv6 adoption and reach the point.

Moreover, some progressive ICPs are willing to maintain a separated IPv6-only system, which will lower the risk of the potential impact on their existing wildly used IPv4 system in the early phase. Besides, single-stack system is also easy for operation, management and troubleshooting. There are no duplicated policies need to be applied, including e.g. ACL control, accounting, authentication, etc. In this case, it is also the requirement to offer IPv4 continuity to IPv6-only contents.

Therefore, the transition system provided by operators in data centers will not only help promote ICP transition in a step-by-step way, but also break out the chicken or egg dilemma for the whole IPv6 industry.

3.1. Transition As A Service

In China Telecom, we have deployed a transition platform in our IDC network. It can be regarded as transition services offered by the operators, to small-to-medium size ICPs (e.g., those who rent servers from the operators).

The ICPs can choose to take different approaches according to their scenarios and business strategies. For the conservative ones, the IPv4 services can be still offered natively, and the IPv6 services can be offered by the stateful IPv4/IPv6 translation [RFC6146]. While for progressive ones and newly incomers, the stateless IVI [RFC6219], [RFC6052] can be employed to offer native IPv6 services reachable via IPv4.

3.2. Guiding the traffic to IPv6 network

IPv4 address shortage has driven some network providers began to run IPv6 in part or the whole network. However, even if IPv6 is ready in the IP network, most ICPs in IDC have not been ready to provide IPv6 services. As a result, almost all the traffic is still IPv4-based, which makes the IPv6 network nearly empty. With this in mind, IPv4/IPv6 translation system deployed in IDC can translate the IPv4 packets sourced from the existing servers into IPv6 packets, and forward them into IPv6 network, which is equal to move the traffic from IPv4 network to IPv6 network. and encourage the customers to use IPv6 from the beginning. Furthermore, only translation will be performed on the edge of the network and it is independent of user-side transition mechanisms.

4. Deployment practice one: Communication from IPv6 users to IPv4 server

4.1. Deployment scenario

We have deployed transition service gateway in the exit of our IDCs. It is a shared platform which can serve multiple servers simultaneously. It can be integrated with existing network element of our IDC, e.g. egress router, load balancer, etc., or can be deployed as a new standalone device. The integrated deployment scenario would have little impact on existing network topology; however, it is highly coupled with existing devices. The standalone deployment scenario would be easier to implement on existing network incrementally. However, it will result in extra cost for new devices.

The egress router of our IDC is IPv6-reachable, however, either the content servers or the whole IDC infrastructure have been upgraded to IPv6 directly. With the help of transition gateway, we can provide IPv6 reachable content to customers in a quick manner. Our deployment model is depicted in the following picture.

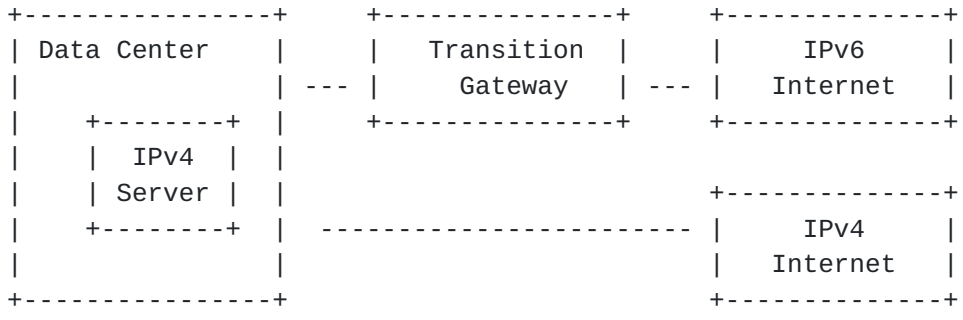

Figure 1: Deployment Model 1

In this deployment model, the Stateful NAT64 is performed to translate IPv6 packets to IPv4 and vice versa. The guidance in [RFC6146] should be followed. The communications are initiated from the IPv6 side. When an IPv6 packet arrives, a lookup of the mapping table will be carried out to get the IPv4 address used for the translation. If there is no one matched, a new entry will be created.

The server-side deployment model is independent of user-side transition When a dual-stack user gets both A and AAAA records for a remote server, it will be encouraged to reach IPv4 content via IPv6 connectivity through the only NAT64 gateway along the path. So even if there are some other CGNs deployed in the customer-side, IPv6 traffic will be forwarded in a traditional way. Therefore, there will be no double-translation problems around here.

Up to now, there are 8 sites including the official website of China Telecom have been upgrading to IPv6 with this mechanism. More than 15 thousands different IPv6 users ever accessing the above eight ICPs through the transition box totally, with 4000 to 6000 active users every day. www.voc.com.cn is the most popular one accessed by more than 4000 IPv6 users daily, and www.chinatelecom.com.cn (the official website of china telecom) has amounts of access from 1200 IPv6 users on average every day.

4.2. Mapping and Addressing

The Stateful NAT64 can support the following two mapping modes:

- 1:1, one IPv6 address is mapped to one IPv4 address (exclusively for given lifetime);

- N:1, each of the IPv4 addresses (i.e. IPv4 address pool) will be shared by multiple IPv6 users from Internet.

To save global IPv4 addresses which has become scarce resource, private blocks, for instance 10.0.0.0/8 may be used for the Stateful NAT64. This private address block can only be seen within the IDC network.

Considering the scale of traffic in the foreseeable future, the 1:1 Mapping Mode with private blocks (one IPv6 address mapped to one private IPv4 address within 10.0.0.0/8) is selected as the default mode for the Stateful NAT64. In this mode, there is only address-layer mapping and no TCP/UDP session maintenance anymore. By this mean, the efficiency of stateful operations could be improved and the

problems introduced by the address sharing could be alleviated (for example, the burden of logging will be reduced in this mode).

However, there may be conflicts if the same private space is used internally for the interconnection of servers (e.g. multiple servers for load balancing). In this case, N:1 mode with public blocks can be used. In order to reduce state management burden in N:1 stateful NAT64 gateway as well as logging system, a bulk of ports can be allocated for each subscriber. In this port-set based mapping mode, one IPv6 address will be mapped to the same IPv4 address and a given port-set.

In addition, an IPv6 prefix is used to serve the IPv4 servers in the IDC, and the route of the prefix has been advertised to the IPv6 Internet. The IPv4 address of the server can be embedded in the IPv6 prefix following the algorithm specified in [RFC6052].

4.3. DNS

To make sure the addresses of servers can be retrieved by IPv6 users before initiating sessions, the AAAA records which formed through IPv4-translated addresses have been added directly on the domain's authoritative DNS, or upgrade authoritative DNS to support DNS64. In this way, the AAAA records under one domain name could be retrieved by IPv6 users around the world.

Please note that if the authoritative DNS of given ICPs' domain names are maintained by some third-party DNS Providers but not by themselves or the operator from whom this transition service (i.e. the deployment model of Stateful NAT64 discussed herein) is purchased, the ICPs must make sure the authoritative AAAA records can be added.

4.4. Fragmentation

Basically, the processing of packets carrying fragments follows the guidance specified in [RFC6145] and [RFC6146] with exceptions that fragmented IPv4/IPv6 packets will be firstly reassembled to an integrated packet before doing packet translation and so on.

4.5. Logging

The logging is essential for tracing back specific users in stateful NAT64. In 1:1 mode, only per-user logging events need to be recorded as {IPv6 address, IPv4 address, timestamp}. For N:1 mode, in order to reduce the number of sessions need to be logged, we adopt port-set based mechanism to assign a bulk of ports to each subscriber. Therefore, one subscriber will only create one corresponding log

report, e.g. {IPv4 address, IPv6 address, port-set, timestamp}.

4.6. Geographically aware services

Since converted IPv4 address would not represent any geographical feature anymore, applications that assume such geographic information may not work as intended.

Two solutions were designed and implemented, one is to maintain the above logging information in geographic server as well, and offer an open API to ICPs to retrieve its original IPv6 address when necessary. It will have little impact on NAT64 gateway since there is no application-layer procedure. However, due to the transmission and computational latency in geographic servers, it is more suitable for ICPs to retrieve IPv6 users' source address offline. Another way is to embed user's source IPv6 address in x-forward field of user's request when it traverses NAT64 gateway. This involves application-layer process which will bring extra burden on NAT64 gateway. So only for ICPs who really need online users' source address will be offered with this additional service.

4.7. ALG issues

Since the types of applications are relatively limited due to the deployment policy, it would be easier to solve the ALG issue compared to client-side deployment. For example, Web-based ICPs might be introduced in the first stage, and so specific ALGs can be applied accordingly.

Since video traffic constitutes a great portion of the whole Internet traffic, we have implemented HTTP AGLs for video traffic in particular.

In our test for TOP100 Websites in China, there are basically three types of HTTP ALGs for video traffic.

HTTP/1.1 302 Found: This is a common way of performing a redirection. Usually, IPv4 address literals for redirected server will be embedded in Location header.

HTTP/1.1 301 Moved Permanently: This is also a redirect way indicating the requested resource has been assigned a new permanent place, and the IPv4 address literals for redirected server will also be embedded in Location header.

HTTP/1.1 200 ok: This code means the request has succeeded. However, some ICPs will still embed the IPv4 address literals to indicate the redirected server in the following communication, and

they will use a great variety of keywords. For example, www.sina.com.cn uses the keyword "CDATA[http://" followed by a list of IPv4 addresses, and v.6.cn use "watchip" as its keyword.

Since the first two types occupy the great majority of existing ALGs for HTTP-based videos traffic, we have implemented the ALG for the first two cases to synchronize an IPv4-translated address if the server of the embedded IPv4 address is located within the NAT64 region.

4.8. High Availability

In general, there are two mechanisms to achieve high reliability, i.e. cold-standby and hot-standby. In cold-standby mode, the NAT64 states are not replicated from the Primary NAT64 gateway to the Backup NAT64 gateway. When the Primary NAT64 gateway fails, all the existing established sessions will be flushed out. The hosts are required to re-establish sessions with the external hosts. Another high availability option is the hot standby mode. In this mode the NAT64 gateway keeps established sessions while failover happens. The 1:1 mapping mode will greatly reduce the amount of sessions needed to be replicated on-the-fly from the Primary NAT64 gateway to the Backup gateway. Another option is to deploy an Anycast NAT64 prefix. This is similar to cold-standby that NAT64 states are not replicated between Primary gateway and Backup gateway, except that the heartbeat line is not needed anymore.

4.9. Security

The security issues and considerations discussed in [RFC6146] apply to the deployment model described in this document. However, when deploying stateful NAT64 in server side, it is hard to apply source-based filtering policy. As a result, we have introduced alarming mechanism to report the current status of state-consuming speed in NAT64 gateway.

Besides, both 1:1 mapping mode and port-set based N:1 mapping mode can guarantee that one IPv6 source address will be mapped to a single IPv4 address. Therefore, the ICP can identify a single subscriber either by IPv4 source address in 1:1 mapping, or IPv4 source address plus port-set in N:1 mapping.

4.10. Deployment practices

Up to now, there are 8 sites including the official website of China Telecom have been upgrading to IPv6 with this mechanism. More than15 thousands different IPv6 users ever accessing the above eight Content Providers through the transition box totally, with 4000 to 6000

active users every day. www.voc.com.cn is the most popular one accessed by more than 4000 IPv6 users daily, and www.chinatelecom.com.cn (the official website of china telecom) has amounts of access from 1200 IPv6 users on average every day.

5. Deployment practice two: communications from IPv4 users to IPv6 server

5.1. Deployment scenario

Considering in the foreseeable future, IPv6 will be a widely accepted protocol in the Internet, some ICPs, especially newcomers, will setup IPv6-only servers, to reduce the operation and maintenance complexity. When the server in question itself is IPv6-capable,communications initiated from IPv6 users will not encounter any transition problem. What we are concerned is the communications initiated from IPv4 users. To mitigate this problem, IPv4/IPv6 translation is utilized in the IDC that the server resides. In this scenario, the IPv4 node will firstly get A/AAAA records of the server from DNS, and then the communication will follow the path to NAT64 Gateway. When an IPv4 packet arrives at NAT64 Gateway, it would be translated to an IPv6 packet based on stateless 1:1 mapping algorithm [RFC6219].

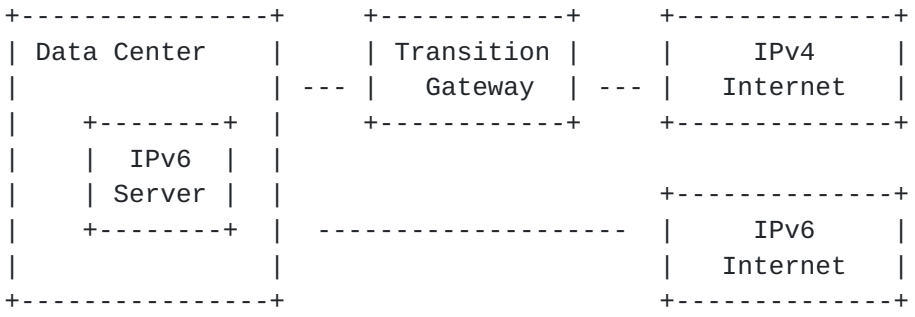

Figure 2: Deployment Model2

5.2. Mapping and Addressing

To eliminate the state management burden, we adopted stateless transition gateway to do the Interworking between IPv4 Internet and IPv6-only server within IDC, IPv6-only server should be configured with an IPv4-translatable address. Then both source address and destination address are applied with 1:1 mapping to keep the simplicity and transparency.

In addition, an IPv4 address within the range of a given IPv4 prefix is used to represent the IPv6 server, and the route of the IPv4

prefix has been advertised to the IPv4 Internet. An IPv6 prefix will be assigned to the IDC to represent the whole IPv4 Internet, when IPv4 packet traverse the transition gateway, IPv6 addresses, e.g., source address and destination address, will be formed by combine the IPv4 address with a IPv6 prefix following the algorithm specified in [RFC6052]. In this way, the server can be reachable from IPv4 Internet without mapping states in transition gateway.

5.3. DNS

To make sure that addresses of servers can be retrieved by IPv4 users before initiating sessions, the A records which are extracted from IPv4-translated addresses should be added directly on the domain's authoritative DNS, or upgrade authoritative DNS to support DNS64. Other considerations are actually the same with Section 4.

5.4. Logging

There is no logging issue in stateless transition solution.

5.5. Geographically aware services

When a ICP gets an IPv4-converted IPv6 addresses with a pre-defined Prefix, it should extract the embedded IPv4 address which would reflects its original geographical information.

5.6. ALG issues

ALG issues would be the same with section 4.6.

5.7. High Availability

Since there is no state maintained in the transition gateway, state replication or re-establishment encountered in the HA of the first deployment model will not exist in the second one.

5.8. Security

IPv4/IPv6 translators which can be modeled as special routers, are subject to the same risks, and can implement the same mitigations. (The discussion of generic threats to routers and their mitigations is beyond the scope of this document.) There is, however, a particular risk that often happens in IPv4 Internet: address spoofing.

An attacker could use a faked IPv4 address as the source address of malicious packets. After translation, the packets will appear as IPv6 packets from the specified source, and the attacker may be hard

to track. If left without mitigation, the attack would allow malicious IPv4 nodes to spoof arbitrary IPv4 addresses.

The mitigation is to implement reverse path checks and to verify throughout the network that packets are coming from an authorized location.

5.9. Deployment practices

The following IPv6-only websites has been setup to provide native IPV6 service to IPv6 users, all of them are hosted in a dual-stack IDC.

http://iptv.bupt.edu.cn

http://www.mayan.cn

http://www.ivi.buptnet.edu.cn

In order to accommodate the access of great volume of existing IPv4-only users, stateless transition gateway was deployed to provide translation in the exit of the IDC. Currently, the peak of the traffic is around 900Mbps.

6. Additional Author List

Qiong Sun

China Telecom

Room 708 No.118, Xizhimenneidajie

Beijing, 100035

P.R.China

Phone: +86 10 5855 2923

Email: sunqiong@ctbri.com.cn

Qian Liu

China Telecom

No.359 Wuyi Rd.,

Changsha, Hunan 410011

P.R.China

Phone: +86 731 8226 0127

Email: 18973133999@189.cn

Qin Zhao

BUPT

Beijing 100876

P.R.China

Phone: +86 138 1127 1524

Email: zhaoqin@bupt.edu.cn

7. IANA Considerations

This document includes no request to IANA.

8. Acknowledgements

The authors would like to thank Fred Baker, Joel Jaeggli, Erik Kline, Randy Bush for their comments and feedback.

9. References

9.1. Normative References

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

[RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.

[RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.

[RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, April 2011.

[RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation

Algorithm", RFC 6145, April 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.

[RFC6154] Leiba, B. and J. Nicolson, "IMAP LIST Extension for Special-Use Mailboxes", RFC 6154, March 2011.

[RFC6219] Li, X., Bao, C., Chen, M., Zhang, H., and J. Wu, "The China Education and Research Network (CERNET) IVI Translation Design and Deployment for the IPv4/IPv6 Coexistence and Transition", RFC 6219, May 2011.

9.2. Informative References

[I-D.wing-behave-http-ip-address-literals]
Wing, D., "Coping with IP Address Literals in HTTP URIs with IPv6/IPv4 Translators", draft-wing-behave-http-ip-address-literals-02 (work in progress), March 2010.

[RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.

Authors' Addresses

Chongfeng Xie
China Telecom
Room 708 No.118, Xizhimenneidajie
Beijing, 100035
P.R.China

Phone: +86 10 5855 2116
Email: xiechf@ctbri.com.cn

Xing Li
Tsinghua University
Room 225, Main Building
Beijing 100084
P.R.China

Phone: +86 10 6278 5983
Email: xing@cernet.edu.cn

Jacni Qin
Consultant
Shanghai,
China

Phone: +86 1391 861 9913
Email: jacniq@gmail.com

Maoke Chen
FreeBit Co., Ltd.
13F E-space Tower, Maruyama-cho 3-6
Shibuya-ku, Tokyo 150-0044
Japan

Email: fibrib@gmail.com

Alain Durand
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089-1206
USA

EMail: adurand@juniper.net