Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2012

X. Li
C. Bao
CERNET Center/Tsinghua University
W. Dec
R. Asati
Cisco Systems
C. Xie
Q. Sun
China Telecom

October 31, 2011

dIVI-pd: Dual-Stateless IPv4/IPv6 Translation with Prefix Delegation
draft-xli-softwire-divi-pd-01

Abstract

This document presents the address specifications and deployment considerations of address-sharing dual stateless IPv4/IPv6 translation with prefix delegation (dIVI-pd). The dIVI-pd keeps the features of stateless, end-to-end address transparency and bidirectional-initiated communications of the original stateless IPv4/IPv6 translation, while it can utilize the IPv4 addresses more effectively. In addition, it does not require the DNS64 and ALG, and can be used with prefix delegation.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The experiences for the IPv6 deployment in the past 10 years strongly indicate that for a successful transition, the communication between IPv4 and IPv6 address families should be supported.

Recently, the stateless and stateful IPv4/IPv6 translation methods are developed and became the IETF standards. The original stateless IPv4/IPv6 translation (stateless 1:1 IVI) is scalable, maintains the end-to-end address transparency and support both IPv6 initiated and IPv4 initiated communications [RFC6052], [RFC6144], [RFC6145], [RFC6147], [RFC6219]. But it can not use the IPv4 addresses effectively. The stateful IPv4/IPv6 translation can share the IPv4 addresses among IPv6 hosts, but it only supports IPv6 initiated communication [RFC6052], [RFC6144], [RFC6145], [RFC6146], [RFC6147]. In addition, both stateless and stateful IPv4/IPv6 translation technologies require the application layer gateway (ALG) for the applications which embed IP address literals. Furthermore, in ADSL and 3G environment, it requires the prefix delegation (assigning an IPv6 /64 or shorter) to the customer router/L3-device rather than assigning a single IPv4-translatable address to the customer device defined in [RFC6052].

In this document, we present address specifications and deployment considerations for address-sharing dual stateless IPv4/IPv6 translation with prefix delegation (dIVI-pd), which is based on basic dIVI model [I-D.xli-behave-divi] with the support of prefix delegation. The dIVI-pd can solve the IPv4 address sharing, the ALG and prefix delegation problems mentioned above, though still keeps the stateless, end-to-end address transparency and supporting of both IPv6 initiated and IPv4 initiated communications.

Due to the introduction of the second translation and the prefix delegation, the dIVI-PD is 4-6-4 model and there is a strong correlation to the stateless encapsulation approach [I-D.murakami-softwire-4rd]. This document uses the address format, the port mapping algorithm and DHCP options defined in [I-D.mdt-softwire-mapping-address-and-port]. [I-D.mdt-softwire-map-dhcp-option], which are the joint design works of stateless encapsulation and dual stateless translation.

2. Applicability

The address-sharing dual stateless IPv4/IPv6 translation with prefix delegation (dIVI-pd) can be used in ADSL or 3G environment when prefix delegation is required. An ADSL example is shown in the following figure.

```
        ----             -----
      //    \\         //     \\         -----
     /        \       /         \      //      \|------[CPE.1]--{H.1}
    |         +-----+            +----+         |
    |         |     |   Metro    |BRAS|         |------[CPE.2]--{H.2}
    | Backbone|Core |   Area     |/SR | Access  |
    | Network |Route|  Network   +----+ Network |------[CPE.k]--{H.k}
    |         |     |            |AAA |         |
    |         +-----+            +----+         |------[CPE.n]--{H.n}
     \        /       \          /      \\     /|
      \\    //         \\      //         ----
        ----            ----|--                         |
                            |                           |
    IPv4/IPv6           [XLAT1]       IPv6-only     [XLAT2.x] IPv4/IPv6
    Internet                |         Network           |      Hosts
                            |                           |
    Data path:              |                           |
    IPv6 -------------------------------IPv6-----------|------IPv6----
    IPv4 ------------------|------------IPv6-----------|------IPv4----
                            |                           |
    Address assignment:     |                           |
    IPv6                    |  [BRAS]->(DHCPv6 PD)->[CPE]->SLAAC->[Host]
    IPv4                    |                           | \-DHCP-/
```

Figure 1: BRAS

Where the ISP's backbone network is dual stack, as well as part of the metro-area network. The core IPv4/IPv6 translator (XLAT1) is performing the IPv4 address-sharing stateless IPv4/IPv6 translation and connects the dual-stack part and the IPv6-only part of the metro-area networks. The access network is IPv6-only and multiple IPv4/IPv6 translators (XLAT2.x) are connected to the access network and provide dual-stack access to the customer devices. Each dual-stack customer get a whole IPv6 /64 (or shorter) and a fractional public IPv4 address.

The data path of this user case are: The IPv6 packets from customer devices and the IPv6 Internet are not translated, while the IPv4 packets from customer devices and the IPv4 Internet are translated twice via stateless IPv4/IPv6 translation technology. Due to the stateless nature, the dual stateless IPv4/IPv6 translation is almost equivalent to tunneling with header compression.

There are two address assignment processes: (1) From BRAS to CPE is via IPv6CP and DHCPv6 prefix delegation; (2) From CPE to customer device, the IPv6 is via SLAAC and the IPv4 is via DHCP. Note that if more than one customer device requires IPv4 addresses, a built-in NAT44 in each CPE can be used to translate a fractional IPv4 address

to several [RFC1918] defined IPv4 addresses.

3. Terminologies

This document uses the terminologies defined in [I-D.mdt-softwire-mapping-address-and-port].

This document uses the terminologies defined in [RFC6144].

Since [I-D.mdt-softwire-mapping-address-and-port] is used for both encapsulation and stateless translation, the equivalent terminologies in [RFC6144] are:

MAP Border Relay (BR) Address: The MAP Border Relay (BR) Address is the IPv4-converted address defined in [RFC6144] and in [RFC6052].

MAP Customer Edge (CE) Address: The MAP Customer Edge (CE) Address is the IPv4-translatable address defined in [RFC6144] and in [RFC6052].

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

4. Port Mapping Algorithm and Address Format

The port mapping algorithm and address format are defined in [I-D.mdt-softwire-mapping-address-and-port].

4.1. Port Mapping Algorithm

Port mapping algorithm is defined in Section 4.1 of [I-D.mdt-softwire-mapping-address-and-port].

For given sharing ratio (R) and the maximum number of continue ports (M), the generalized modulus algorithm is defined as

1. The port number (P) of a given PSID (K) is composed of

   P = R*M*j + M*K + i

   Where
   o PSID: K=0 to R-1
   o Port range index: j = (1024/M)/R to ((65536/M)/R)-1, if the well-known port numbers (0-1023) are excluded.
   o Port continue index: i=0 to M-1

2. The PSID (K) of a given port number (P) is determined by

   K = (floor(P/M)) % R

   Where
   o % is modular operator
   o floor(arg) is a function returns the largest integer not
   greater than arg

4.2. Basic Mapping Rule (BMR)

Basic mapping rule is used for IPv4 prefix, address or port set assignment.

```
 |     n bits          |  o bits    | m bits  |   128-n-o-m bits      |
 +--------------------+-----------+---------+------------+----------+
 | Domain IPv6 prefix |  EA bits  |subnet ID|     interface ID      |
 +--------------------+-----------+---------+----------------------+
 |<---  End-user IPv6 prefix  --->|
```

Figure 2: IPv6 address format

The Embedded Address bits (EA bits) are unique per end user within a Domain IPv6 prefix. The EA bits encode the CE specific IPv4 address and port information. The EA bits can contain a full or part of an IPv4 prefix or address, and in the shared IPv4 address case contains a Port Set Identifier (PSID).

```
                |<------------- EA bits----------->|
  |   r bits    |        p bits        |   q bits   |
  +-------------+---------------------+------------+
  | Domain IPv4 | IPv4 Address suffix |Port Set ID |
  +-------------+---------------------+------------+
  |            32 bits                |
```

Figure 3: Shared IPv4 address

The interface ID is defined as

```
<-8-><--------- L>=32 -------><48-L><8->
+---+----------------+------+-----+---+
| u |  IPv4 address  | PSID |  0  | L |
+---+----------------+------+-----+---+
```

Figure 4: Interface ID

Forwarding Mapping Rule (FMR) is used for forwarding, whcih has similar address format to BMR. Specifying forwarding mapping rule determines the prefix of the IPv4-translatable addresses for other CEs, which results in different routing behaviors (Hubs and Spokes or Mesh).

4.3. Default Mapping Rule (DMR)

The Defualt mapping rule defines an IPv6 prefix (BR's IPv6 prefix). The full destination IPv4 address must be encoded in the IPv6 address.

4.4. Address Specifications

Based on the above discussion, the addresses are defined in the following figure.

```
            Source address from a CE to any destination
                   (IPv4-translatable address)

   <--------- 64 ------------><8 ><-------- L>=32 -----<>-44-L-<>8 <
  +-------------+--------+---+---+----------------+----+-----+-+---+
  |Domain prefix|EA bits | 0 | u |  IPv4 address  |PSID|  0    | L |
  +-------------+--------+---+---+----------------+----+-----+-+---+


      Destination address from a CE to the outside IPv4 Internet
                     (IPv4-converted address)

   <--------- 64 ------------>< 8 ><----   32 ----><----- 24 ----- >
  +-------------------------+----+---------------+----------------+
  |          BR prefix       | u  |  IPv4 address |        0       |
  +-------------------------+----+---------------+----------------+
```

Figure 5: Extended IPv4-translatable address format

5. Header Translation and MTU Handling

The general header and ICMP translation specifications are defined in [RFC6145].

Special MTU and fragmentation actions must be taken in the case of dual translation.

6. Dual Stateless Translation

When dual stateless IPv4/IPv6 translation is deployed, its behavior is similar to tunneling. Tunneling do not require DNS64 and ALG., because the communication occurs in same address family. Dual translation don't need DNS64 and ALG as well, even in each translator the communication occurs between different address families. However, there are following differences:

- Scalability. Dual stateless translation is based on routing, there is nothing needed to maintain in the translator, operator's management loads are minimum compared with tunneling scheme, which has to maintain tunnel states.

- Low OPEX. Dual stateless translation can do traffic engineering and flow analysis without decapsulation which is a must in tunnel case.

- Header Compression. The dual stateless IPv4/IPv6 translation does not need to do encapsulation and 12 octets header overhead are reduced.

- Transparent transition to IPv6. The dual stateless translation can be treated as a special case of single stateless translation, the first XLAT performances exactly the same function, no matter there is a XLAT.x or not. Hence it is a unified approach, rather than special setup for the coexistence and transition. This is to say that the ISP can deploy IPv6-only network with XLAT, so the IPv6-only hosts can communicate with both the IPv6 Internet and the IPv4 Internet. However, if for some reason a specific ALG cannot be supported, and for users, who need that specific application, can deploy XLAT.x. When the application is updated, the XLAT.x can be removed. There is nothing to change to XLAT. with more and more contents and users move to IPv6, the working load of XLAT will be less and less, and eventually can be removed. The whole process is transparent, smooth and incremental.

Due to the differences between the IPv4 header and the IPv6 header, the dual stateless IPv4/IPv6 translation cannot be entirely lossless [RFC6145], for example the IPv4 options are lost. The experimental data shows that the IPv4 packets which contain options are very few (10e-6) and causes no harm. Another corner case is the fragmetation handling. For IPv4 packets with DF=1 and MF=1, the dual stateless translation will results in DF=0. The experimental data shows that the IPv4 packets with DF=1 and MF=1 are very few (10e-5) and causes no harm.

Note that for dual stateless translation, the encapsulation (from

IPv4 to IPv6) and decapsulation (from IPv6 to IPv4) defined by [RFC2473] can be implemented in the translators. In this case, the dual stateless translation processes are entirely lossless, it still has the operation and management conveniences of the dual stateless translation in layer 3, but the control in layer 4 is lost.

7. Deployment Considerations

Given:

1. The total number of CEs in this domain.

2. The sharing ratio R.

3. The port continue parameter M.

4. The customer prefix length.

5. The ISP's IPv6 prefix.

6. The ISP's IPv4 prefix.

7. The BR IPv6 prefix.

Other dIVI-PD configuration parameters can be derived using the port mapping algorithm and address format defined in this document.

8. CE Configuration via DHCP Option

Based on the address format and the port mapping algorithm defined in this document, the CE needs to get the corresponding parameters via DHCPv6 [RFC3315][RFC3633] or others signaling scheme. These parameters are:

1. The IPv6 prefix

2. The IPv6 prefix length

3. The IPv4 prefix

4. The IPv4 prefix length

5. The sharing ratio (R)

6. The maximum number of continue ports (M)

7. The PSID (K)

8. The PSID length (c)

9. Experimental Evaluation

The basic stateless IPv4/IPv6 translation (IVI) has been deployed since 2007. It connects [CERNET] and [CNGI-CERNET2].

The dual stateless translation with IPv4 address sharing (dIVI) has been deployed in [CERNET] and [CNGI-CERNET2] since 2009. The design and implementation results are presented in [I-D.xli-behave-divi].

The dIVI has also been tested in China Telecom. The [I-D.sunq-v6ops-ivi-sp] summarizes the testing results.

The dIVI-pd presented in this document has been running in [CERNET] and [CNGI-CERNET2] since Jan. 2011. The experimental results indicate that the CPE index coding, the suffix coding and port-set ID mapping algorithm work for existing applications without any problem.

10. Security Considerations

See security considerations presented in [RFC6052] and [RFC6145].

11. IANA Considerations

This memo adds no new IANA considerations.

Note to RFC Editor: This section will have served its purpose if it correctly tells IANA that no new assignments or registries are required, or if those assignments or registries are created during the RFC publication process. From the author's perspective, it may therefore be removed upon publication as an RFC at the RFC Editor's discretion.

12. Acknowledgments

The authors would like to acknowledge the following contributors in the different phases of the address-sharing IVI and dIVI development: Hong Zhang, Yu Zhai, Wentao Shang, Weifeng Jiang, Bizhen Fu, Guoliang Han and Weicai Wang.

The authors would like to acknowledge the following contributors who

provided helpful inputs: Heyu Wang, Lu Yan, Dan Wing, Fred Baker, Dave Thaler, Randy Bush, Kevin Yin and Bobby Li.

The authors would like to thank the MAP team for the technical discussions, which make the continue improvements of dIVI-PD.

13. References

13.1. Normative References

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

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

[RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, December 1998.

[RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.

[RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.

[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.

[RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 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.

[RFC6296] Wasserman, M. and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, June 2011.

13.2. Informative References

[CERNET] "CERNET Homepage: http://www.edu.cn/english_1369/index.shtml".

[CNGI-CERNET2] "CNGI-CERNET2 Homepage: http://www.cernet2.edu.cn/index_en.htm".

[I-D.bcx-behave-address-fmt-extension] Bao, C. and X. Li, "Extended IPv6 Addressing for Encoding Port Range", draft-bcx-behave-address-fmt-extension-01 (work in progress), October 2011.

[I-D.mdt-softwire-map-dhcp-option] Mrugalski, T., Boucadair, M., and O. Troan, "DHCPv6 Options for Mapping of Address and Port", draft-mdt-softwire-map-dhcp-option-00 (work in progress), October 2011.

[I-D.mdt-softwire-mapping-address-and-port] Troan, O., "Mapping of Address and Port (MAP)", draft-mdt-softwire-mapping-address-and-port-00 (work in progress), October 2011.

[I-D.murakami-softwire-4rd] Murakami, T., Troan, O., and S. Matsushima, "IPv4 Residual Deployment on IPv6 infrastructure - protocol specification", draft-murakami-softwire-4rd-01 (work in progress), September 2011.

[I-D.sunq-v6ops-ivi-sp] Sun, Q., Xie, C., Li, X., Bao, C., and M. Feng, "Considerations for Stateless Translation (IVI/dIVI) in Large SP Network", draft-sunq-v6ops-ivi-sp-02 (work in progress), March 2011.

[I-D.xli-behave-divi] Bao, C., Li, X., Zhai, Y., and W. Shang, "dIVI: Dual-Stateless IPv4/IPv6 Translation", draft-xli-behave-divi-04

(work in progress), October 2011.

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

Authors' Addresses

Xing Li
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing 100084
CN

Phone: +86 10-62785983 begin_of_the_skype_highlighting +86
10-62785983 end_of_the_skype_highlighting
Email: xing@cernet.edu.cn

Congxiao Bao
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing 100084
CN

Phone: +86 10-62785983 begin_of_the_skype_highlighting +86
10-62785983 end_of_the_skype_highlighting
Email: congxiao@cernet.edu.cn

Wojciech Dec
Cisco Systems
Haarlerberdweg 13-19
Amsterdam 1101 CH
NL

Email: wdec@cisco.com

Rajiv Asati
Cisco Systems
7025-6 Kit Creek Road
Research Triangle Park NC 27709
USA

Email: rajiva@cisco.com

Chongfeng Xie
China Telecom
Room 708, No.118, Xizhimennei Street
Beijing 100035
CN

Phone: +86-10-58552116 begin_of_the_skype_highlighting
+86-10-58552116 end_of_the_skype_highlighting
Email: xiechf@ctbri.com.cn

Qiong Sun
China Telecom
Room 708, No.118, Xizhimennei Street
Beijing 100035
CN

Phone: +86-10-58552936 begin_of_the_skype_highlighting
+86-10-58552936 end_of_the_skype_highlighting
Email: sunqiong@ctbri.com.cn