

Softwire
Internet-Draft
Intended status: Standards Track
Expires: November 17, 2013

L. Cai
ZTE
J. Qin
S. Tsuchiya, Ed.
Cisco Systems
May 16, 2013

Definitions of Managed Objects for 6rd
draft-cai-softwire-6rd-mib-05

Abstract

This document defines a portion of the Management Information Base (MIB) for use with network management protocols. In particular, it defines objects for managing 6rd devices.

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 November 17, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. The Internet-Standard Management Framework	2
3. Conventions	2
4. Structure of the MIB Module	3
4.1. sixRdTable	3
4.2. sixRdBrdIpv4AddressTable	3
4.3. sixRdSecurityCeck	3
5. Relationship to Other MIB Modules	3
5.1. Relationship to the SNMPv2-MIB	3
5.2. Relationship to the IP Tunnel MIB	3
5.3. Relationship to the Interfaces MIB	4
5.4. Relationship to the IP MIB	4
5.5. MIB modules required for IMPORTS	4
6. Definitions	4
7. Security Considerations	7
8. IANA Considerations	8
9. References	8
9.1. Normative References	8
9.2. Informative References	9
Authors' Addresses	9

1. Introduction

This draft describes the Management Information Base (MIB) module for 6rd (IPv6 Rapid Deployment, [RFC5969]), which specifies an automatic tunneling mechanism to deploy IPv6 to sites via a operator's IPv4 network.

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC 3410 [RFC3410].

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIV2, which is described in STD 58, RFC 2578 [RFC2578], STD 58, RFC 2579 [RFC2579] and STD 58, RFC 2580 [RFC2580].

3. Conventions

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].

4. Structure of the MIB Module

The MIB Module specified herein provides one way to manage the 6rd devices through SNMP.

4.1. sixRdTable

This table contains the configuration information for 6rd.

4.2. sixRdBriPv4AddressTable

This table contains the BR IPv4 Address for configurations on given 6rd CE device.

4.3. sixRdSecurityCeck

This table contains counter of packets drop by 6rd receiving rule.

5. Relationship to Other MIB Modules

5.1. Relationship to the SNMPv2-MIB

The 'system' group in the SNMPv2-MIB [RFC3418] is defined as being mandatory for all systems, and the objects apply to the entity as a whole. The 'system' group provides identification of the management entity and certain other system-wide data. The SAMPLE-MIB does not duplicate those objects.

5.2. Relationship to the IP Tunnel MIB

The IP Tunnel MIB [RFC4087] contains objects common to all IP tunnels, including 6rd. Additionally, tunnel encapsulation specific MIB (like what is defined in this document) extend the IP tunnel MIB to further describe encapsulation specific information, for example (in case of 6rd): 6rd prefix, 6rd Prefix Length, IPv4Mask Length and BR IPv4 Address.

The implementation of the IP Tunnel MIB is required for 6rd. The tunnelIfEncapsMethod in the tunnelIfEntry should be set to sixRd("xx"), and an entry in the 6rd MIB module will exist for every tunnelIfEntry with this tunnelIfEncapsMethod. The tunnelIfRemoteAddress must be set to 0.0.0.0.

[Ed.Note:] This is similar to the situation of L2TP MIB [RFC3371] case, since the IANA is requested to assign a value for sixRdMIB under the "transmission" subtree. Also, a new IANAtunnelType (rather than IANAifType) value is needed and should be recorded in the IANAifType-MIB registry, refer to Section 8.

5.3. Relationship to the Interfaces MIB

Each logical interface (physical or virtual) has an ifEntry in the Interfaces MIB [RFC2863]. Tunnels are handled by creating a logical interface (ifEntry) for each tunnel.

5.4. Relationship to the IP MIB

IP MIB [RFC4293] provides traffic statistics counter and status for 6rd virtual interface.

5.5. MIB modules required for IMPORTS

This MIB module IMPORTs objects from [RFC4087], [RFC2580], [RFC2578], [RFC2863], [RFC3411].

6. Definitions

SIXRD-MIB DEFINITIONS ::= BEGIN

IMPORTS

OBJECT-TYPE, transmission, Integer32
FROM SNMPv2-SMI

ifIndex
FROM IF-MIB

InetAddressIPv4, InetAddressPrefixLength, InetAddressIPv6
FROM INET-ADDRESS-MIB;

sixRdMIB MODULE-IDENTITY

LAST-UPDATED "201208120000Z" -- August 12, 2012

ORGANIZATION "IETF Softwire Working Group"

CONTACT-INFO

"Lei Cai
ZTE
No. 68 Zijinhua Rd.,
Nanjing, 210012
China
Email: cai.lei3@zte.com.cn

Jacni Qin

Cisco Systems
Shanghai,
China
Email: jacni@jacni.com

Shishio Tsuchiya
Cisco Systems
Midtown Tower, 9-7-1, Akasaka
Minato-Ku, Tokyo 107-6227
Japan
Email: shtsuchi@cisco.com"

DESCRIPTION

"The MIB module defines managed objects for 6rd."

:: = { transmission XX } ---xx to be replaced

sixRdDevice OBJECT-TYPE

SYNTAX Integer32 (0..1)

MAX-ACCESS read-write

STATUS current

DESCRIPTION

"A value of 1 indicates the device is a 6rd BR,
or 0 indicates the device is a 6rd CE."

::= { sixRdMIB 1 }

sixRdTable OBJECT-TYPE

SYNTAX SEQUENCE OF SixRdEntry

MAX-ACCESS not-accessible

STATUS current

DESCRIPTION

"The table contains the configuration information
of 6rd on a particular tunnel."

::= { sixRdMIB 2 }

sixRdEntry OBJECT-TYPE

SYNTAX SixRdEntry

MAX-ACCESS not-accessible

STATUS current

DESCRIPTION

"An entry containing the configuration
information of 6rd on a particular tunnel."

INDEX {ifIndex}

::= { sixRdTable 1 }

SixRdEntry ::= SEQUENCE {

sixRdPrefix InetAddressIPv6,

sixRdPrefixLen InetAddressPrefixLength,

```
        sixRdIpv4MaskLen      Integer32
    }

sixRdPrefix OBJECT-TYPE
    SYNTAX      InetAddressIPv6
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "The 6rd prefix of this 6rd domain."
    ::= { sixRdEntry 1 }

sixRdPrefixLen OBJECT-TYPE
    SYNTAX      InetAddressPrefixLength
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "The length of 6rd prefix."
    ::= { sixRdEntry 2 }

sixRdIpv4MaskLen OBJECT-TYPE
    SYNTAX      Integer32 (0..32)
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "The number of high-order bits that are
         identical across all CE IPv4 addresses within
         this 6rd domain."
    ::= { sixRdEntry 3 }

sixRdBrIpv4AddressTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF SixRdBrIpv4AddressEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "The table contains the BR IPv4 Address of given
         6rd domain if the value of 6rdDevice is 0 (i.e.,
         6rd CE), or should be omitted if the value of
         6rdDevice is 1 (i.e., 6rd BR)."
    ::= { sixRdMIB 3 }

sixRdBrIpv4AddressEntry OBJECT-TYPE
    SYNTAX      SixRdBrIpv4AddressEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "An entry containing the BR IPv4 Address of given
         6rd domain."
    INDEX      {ifIndex,
```

```

        sixRdBriPv4Address
    }
    ::= { sixRdBriPv4AddressTable 1 }

SixRdBriPv4AddressEntry ::= SEQUENCE {
    sixRdBriPv4Address          InetAddressIPv4
}

sixRdBriPv4Address OBJECT-TYPE
    SYNTAX      InetAddressIPv4
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "The BR IPv4 Address of this 6rd domain."
    ::= { sixRdBriPv4AddressEntry 1 }

sixRdSecurityCeck OBJECT-TYPE
    SYNTAX      SEQUENCE OF sixRdSecurityCeckInvalidPackets
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "This table contains counter of packets drop by 6rd
        receiving rule."
    ::= { sixRdMIB 4 }

sixRdSecurityCeckInvalidPackets OBJECT-TYPE
    SYNTAX      Counter64
    MAX-ACCESS  read-only
    STATUS      current
    DESCRIPTION
        "6rd BR/CE MUST validate the embedded IPv4 source
        address of the encapsulated IPv6 packet with the IPv4
        source address it is encapsulated by according to the
        configured parameters of the 6rd domain. If the two
        source addresses do not match, the packet MUST be dropped
        and a counter incremented. This counter indicates the tota
        number of octets dropped packets by the receiving rules."
    INDEX      {ifIndex}
    ::= { sixRdSecurityCeckInvalidPackets 1 }

END

```

7. Security Considerations

This document does not introduce any new security concern in addition to what is discussed in Section 6 of [RFC4087].

8. IANA Considerations

The MIB module in this document uses the following IANA-assigned OBJECT IDENTIFIER values recorded in the SMI Numbers registry, and the following IANA-assigned tunnelType values recorded in the IANAifType-MIB registry:

Descriptor	OBJECT IDENTIFIER value
-----	-----

sixRdMIB	{ transmission XXX }
----------	----------------------

IANA tunnelType ::= TEXTUAL-CONVENTION
SYNTAX INTEGER {

sixRd ("XX") -- 6rd encapsulation

}

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.
- [RFC2578] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.
- [RFC2579] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Textual Conventions for SMIv2", STD 58, RFC 2579, April 1999.
- [RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Conformance Statements for SMIv2", STD 58, RFC 2580, April 1999.
- [RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, June 2000.
- [RFC3371] Caves, E., Calhoun, P., and R. Wheeler, "Layer Two Tunneling Protocol "L2TP" Management Information Base", RFC 3371, August 2002.

- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart,
"Introduction and Applicability Statements for Internet-
Standard Management Framework", RFC 3410, December 2002.
- [RFC3411] Harrington, D., Presuhn, R., and B. Wijnen, "An
Architecture for Describing Simple Network Management
Protocol (SNMP) Management Frameworks", STD 62, RFC 3411,
December 2002.
- [RFC3418] Presuhn, R., "Management Information Base (MIB) for the
Simple Network Management Protocol (SNMP)", STD 62, RFC
3418, December 2002.
- [RFC4087] Thaler, D., "IP Tunnel MIB", RFC 4087, June 2005.
- [RFC4293] Routhier, S., "Management Information Base for the
Internet Protocol (IP)", RFC 4293, April 2006.
- [RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4
Infrastructures (6rd) -- Protocol Specification", RFC
5969, August 2010.

9.2. Informative References

- [RFC4181] Heard, C., "Guidelines for Authors and Reviewers of MIB
Documents", BCP 111, RFC 4181, September 2005.

Authors' Addresses

Lei Cai
ZTE
No. 68 Zijinhua Rd.,
Nanjing 210012
China

Phone: +86 25 5287 2205
Email: cai.lei3@zte.com.cn

Jacni Qin
Cisco Systems
Shanghai
China

Phone: +86 1891 836 3666
Email: jacni@jacni.com

Shishio Tsuchiya (editor)
Cisco Systems
Midtown Tower, 9-7-1, Akasaka
Minato-Ku, Tokyo 107-6227
Japan

Phone: +81 3 6434 6543
Email: shtsuchi@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2013

G. Chen
China Mobile
July 9, 2012

Parameter Provisioning with non-DHCP-PD in Carrier-side Stateless
Solution
draft-chen-software-ce-non-pd-00

Abstract

The deployment of carrier-side stateless solution requires some particular considerations in mobile network contexts. The widely existed legacy network restricts the development of IP address provisioning based on DHCP-PD. The memo provided several potential approaches to facilitate the issues and provide the possibilities leveraging 3GPP mechanism. The stateless algorithm is required to be updated depending to the preferred solution accordingly.

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 January 10, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Possible Solutions	3
2.1. Design Principles	3
2.2. Using IID to carry EA-bits	4
2.3. Changing ND behaviour to deliver IPv6 prefix	4
2.4. Combination of DHCPv4v6 Approach	5
3. Conclusions	5
4. IANA Considerations	5
5. Security Considerations	6
6. References	6
6.1. Normative References	6
6.2. Informative References	6
Author's Address	7

1. Introduction

The work item of carrier-side stateless has been chartered and motivation has been described in [I-D.ietf-softwire-stateless-4v6-motivation]. Several candidates solutions have been designed to meet the demands, such as 4rd[I-D.ietf-softwire-4rd] and MAP-E, MAP-T[I-D.ietf-softwire-map]. Regarding the configuration rule provisioning, a particular DHCPv6 option have been proposed to assign mapping rules to CE nodes. Mapping rules may serve to derive IPv4 address depending on delegated CE IPv6 prefix. Since the potential dependency between IPv4 and IPv6 is existing, CE IPv6 prefix inserted with EA-bits is presumably delegated using DHCP-PD[RFC3633].

IPv6 prefix delegation is a feature of 3GPP Release-10 and is not covered by any earlier releases[RFC6459]. In another words, when the case is coming down to a mobile network contexts, those provisioning should require 3GPP Release 10 or beyond network, which is not deployed anywhere today. Some implementations may consider the lack of the functionalities to use static configuration on mobile terminals, but such solution is not scale well, especially facing several million of customers. Therefore, the absence of DHCP-PD may become a push-back during the deployment in 3GPP network.

This document proposed a number of alternative solutions to the problem, such that the softwires wg can discuss both of them. It is expected that once the softwires wg converges on a preferred solution, the other ones will be removed from the document.

2. Possible Solutions

2.1. Design Principles

In earlier Release 3GPP network, only the /64 prefix could be allocated for each user via SLAAC[RFC4862]. Stateless DHCPv6[RFC3736] is available in those network, but stateful DHCPv6[RFC3315] is not supported. The following solutions is trying to leverage current mobile network. Considering operational approaches to resolve the problems, the document avoid introduce additional protocol changes to existing devices in network sides. The potential changes on UE sides could be easily implemented using user-land codes. A address block could be reserved from addressing plan corresponding to Rule IPv6 prefix. Such CE-specific IPv6 prefix can't be delivered through DHCP-PD, but could be achieved through following approaches.

2.2. Using IID to carry EA-bits

Without IPv6 prefix delegation, network could not assign a prefix shorter than /64. Therefore, a IPv6 prefix with EA-bits might be nowhere to get. One possible approach is to change current port algorithm a bit to get EA-bits from interface identifier (IID), because 3GPP network would assign a IID to user before a global IPv6 address is delivered. Referring to Section 9.2.1.1 in 3GPP [TS23.060], IPv6 address allocation is experiencing two phases. Mobile Gateway shall provide an interface identifier to the user in advance to ensure that the link-local address generated by the user does not collide with the link-local address of the mobile gateway. Mobile gateway could deliver such IID so as to carry EA-bits. When mapping rule has delivered to CE at second phase, CE nodes could derive the EA-bits by shifting with length of u-bits and digging the bits with EA length.

CE could synthesize a CE IPv6 prefix and address according the algorithm respectively. The translation could be performed afterwards with such CE-generated IPv6 address. Since the system already reserve an address block, there is no address collision risks when transmitting the packages.

The pros of the approach is this mechanism doesn't require any changes to 3GPP network. The cons is that requires updates on current algorithm. Distinguishing DHCP-PD and non-PD case is needed to perform EA-bits generation.

2.3. Changing ND behaviour to deliver IPv6 prefix

IPv6 Stateless Address Autoconfiguration (SLAAC), as specified in [RFC4861] and [RFC4862], is widely available in earlier 3GPP network. Since a IPv6 block have been reserved for CE prefix, this block can be treated as on-line for specific user. Referring to Section 9.2.1.1 in 3GPP [TS23.060], UE may send router solicitation message to the mobile gateway on point-to-point link at second phase. Normal process of gateway is to assign a router advertisement message containing /64 prefix heading to UE with A bit set in Prefix Information option. In order to delivery the CE prefix through ND, the gateway should response RA with prefix information option including reserved CE prefix, length of which is usually shorter than /64. The configuration of this prefix option should set O bit and unset A bit.

CE receiving the option could set the prefix as CE IPv6 prefix. The process of algorithm can be remained as-is.

The pros of the approach is to keep current algorithm with no

changes. The cons for that is to increase the provisioning complexity on mobile gateway.

2.4. Combination of DHCPv4v6 Approach

3GPP network can't support the stateful DHCPv6[RFC3315] process. Any process related to stateful would be failed in 3GPP network. Therefore, the provisioning of mapping rules through DHCPv6 would be restricted to stateless parameters provisioning, like rule IPv4 prefix and rule IPv6 prefix. Dynamic information should be retrieved through another way. 3GPP network could support DHCPv4[RFC2131] by indicating to the network within PDP(Packet Data Protocol) activation protocol negotiations. That provides a possibility to carry port-restricted information from DHCPv4 options. In such consideration, port delegation with port mask allocation in [I-D.bajko-pripaddrassign] can be combined with stateless DHCPv6[RFC3736] to populate CE with complete mapping rules.

CE could synthesize a CE IPv6 prefix and address according the algorithm respectively, because CE already received sufficient information to generate IPv6 prefix automatically. The translation could be performed afterwards with such CE-generated IPv6 address. Since the system already reserve the address block, there is no duplicated risks when transmitting the packages.

The combination of DHCPv4v6 would leverage 3GPP process and doesn't require any change to existing 3GPP system. The cons for this approach would potentially cause doubling PDP counts for earlier release terminals before Release 8. The algorithm should also be needed to update accordingly.

3. Conclusions

Three aforementioned solutions have been proposed to address issues of non-DHCP-PD provisioning in mobile contexts. It required WG to evaluate the solutions and determine the workaround for the issues. The minimal impacts to 3GPP system are desirable for the preferable solution.

4. IANA Considerations

This document makes no request of IANA.

5. Security Considerations

TBD

6. References

6.1. Normative References

- [I-D.bajko-pripaddrassign]
Bajko, G., Savolainen, T., Boucadair, M., and P. Levis,
"Port Restricted IP Address Assignment",
draft-bajko-pripaddrassign-04 (work in progress),
April 2012.
- [I-D.ietf-softwire-4rd]
Despres, R., Penno, R., Lee, Y., Chen, G., and S. Jiang,
"IPv4 Residual Deployment via IPv6 - a unified Stateless
Solution (4rd)", draft-ietf-softwire-4rd-02 (work in
progress), June 2012.
- [I-D.ietf-softwire-map]
Troan, O., Dec, W., Li, X., Bao, C., Zhai, Y., Matsushima,
S., and T. Murakami, "Mapping of Address and Port (MAP)",
draft-ietf-softwire-map-01 (work in progress), June 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2131] Droms, R., "Dynamic Host Configuration Protocol",
RFC 2131, March 1997.
- [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.
- [RFC3736] Droms, R., "Stateless Dynamic Host Configuration Protocol
(DHCP) Service for IPv6", RFC 3736, April 2004.
- [TS23.060]
"General Packet Radio Service (GPRS); Service description;
Stage 2", June 2012.

6.2. Informative References

- [I-D.ietf-softwire-stateless-4v6-motivation]
Boucadair, M., Matsushima, S., Lee, Y., Bonness, O.,
Borges, I., and G. Chen, "Motivations for Carrier-side

Stateless IPv4 over IPv6 Migration Solutions",
draft-ietf-softwire-stateless-4v6-motivation-03 (work in
progress), June 2012.

- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC6459] Korhonen, J., Soininen, J., Patil, B., Savolainen, T., Bajko, G., and K. Iisakkila, "IPv6 in 3rd Generation Partnership Project (3GPP) Evolved Packet System (EPS)", RFC 6459, January 2012.

Author's Address

Gang Chen
China Mobile
No.32 Xuanwumen West Street
Xicheng District
Beijing 100053
China

Email: phdgang@gmail.com

Softwire Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 29, 2013

Y. Cui
Tsinghua University
Q. Sun
China Telecom
M. Boucadair
France Telecom
T. Tsou
Huawei Technologies
Y. Lee
Comcast
I. Farrer
Deutsche Telekom AG
February 25, 2013

Lightweight 4over6: An Extension to the DS-Lite Architecture
draft-cui-softwire-b4-translated-ds-lite-11

Abstract

DS-Lite [RFC6333] describes an architecture for transporting IPv4 packets over an IPv6 network. This document specifies an extension to DS-Lite called Lightweight 4over6 which moves the Network Address Translation function from the DS-Lite AFTR to the B4, removing the requirement for a Carrier Grade NAT function in the AFTR. This reduces the amount of centralized state that must be held to a per-subscriber level. In order to delegate the NAPT function and make IPv4 Address sharing possible, port-restricted IPv4 addresses are allocated to the B4s.

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 August 19, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions	4
3. Terminology	4
4. Lightweight 4over6 Architecture	5
5. Lightweight B4 Behavior	7
5.1. Lightweight B4 Provisioning	7
5.2. Lightweight B4 Data Plane Behavior	8
6. Lightweight AFTR Behavior	9
6.1. Binding Table Maintenance	9
6.2. lwAFTR Data Plane Behavior	10
7. Provisioning of IPv4 address and Port Set	11
8. ICMP Processing	12
9. Security Considerations	13
10. IANA Considerations	13
11. Author List	13
12. Acknowledgement	16
13. References	16
13.1. Normative References	16
13.2. Informative References	17
Authors' Addresses	18

1. Introduction

Dual-Stack Lite (DS-Lite, [RFC6333]) defines a model for providing IPv4 access over an IPv6 network using two well-known technologies: IP in IP [RFC2473] and Network Address Translation (NAT). The DS-Lite architecture defines two major functional elements as follows:

Basic Bridging BroadBand element: A B4 element is a function implemented on a dual-stack capable node, either a directly connected device or a CPE, that creates a tunnel to an AFTR.

Address Family Transition Router: An AFTR element is the combination of an IPv4-in-IPv6 tunnel endpoint and an IPv4-IPv4 NAT implemented on the same node.

As the AFTR performs the centralized NAT44 function, it dynamically assigns public IPv4 addresses and ports to requesting host's traffic (as described in [RFC3022]). To achieve this, the AFTR must dynamically maintain per-flow state in the form of active NAT sessions. For service providers with a large number of B4 clients, the size and associated costs for scaling the AFTR can quickly become prohibitive. It can also place a large NAT logging overhead upon the service provider in countries where legal requirements mandate this.

This document describes a mechanism called Lightweight 4 over 6 (lw4o6), which provides a solution for these problems. By relocating the NAT functionality from the centralized AFTR to the distributed B4s, a number of benefits can be realised:

- o NAT44 functionality is already widely supported and used in today's CPE devices. Lw4o6 uses this to provide private<->public NAT44, meaning that the service provider does not need a centralized NAT44 function.
- o The amount of state that must be maintained centrally in the AFTR can be reduced from per-flow to per-subscriber. This reduces the amount of resources (memory and processing power) necessary in the AFTR.
- o The reduction of maintained state results in a greatly reduced logging overhead on the service provider.

Operator's IPv6 and IPv4 addressing architectures remain independent of each other. Therefore, flexible IPv4/IPv6 addressing schemes can

be deployed.

Lightweight 4over6 provides a solution for a hub-and-spoke softwire architecture only. It does not offer direct, meshed IPv4 connectivity between subscribers without packets traversing the AFTR. If this type of meshed interconnectivity is required, [I-D.ietf-softwire-map] provides a suitable solution.

The tunneling mechanism remains the same for DS-Lite and Lightweight 4over6. This document describes the changes to DS-Lite that are necessary to implement Lightweight 4over6. These changes mainly concern the configuration parameters and provisioning method necessary for the functional elements.

Lightweight 4over6 features keeping per-subscriber state in the service provider's network. It is categorized as Binding approach in [I-D.bfmk-softwire-unified-cpe] which defines a unified IPv4-in-IPv6 Softwire CPE.

This document is an extended case, which covers address sharing for [I-D.ietf-softwire-public-4over6]. It is also a variant of A+P called Binding Table Mode (see Section 4.4 of [RFC6346]).

This document focuses on architectural considerations and particularly on the expected behavior of the involved functional elements and their interfaces. Deployment-specific issues are discussed in a companion document. As such, discussions about redundancy and provisioning policy are out of scope.

2. Conventions

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. Terminology

The document defines the following terms:

Lightweight 4over6 (lw4o6): Lightweight 4over6 is an IPv4-over-IPv6 hub and spoke mechanism, which extends DS-Lite by moving the IPv4 translation (NAPT44) function from the AFTR to the B4.

Lightweight B4 (lwB4): A B4 element (Basic Bridging BroadBand element [RFC6333]), which supports Lightweight 4over6 extensions. An lwB4 is a function implemented on a dual-stack capable node, (either a directly connected device or a CPE), that supports port-restricted IPv4 address allocation, implements NAPT44 functionality and creates a tunnel to an lwAFTR

Lightweight AFTR (lwAFTR): An AFTR element (Address Family Transition Router element [RFC6333]), which supports Lightweight 4over6 extension. An lwAFTR is an IPv4-in-IPv6 tunnel endpoint which maintains per-subscriber address binding only and does not perform a NAPT44 function.

Restricted Port-Set: A non-overlapping range of allowed external ports allocated to the lwB4 to use for NAPT44. Source ports of IPv4 packets sent by the B4 must belong to the assigned port-set. The port set is used for all port aware IP protocols (TCP, UDP, SCTP etc.)

Port-restricted IPv4 Address: A public IPv4 address with a restricted port-set. In Lightweight 4over6, multiple B4s may share the same IPv4 address, however, their port-sets must be non-overlapping.

Throughout the remainder of this document, the terms B4/AFTR should be understood to refer specifically to a DS-Lite implementation. The terms lwB4/lwAFTR refer to a Lightweight 4over6 implementation.

4. Lightweight 4over6 Architecture

The Lightweight 4over6 architecture is functionally similar to DS-Lite. lwB4s and an lwAFTR are connected through an IPv6-enabled network. Both approaches use an IPv4-in-IPv6 encapsulation scheme to deliver IPv4 connectivity services. The following figure shows the data plane with main functional change between DS-Lite and lw4o6:

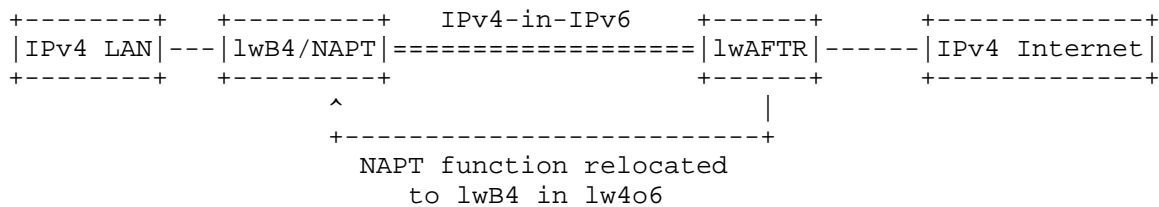


Figure 1 Lightweight 4over6 Data Plane Overview

There are three main components in the Lightweight 4over6 architecture:

- o The lwB4, which performs the NAPT function and encapsulation/de-capsulation IPv4/IPv6.
- o The lwAFTR, which performs the encapsulation/de-capsulation IPv4/IPv6.
- o The provisioning system, which tells the lwB4 which IPv4 address and port set to use.

The lwB4 differs from a regular B4 in that it now performs the NAPT functionality. This means that it needs to be provisioned with the public IPv4 address and port set it is allowed to use. This information is provided through a provisioning mechanism such as DHCP, PCP or TR-69.

The lwAFTR needs to know the binding between the IPv6 address of each subscriber and the IPv4 address and port set allocated to that subscriber. This information is used to perform ingress filtering upstream and encapsulation downstream. Note that this is per-subscriber state as opposed to per-flow state in the regular AFTR case.

The consequence of this architecture is that the information maintained by the provisioning mechanism and the one maintained by the lwAFTR MUST be synchronized (See figure 2). The details of this synchronization depend on the exact provisioning mechanism and will be discussed in a companion draft.

The solution specified in this document allows to assign either a full IPv4 address or shared IPv4 address to requesting CPEs. [I-D.ietf-softwire-public-4over6] provides a mechanism supporting to assign a full IPv4 address only, which could be referred to in this case.

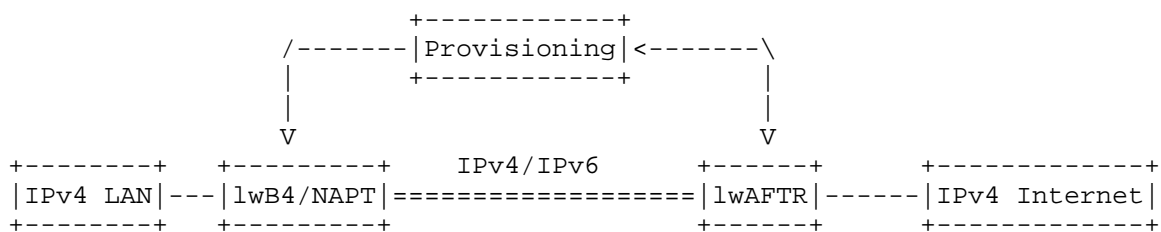


Figure 2 Lightweight 4over6 Provisioning Synchronization

5. Lightweight B4 Behavior

5.1. Lightweight B4 Provisioning

With DS-Lite, the B4 element only needs to be configured with a single DS-Lite specific parameter so that it can set up the software (the IPv6 address of the AFTR). Its IPv4 address can be taken from the well-known range 192.0.0.0/29.

In lw4o6, due to the distributed nature of the NAPT function, a number of lw4o6 specific configuration parameters must be provisioned to the lwB4. These are:

- o IPv6 Address for the lwAFTR
- o IPv4 External (Public) Address for NAPT44
- o Restricted port-set to use for NAPT44

An IPv6 address from an assigned prefix is also required for the lwB4 to use as the encapsulation source address for the softwire. Normally, this is the lwB4's globally unique WAN interface address which can be obtained via an IPv6 address allocation procedure such as SLAAC, DHCPv6 or manual configuration.

In the event that the lwB4's encapsulation source address is changed for any reason (such as the DHCPv6 lease expiring), the lwB4's dynamic provisioning process must be re-initiated.

For learning the IPv6 address of the lwAFTR, the lwB4 SHOULD implement the method described in section 5.4 of [RFC6333] and implement the DHCPv6 option defined in [RFC6334]. Other methods of learning this address are also possible.

An lwB4 MUST support dynamic port-restricted IPv4 address provisioning. The potential port set algorithms are described in

[I-D.sun-dhc-port-set-option], and Section 5.1 of [I-D.ietf-softwire-map]. Several different mechanisms can be used for provisioning the lwB4 with its port-restricted IPv4 address such as: DHCPv4, DHCPv6, PCP and PPP. Some alternatives are mentioned in Section 7 of this document.

In this document, lwB4 can be a binding mode CPE. Its provisioning method is RECOMMENDED to follow that is specified in section 3.3 of [I-D.bfmk-softwire-unified-cpe], which will evolve to reflect the consensus from DHC Working Group.

In the event that the lwB4 receives an ICMPv6 error message (type 1, code 5) originating from the lwAFTR, the lwB4 SHOULD interpret this to mean that no matching entry in the lwAFTR's binding table has been found. The lwB4 MAY then re-initiate the dynamic port-restricted provisioning process. The lwB4's re-initiation policy SHOULD be configurable.

The DNS considerations described in Section 5.5 and Section 6.4 of [RFC6333] SHOULD be followed.

5.2. Lightweight B4 Data Plane Behavior

Several sections of [RFC6333] provide background information on the B4's data plane functionality and MUST be implemented by the lwB4 as they are common to both solutions. The relevant sections are:

- | | |
|-----------------------------------|--|
| 5.2. Encapsulation | Covering encapsulation and de-capsulation of tunneled traffic |
| 5.3. Fragmentation and Reassembly | Covering MTU and fragmentation considerations (referencing [RFC2473]) |
| 7.1. Tunneling | Covering tunneling and traffic class mapping between IPv4 and IPv6 (referencing [RFC2473] and [RFC4213]) |

The lwB4 element performs IPv4 address translation (NAPT44) as well as encapsulation and de-capsulation. It runs standard NAPT44 [RFC3022] using the allocated port-restricted address as its external IPv4 address and port numbers.

The lwB4 should behave as is depicted in (2.2) of section 3.2 of [I-D.bfmk-softwire-unified-cpe] when it starts up. The working flow of the lwB4 is illustrated with figure 3.

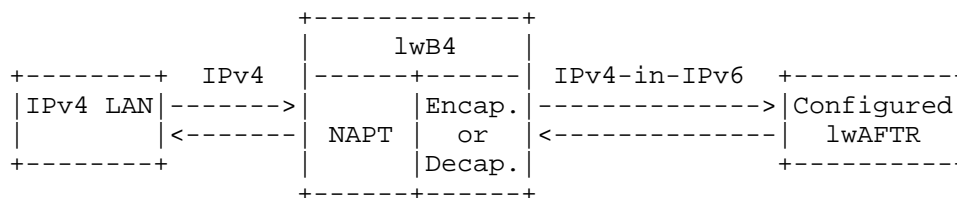


Figure 3 Working Flow of the lwB4

Internally connected hosts source IPv4 packets with an [RFC1918] address. When the lwB4 receives such an IPv4 packet, it performs a NAPT44 function on the source address and port by using the public IPv4 address and a port number from the allocated port-set. Then, it encapsulates the packet with an IPv6 header. The destination IPv6 address is the lwAFTR's IPv6 address and the source IPv6 address is the lwB4's IPv6 tunnel endpoint address. Finally, the lwB4 forwards the encapsulated packet to the configured lwAFTR.

When the lwB4 receives an IPv4-in-IPv6 packet from the lwAFTR, it de-capsulates the IPv4 packet from the IPv6 packet. Then, it performs NAPT44 translation on the destination address and port, based on the available information in its local NAPT44 table.

The lwB4 is responsible for performing ALG functions (e.g., SIP, FTP), and other NAPT traversal mechanisms (e.g., UPnP, NAPT-PMP, manual binding configuration, PCP) for the internal hosts. This requirement is typical for NAPT44 gateways available today.

It is possible that a lwB4 is co-located in a host. In this case, the functions of NAPT44 and encapsulation/de-capsulation are implemented inside the host.

6. Lightweight AFTR Behavior

6.1. Binding Table Maintenance

The lwAFTR maintains an address binding table containing the binding between the lwB4's IPv6 address, the allocated IPv4 address and restricted port-set. Unlike the DS-Lite extended binding table defined in section 6.6 of [RFC6333] which is a 5-tuple NAT table, each entry in the Lightweight 4over6 binding table contains the following 3-tuples:

- o IPv6 Address for a single lwB4

- o Public IPv4 Address
- o Restricted port-set

The entry has two functions: the IPv6 encapsulation of inbound IPv4 packets destined to the lwB4 and the validation of outbound IPv4-in-IPv6 packets received from the lwB4 for de-capsulation.

The lwAFTR does not perform NAT and so does not need session entries.

The lwAFTR MUST synchronize the binding information with the port-restricted address provisioning process. If the lwAFTR does not participate in the port-restricted address provisioning process, the binding MUST be synchronized through other methods (e.g. out-of-band static update).

If the lwAFTR participates in the port-restricted provisioning process, then its binding table MUST be created as part of this process.

For all provisioning processes, the lifetime of binding table entries MUST be synchronized with the lifetime of address allocations.

6.2. lwAFTR Data Plane Behavior

Several sections of [RFC6333] provide background information on the AFTR's data plane functionality and MUST be implemented by the lwAFTR as they are common to both solutions. The relevant sections are:

- | | |
|-----------------------------------|--|
| 6.2. Encapsulation | Covering encapsulation and de-capsulation of tunneled traffic |
| 6.3. Fragmentation and Reassembly | Fragmentation and re-assembly considerations (referencing [RFC2473]) |
| 7.1. Tunneling | Covering tunneling and traffic class mapping between IPv4 and IPv6 (referencing [RFC2473] and [RFC4213]) |

When the lwAFTR receives an IPv4-in-IPv6 packet from an lwB4, it de-capsulates the IPv6 header and verifies the source addresses and port in the binding table. If both the source IPv4 and IPv6 addresses match a single entry in the binding table and the source port in the allowed port-set for that entry, the lwAFTR forwards the packet to

the IPv4 destination.

If no match is found (e.g., no matching IPv4 address entry, port out of range, etc.), the lwAFTR MUST discard or implement a policy (such as redirection) on the packet. An ICMPv6 type 1, code 5 (source address failed ingress/egress policy) error message MAY be sent back to the requesting lwB4. The ICMP policy SHOULD be configurable.

When the lwAFTR receives an inbound IPv4 packet, it uses the IPv4 destination address and port to lookup the destination lwB4's IPv6 address in its binding table. If a match is found, the lwAFTR encapsulates the IPv4 packet. The source is the lwAFTR's IPv6 address and the destination is the lwB4's IPv6 address from the matched entry. Then, the lwAFTR forwards the packet to the lwB4 natively over the IPv6 network.

If no match is found, the lwAFTR MUST discard the packet. An ICMPv4 type 3, code 1 (Destination unreachable, host unreachable) error message MAY be sent back. The ICMP policy SHOULD be configurable.

The lwAFTR MUST support hairpinning of traffic between two lwB4s, by performing de-capsulation and re-encapsulation of packets. The hairpinning policy MUST be configurable.

7. Provisioning of IPv4 address and Port Set

There are several dynamically provisioning protocols for IPv4 address and port set. These protocols MAY be implemented. Some possible alternatives include:

- o DHCP: Extending DHCP protocol MAY be used for the provisioning [I-D.ietf-dhc-dhcpv4-over-ipv6] [I-D.ietf-softwire-map-dhcp].
- o PCP[I-D.ietf-pcp-base]: a lwB4 MAY use [I-D.tsou-pcp-natcoord] to retrieve a restricted IPv4 address and a set of ports.

In a Lightweight 4over6 domain, the same provisioning mechanism MUST be enabled in the lwB4s, the AFTRs and the provisioning server.

DHCP-based provisioning mechanism (DHCPv4/DHCPv6) is RECOMMENDED in this document. The provisioning mechanism for port-restricted IPv4 address will evolve according to the consensus from DHC Working Group.

8. ICMP Processing

ICMP does not work in an address sharing environment without special handling [RFC6269]. Due to the port-set style address sharing, Lightweight 4over6 requires specific ICMP message handling not required by DS-Lite.

The following behavior SHOULD be implemented by the lwAFTR to provide ICMP error handling and basic remote IPv4 service diagnostics for a port restricted CPE: for inbound ICMP messages, the lwAFTR MAY behave in two modes:

Either:

1. Check the ICMP Type field.
2. If the ICMP type is set to 0 or 8 (echo reply or request), then the lwAFTR MUST take the value of the ICMP identifier field as the source port, and use this value to lookup the binding table for an encapsulation destination. If a match is found, the lwAFTR forwards the ICMP packet to the IPv6 address stored in the entry; otherwise it MUST discard the packet.
3. If the ICMP type field is set to any other value, then the lwAFTR MUST use the method described in REQ-3 of [RFC5508] to locate the source port within the transport layer header in ICMP packet's data field. The destination IPv4 address and source port extracted from the ICMP packet are then used to make a lookup in the binding table. If a match is found, it MUST forward the ICMP reply packet to the IPv6 address stored in the entry; otherwise it MUST discard the packet.

Or:

- o Discard all inbound ICMP messages.

The ICMP policy SHOULD be configurable.

The lwB4 SHOULD implement the requirements defined in [RFC5508] for ICMP forwarding. For ICMP echo request packets originating from the private IPv4 network, the lwB4 SHOULD implement the method described in [RFC6346] and use an available port from its port-set as the ICMP Identifier.

For both the lwAFTR and the lwB4, ICMPv6 MUST be handled as described in [RFC2473].

9. Security Considerations

As the port space for a subscriber shrinks due to address sharing, the randomness for the port numbers of the subscriber is decreased significantly. This means it is much easier for an attacker to guess the port number used, which could result in attacks ranging from throughput reduction to broken connections or data corruption.

The port-set for a subscriber can be a set of contiguous ports or non-contiguous ports. Contiguous port-sets do not reduce this threat. However, with non-contiguous port-set (which may be generated in a pseudo-random way [RFC6431]), the randomness of the port number is improved, provided that the attacker is outside the Lightweight 4over6 domain and hence does not know the port-set generation algorithm.

More considerations about IP address sharing are discussed in Section 13 of [RFC6269], which is applicable to this solution.

10. IANA Considerations

This document does not include an IANA request.

11. Author List

The following are extended authors who contributed to the effort:

Jianping Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-62785983
Email: jianping@cernet.edu.cn

Peng Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-62785822
Email: pengwu.thu@gmail.com

Qi Sun
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-62785822
Email: sunqi@csnet1.cs.tsinghua.edu.cn

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

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

Xiaohong Deng
France Telecom

Email: xiaohong.deng@orange.com

Cathy Zhou
Huawei Technologies
Section B, Huawei Industrial Base, Bantian Longgang
Shenzhen 518129
P.R.China

Email: cathyzhou@huawei.com

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

Email: adurand@juniper.net

Reinaldo Penno
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, California 95134
USA

Email: repenno@cisco.com

Alex Clauberg
Deutsche Telekom AG
GTN-FM4
Landgrabenweg 151
Bonn, CA 53227
Germany

Email: axel.clauberg@telekom.de

Lionel Hoffmann
Bouygues Telecom
TECHNOPOLE
13/15 Avenue du Marechal Juin
Meudon 92360
France

Email: lhoffman@bouyguestelecom.fr

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

Email: fibrib@gmail.com

12. Acknowledgement

The authors would like to thank Ole Troan, Ralph Droms and Suresh Krishnan for their comments and feedback.

This document is a merge of three documents:

[I-D.cui-softwire-b4-translated-ds-lite], [I-D.zhou-softwire-b4-nat] and [I-D.penno-softwire-sdnat].

13. References

13.1. Normative References

- [I-D.bfmk-softwire-unified-cpe]
Boucadair, M. and I. Farrer, "Unified IPv4-in-IPv6 Softwire CPE", draft-bfmk-softwire-unified-cpe-02 (work in progress), January 2013.
- [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.
- [RFC3022] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, January 2001.
- [RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.
- [RFC5508] Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT Behavioral Requirements for ICMP", BCP 148, RFC 5508, April 2009.
- [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-

Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.

- [RFC6334] Hankins, D. and T. Mrugalski, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6) Option for Dual-Stack Lite", RFC 6334, August 2011.
- [RFC6346] Bush, R., "The Address plus Port (A+P) Approach to the IPv4 Address Shortage", RFC 6346, August 2011.
- [RFC6431] Boucadair, M., Levis, P., Bajko, G., Savolainen, T., and T. Tsou, "Huawei Port Range Configuration Options for PPP IP Control Protocol (IPCP)", RFC 6431, November 2011.

13.2. Informative References

- [I-D.cui-software-b4-translated-ds-lite]
Cui, Y., Sun, Q., Boucadair, M., Tsou, T., Lee, Y., and I. Farrer, "Lightweight 4over6: An Extension to the DS-Lite Architecture", draft-cui-software-b4-translated-ds-lite-10 (work in progress), February 2013.
- [I-D.ietf-dhc-dhcpv4-over-ipv6]
Cui, Y., Wu, P., Wu, J., and T. Lemon, "DHCPv4 over IPv6 Transport", draft-ietf-dhc-dhcpv4-over-ipv6-05 (work in progress), September 2012.
- [I-D.ietf-pcp-base]
Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-29 (work in progress), November 2012.
- [I-D.ietf-software-map]
Troan, O., Dec, W., Li, X., Bao, C., Matsushima, S., and T. Murakami, "Mapping of Address and Port with Encapsulation (MAP)", draft-ietf-software-map-04 (work in progress), February 2013.
- [I-D.ietf-software-map-dhcp]
Mrugalski, T., Troan, O., Dec, W., Bao, C., leaf.yeh.sdo@gmail.com, l., and X. Deng, "DHCPv6 Options for Mapping of Address and Port", draft-ietf-software-map-dhcp-03 (work in progress), February 2013.
- [I-D.ietf-software-public-4over6]
Cui, Y., Wu, J., Wu, P., Vautrin, O., and Y. Lee, "Public IPv4 over IPv6 Access Network",

draft-ietf-softwire-public-4over6-04 (work in progress),
October 2012.

[I-D.penno-softwire-sdnat]

Penno, R., Durand, A., Hoffmann, L., and A. Clauberg,
"Stateless DS-Lite", draft-penno-softwire-sdnat-02 (work
in progress), March 2012.

[I-D.sun-dhc-port-set-option]

Sun, Q., Lee, Y., Sun, Q., Bajko, G., and M. Boucadair,
"Dynamic Host Configuration Protocol (DHCP) Option for
Port Set Assignment", draft-sun-dhc-port-set-option-00
(work in progress), October 2012.

[I-D.tsou-pcp-natcoord]

Sun, Q., Boucadair, M., Deng, X., Zhou, C., Tsou, T., and
S. Perreault, "Using PCP To Coordinate Between the CGN and
Home Gateway", draft-tsou-pcp-natcoord-09 (work in
progress), November 2012.

[I-D.zhou-softwire-b4-nat]

Zhou, C., Boucadair, M., and X. Deng, "NAT offload
extension to Dual-Stack lite",
draft-zhou-softwire-b4-nat-04 (work in progress),
October 2011.

Authors' Addresses

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

Phone: +86-10-62603059
Email: yong@csnet1.cs.tsinghua.edu.cn

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

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

Mohamed Boucadair
France Telecom
Rennes 35000
France

Email: mohamed.boucadair@orange.com

Tina Tsou
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1-408-330-4424
Email: tena@huawei.com

Yiu L. Lee
Comcast
One Comcast Center
Philadelphia, PA 19103
USA

Email: yiu_lee@cable.comcast.com

Ian Farrer
Deutsche Telekom AG
GTN-FM4, Landgrabenweg 151
Bonn, NRW 53227
Germany

Email: ian.farrer@telekom.de

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 25, 2013

I. Farrer
Deutsche Telekom AG
A. Durand
Juniper Networks
October 22, 2012

lw4over6 Deterministic Architecture
draft-farrer-softwire-lw4o6-deterministic-arch-01

Abstract

This memo describes an architecture for implementing Lightweight 4over6 (lw4o6) in a scalable and resilient manner. This is achieved using characteristics which are inherent to lw4o6 and dynamic IP routing protocols.

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].

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 25, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Deterministic Architecture	3
2.1. Distribution of the Subscriber Population	3
2.2. AFTR Cluster	4
2.3. IPv4 Address Plus Ports to IPv6 Binding Table Considerations	4
2.4. IPv6 and IPv4 Anycast Considerations	5
2.5. Load Balancing across Multiple Concentrators	7
2.6. DHCPv6 Tunnel End-point Option Considerations	7
2.7. CPE IPv6 Address Management	7
2.8. Binding Table Synchronization	7
2.9. Subscriber Management and Growth	8
2.10. Privacy Extensions	8
3. IANA Considerations	8
4. Security Considerations	8
5. Acknowledgements	8
6. References	9
6.1. Normative References	9
6.2. Informative References	9
Authors' Addresses	9

1. Introduction

DS-Lite [RFC6333] is a solution to deal with the IPv4 exhaustion problem once an IPv6 access network is deployed. It enables unmodified IPv4 applications to access the IPv4 Internet over the IPv6 access network. In the DS-Lite architecture, global IPv4 addresses are shared amongst subscribers as the concentrator (AFTR) performs a Carrier-Grade NAT (CGN) function.

[I-D.cui-software-b4-translated-ds-lite] describes Lightweight 4over6, which extends the original DS-Lite model so that NAT is performed by the CPE and IPv4 address sharing is possible through the use of source port-restrictions. AFTRs which only implement the functionality required for lw4o6 (i.e. tunnel concentration without a CGN function) are referred to as lwAFTRs.

This memo provides an operational architecture for the deployment of Lightweight 4over6, offering scalability and high-availability whilst preserving the per-flow stateless nature of the solution.

The approach presented here is stateless and deterministic. It leverages the stateless properties of Lightweight 4over6 to offer a completely deterministic solution. The bindings between IPv4 addresses, ports and IPv6 addresses are pre-computed and stored identically in the lwAFTRs and DHCP servers. This allows for a very simple fail-over mechanism within a cluster of identically provisioned lwAFTRs.

The deterministic architecture is unsuited to per-flow stateful approaches such as DS-Lite, as by their nature, states are dynamically created / deleted and would need to be synchronised in real time between all of the AFTRs in a single cluster to function correctly. This synchronisation greatly increases the complexity of the AFTR and is not required by lwAFTRs.

2. Deterministic Architecture

2.1. Distribution of the Subscriber Population

In a large deployment, it makes sense to distribute the subscriber population into subscriber groups, managed by a single lwAFTR, or by many lwAFTRs grouped in a cluster. Topological considerations and geographical proximity may also be factors in the grouping of subscribers. The exact size of those groups depend on the capacity characteristics of the lwAFTRs and is out of scope for this memo.

Each subscriber group is assigned an IPv6 anycast address and a pool

of IPv4 addresses which are common to all lwAFTRs in a cluster. The IPv4 pool must be sized to handle the subscriber population. No constraints are placed upon the addresses that are used for this pool, in that they can be taken from a single, contiguous block, multiple non-contiguous blocks or single IPv4 addresses as required by the operator.

The exact ratio subscribers to IPv4 addresses, (e.g. the average number of ports assigned per subscriber) is out of scope for this memo.

2.2. AFTR Cluster

All lwAFTRs within a cluster are configured with identical lw4o6 parameters. In particular, they are configured with the same:

- o IPv6 lwAFTR tunnel end-point address
- o IPv4 public pool
- o IPv6 address to IPv4 address and port binding table

2.3. IPv4 Address Plus Ports to IPv6 Binding Table Considerations

The DHCPv4 over IPv6 [I-D.ietf-dhc-dhcpv4-over-ipv6] server will provide each IPv6 CPE an IPv4 address and port range to use within its local NAT binding table. The DHCPv4 server uses the IPv6 address of the CPE as its identifier. As such, the DHCPv4 server contains a table for assigning a specific IPv4 address and ports based on the IPv6 address of the requesting CPE. To maintain the stateless nature of the architecture, DHCPv4 reservation based address assignment is recommended. The lease time for the IPv4 address is unimportant, although a long lease time (or even infinity) is recommended to reduce the number of DHCPv4 requests.

A similar table (containing the same address/port binding information) is also present on all lwAFTRs in the cluster.

The following table shows sample CPE configuration data for a subscriber group. In order for the system to function coherently, this data needs to be kept synchronised between all of the functional elements (lwAFTRs and DHCPv4o6 servers) serving the subscriber group.

IPv6 address	IPv4 address	port-range
2001:db8::1	1.2.3.4	1000-1999
2001:0:1::2	1.2.3.4	2000-2999
2001:0:5::1	2.3.4.5	1500-3999

Figure 1 DHCP4o6/lwAFTR Per-subscriber configuration table data example

This memo proposes a simple architecture to guarantee the synchronization of those mapping tables and rely on anycast IPv4 and IPv6 technologies to provide failover within the lwAFTRs in the same cluster.

2.4. IPv6 and IPv4 Anycast Considerations

The following diagram shows the architecture for the Lightweight 4over6 cluster deployment.

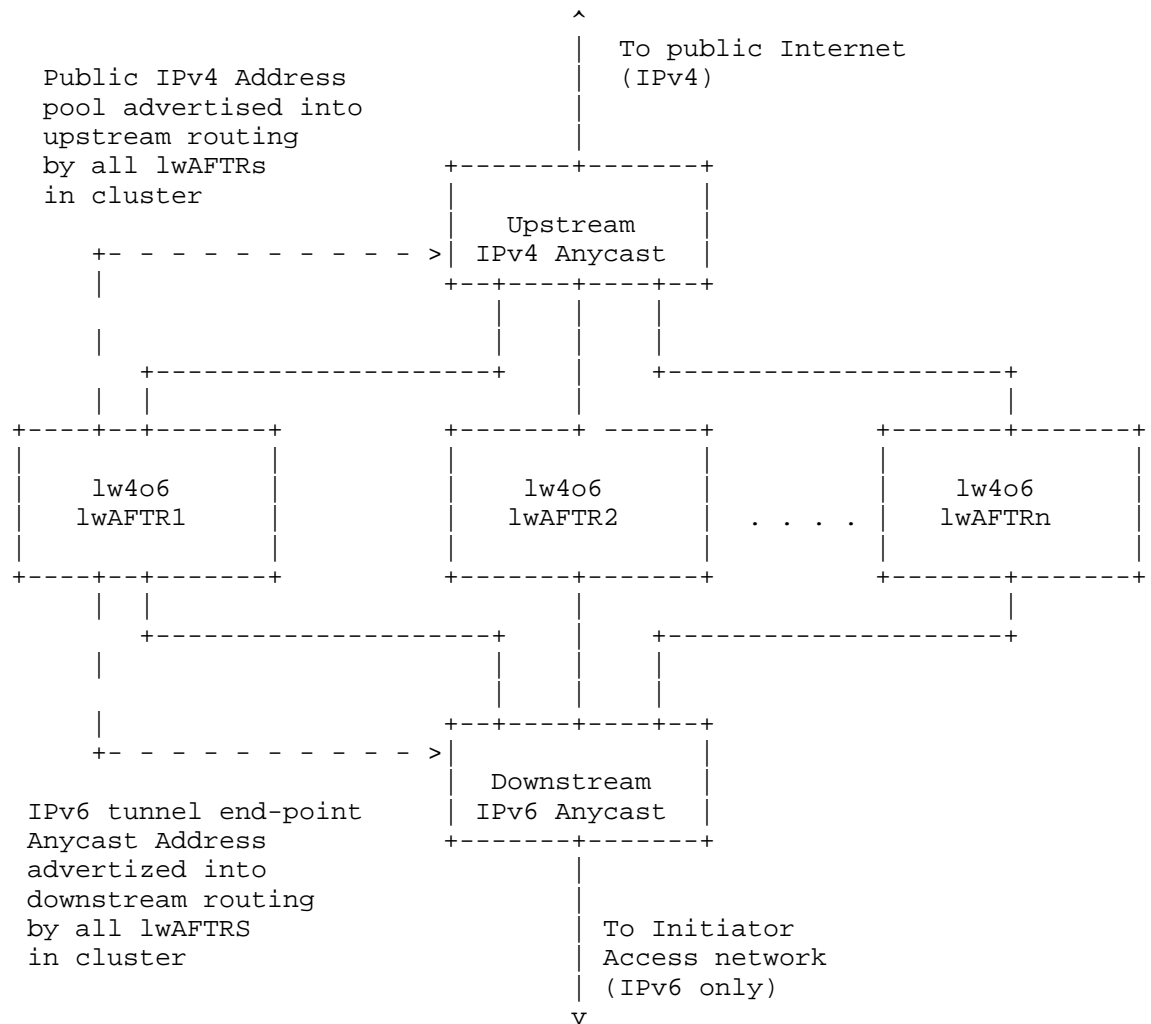


Figure 2 Lightweight 4over6 Cluster

To increase service availability, A simple way to achieve fail-over is to configure both the IPv6 tunnel end point address and the IPv4 address pool as anycast addresses on the lwAFTRs and announce these routes into the IGP run by the ISP, such as OSPF or IS-IS.

The number of lwAFTRs in a cluster provides the degree of redundancy of the solution. In practice, two lwAFTRs are expected to be sufficient in most cases.

2.5. Load Balancing across Multiple Concentrators

lwAFTR functionality can be scaled by load balancing the encapsulation/decapsulation across multiple lwAFTRs in a cluster. Due to the commonality of configuration and stateless nature of the solution, any tunneled packet from any Initiator served by the cluster can arrive at any cluster member and will be processed in the same way. Likewise, for inbound packets originating in the IPv4 realm, a packet that arrives at any of the cluster member will be encapsulated and sent to the correct initiator.

Load balancing could be achieved using specific load balancing infrastructure to distribute the tunnels and inbound v4 traffic across the cluster. It is also possible to use the Equal Cost Multipath inherent in some routing protocols to achieve this.

In order to prevent out-of-sequence packets in the tunnelled traffic, a mechanism for forwarding all packets belonging to a single tunnel through the same cluster member should be used. An example of this would be a source/destination hashing algorithm such as [RFC2992] describes.

2.6. DHCPv6 Tunnel End-point Option Considerations

All CPEs belonging to the same group of subscribers need to receive the same tunnel end-point option (via DHCPv6). This will be set to the IPv6 anycast address of the lwAFTR cluster.

2.7. CPE IPv6 Address Management

The DHCPv4 server uses the IPv6 address of the CPE as its index. In order to keep the overall service architecture flexible and adaptable, it is preferable that the CPE is configured using DHCPv6 out of a specific pool reserved by the ISP.

2.8. Binding Table Synchronization

It is proposed that binding tables be pre-computed and stored statically on the lwAFTRs and the DHCPv4 servers. The method of creating the binding tables is out of the scope of this memo.

These tables are not expected to change regularly. Typical reasons for an update include adding or removing an IPv4 address block, or changing the size of IPv4 ports ranges available to each CPE.

To ensure continuous operation, binding tables have to be updated simultaneously across all lwAFTRs in a cluster by a mechanism such as netconf. It may also be necessary to reconfigure CPEs during this

process (e.g. via a DHCPv6 reconifigure message). The details are out of scope for this memo.

2.9. Subscriber Management and Growth

It is recommended that the ISP predefines all IPv6 addresses and corresponding IPv4 addresses and port ranges for any given subscriber group.

If the ISP runs out of space within a subscriber group, another group is then defined. Customer CPEs can be migrated between different subscriber groups by alternating the CPE configuration over DHCP.

2.10. Privacy Extensions

In some deployments, regulations require that IP addresses allocated to customers can be changed periodically or on demand to protect users privacy.

This can be achieved by rolling over the IPv6 addresses in the DHCPv6 server allocating IPv6 addresses to the CPE. If all subscribers within the subscriber group are allocated the same number of ports in IPv4, then the IPv6 to IPv4 address and port binding may remain the same, the IPv4 address and ports will then roll over automatically at the same time as the IPv6 addresses do.

[I-D.cui-softwire-b4-translated-ds-lite] states that when the IPv6 address of the lwB4 is changed, then DHCPv4over6 configuration process must be re-initiated.

3. IANA Considerations

None.

4. Security Considerations

None.

5. Acknowledgements

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.

6.2. Informative References

- [I-D.cui-software-b4-translated-ds-lite]
Cui, Y., Sun, Q., Boucadair, M., Tsou, T., Lee, Y., and I. Farrer, "Lightweight 4over6: An Extension to the DS-Lite Architecture", draft-cui-software-b4-translated-ds-lite-08 (work in progress), September 2012.
- [I-D.ietf-dhc-dhcpv4-over-ipv6]
Cui, Y., Wu, P., Wu, J., and T. Lemon, "DHCPv4 over IPv6 Transport", draft-ietf-dhc-dhcpv4-over-ipv6-05 (work in progress), September 2012.

Authors' Addresses

Ian Farrer
Deutsche Telekom AG
GTN-FM4
Landgrabenweg 151
Bonn 53227
Germany

Email: ian.farrer@telekom.de

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

Email: adurand@juniper.net

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: November 15, 2013

Y. Fu
S. Jiang
B.Liu
Huawei Technologies Co., Ltd
J.Dong
P. Wu
Tsinghua University
May 14, 2013

Definitions of Managed Objects for MAP-E
draft-fu-softwire-map-mib-05

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 November 15, 2013.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

This memo defines a portion of the Management Information Base (MIB) for using with network management protocols in the Internet community. In particular, it defines managed objects for MAP encapsulation mode.

Table of Contents

1. Introduction	3
2. The Internet-Standard Management Framework	3
3. Terminology	3
4. Structure of the MIB Module	3
4.1. The mapMIBObjects	4
4.1.1. The mapRule Subtree	4
4.1.2. The mapSecurityCheck Subtree	4
4.2. The mapMIBConformance Subtree	4
5. Definitions	4
6. IANA Considerations	12
7. Security Considerations	12
8. Acknowledgments	12
9. References	12
9.1. Normative References	12
9.2. Informative References	13
10. Change Log [RFC Editor please remove]	13
Author's Addresses	14

1. Introduction

MAP [I-D. draft-ietf-softwire-map] is a stateless mechanism for running IPv4 over IPv6-only infrastructure. In particular, it includes two mode, translation mode or encapsulation mode. For the encapsulation mode, it provides an automatic tunnelling mechanism for providing IPv4 connectivity service to end users over a service provider's IPv6 network.

This document defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. This MIB module may be used for monitoring the devices in the MAP scenario, especially, for the encapsulation mode.

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of [RFC3410].

Managed objects are accessed via a virtual information store, termed the MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP).

Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIv2, which is described in [RFC2578], [RFC2579] and [RFC2580].

3. Terminology

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].

4. Structure of the MIB Module

The MAP-E MIB provides a way to configure and manage the devices in MAP encapsulation mode through SNMP.

MAP-E MIB is configurable on a per-interface basis. It depends on several parts of the IF-MIB [RFC2863].

4.1. The mapMIBObjects

4.1.1. The mapRule Subtree

The mapRule subtree describes managed objects used for managing the multiple mapping rules in the MAP encapsulation mode.

According to the MAP specification, the mapping rules are divided into two categories, which are BMR (Basic Mapping Rule), and FMR (Forwarding Mapping Rule).

4.1.2. The mapSecurityCheck Subtree

The mapSecurityCheck subtree is to statistic the number of invalid packets that been identified. There are two kind of invalid packets which are defined in the MAP specification as the following.

- The BR MUST perform a validation of the consistency of the source IPv6 address and source port number for the packet using BMR.
- The CE SHOULD check that MAP received packets' transport-layer destination port number is in the range configured by MAP for the CE.

4.2. The mapMIBConformance Subtree

The mapMIBConformance subtree provides conformance information of MIB objects.

5. Definitions

```
MAP-E-MIB DEFINITIONS ::= BEGIN

IMPORTS
    MODULE-IDENTITY, OBJECT-TYPE, mib-2, transmission,
    Gauge32, Integer32, Counter64
        FROM SNMPv2-SMI    --[RFC2578]

    RowStatus, StorageType, DisplayString
        FROM SNMPv2-TC     --[RFC2579]

    ifIndex, InterfaceIndexOrZero
        FROM IF-MIB        --[RFC2863]

    InetAddressType, InetAddress,
    InetPortNumber, InetAddressPrefixLength
        FROM INET-ADDRESS-MIB --[RFC4001]
```

OBJECT-GROUP, MODULE-COMPLIANCE,
NOTIFICATION-GROUP
FROM SNMPv2-CONF; --[RFC2580]

mapMIB MODULE-IDENTITY
LAST-UPDATED "201302070000Z" -- February 6, 2013
ORGANIZATION "IETF Softwire Working Group"
CONTACT-INFO

"Yu Fu
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd., Hai-Dian District
Beijing, P.R. China 100095
EMail: eleven.fuyu@huawei.com

Sheng Jiang
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd., Hai-Dian District
Beijing, P.R. China 100095
EMail: jiangsheng@huawei.com

Bing Liu
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd., Hai-Dian District
Beijing, P.R. China 100095
EMail: leo.liubing@huawei.com

Jiang Dong
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China
Email: dongjiang@csnet1.cs.tsinghua.edu.cn

Peng Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China
Email: weapon@csnet1.cs.tsinghua.edu.cn"

DESCRIPTION

"The MIB module is defined for management of objects in the
MAP-E BRs or CEs."

REVISION "201305140000Z"

```

 ::= { transmission xxx } --xxx to be replaced with
IANA-assigned value

mapMIBObjects OBJECT IDENTIFIER ::= {mapMIB 1}

mapRule OBJECT IDENTIFIER
 ::= { mapMIBObjects 1 }

mapSecurityCheck OBJECT IDENTIFIER
 ::= { mapMIBObjects 2 }

mapRuleTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF mapRuleEntry
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The (conceptual) table containing rule Information of
        specific mapping rule. It can also be used for row
        creation."
    ::= { mapRule 1 }

mapRuleEntry OBJECT-TYPE
    SYNTAX      MapRuleEntry
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "Each entry in this table contains the information on a
        particular mapping rule."
    INDEX      { mapRuleID }
    ::= { mapRuleTable 1 }

mapRuleEntry ::=
    SEQUENCE {
        mapRuleID                      Integer32,
        mapRuleIPv6PrefixType          InetAddressType,
        mapRuleIPv6Prefix              InetAddress,
        mapRuleIPv6PrefixLen           InetAddressPrefixLength,
        mapRuleIPv4PrefixType          InetAddressType,
        mapRuleIPv4Prefix              InetAddress,
        mapRuleIPv4PrefixLen           InetAddressPrefixLength,
        mapRuleStartPort               InetPortNumber,
        mapRuleEndPort                 InetPortNumber,
        mapRuleEALen                   Integer32,
        mapRuleStatus                  RowStatus,
        mapRuleStorageType              StorageType,

```

```
    mapRuleType Integer32
  }

mapRuleID OBJECT-TYPE
    SYNTAX Integer32 (1..2147483647)
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "An identifier used to distinguish the multiple mapping
        rule which is unique with each CE in the same BR."
    ::= { mapRuleEntry 1 }

mapRuleIPv6PrefixType OBJECT-TYPE
    SYNTAX InetAddressType
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "In this object, it MUST be set to the value of 2 to
        present IPv6 type. It complies the textule convention
        of IPv6 address defined in [RFC4001]."
    ::= { mapRuleEntry 2 }

mapRuleIPv6Prefix OBJECT-TYPE
    SYNTAX InetAddress
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "The IPv6 prefix defined in mapping rule which will be
        assigned to CE ."
    ::= { mapRuleEntry 3 }

mapRuleIPv6PrefixLen OBJECT-TYPE
    SYNTAX InetAddressPrefixLength
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "The length of the IPv6 prefix defined in the mapping rule.
        As a parameter for mapping rule, it will be also assigned
        to CE."
    ::= { mapRuleEntry 4 }

mapRuleIPv4PrefixType OBJECT-TYPE
    SYNTAX InetAddressType
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "In this object, it MUST be set to the value of 1 to
```

present IPv4 type. It complies the textual convention of IPv6 address defined in [RFC4001]."

```
::= { mapRuleEntry 5 }
```

```
mapRuleIPv4Prefix OBJECT-TYPE
    SYNTAX      InetAddress
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        " The IPv4 prefix defined in mapping rule which will be
          assigned to CE."
    ::= { mapRuleEntry 6 }
```

```
mapRuleIPv4PrefixLen OBJECT-TYPE
    SYNTAX      InetAddressPrefixLength
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The length of the IPv4 prefix defined in the mapping
          rule. As a parameter for mapping rule, it will be also
          assigned to CE."
    ::= { mapRuleEntry 7 }
```

```
mapRuleStartPort OBJECT-TYPE
    SYNTAX      InetPortNumber
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The start port number of the port range derived
          from the mapping rule which will be assigned to CE."
    ::= { mapRuleEntry 8 }
```

```
mapRuleEndPort OBJECT-TYPE
    SYNTAX      InetPortNumber
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        " The end port number of the port range derived
          from the mapping rule which will be assigned to CE."
    ::= { mapRuleEntry 9 }
```

```
mapRuleEALen OBJECT-TYPE
    SYNTAX      Integer32
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The length of the Embedded-Address (EA) defined in
```

```
        mapping rule which will be assigned to CE."
 ::= { mapRuleEntry 10 }

mapRuleStatus OBJECT-TYPE
    SYNTAX      RowStatus
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The status of this row, by which new entries may be
         created, or old entries deleted from this table."
 ::= { mapRuleEntry 11 }

mapRuleStorageType OBJECT-TYPE
    SYNTAX      StorageType
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The storage type of this row. If the row is
         permanent(4), no objects in the row need be
         writable."
 ::= { mapRuleEntry 12 }

mapRuleType OBJECT-TYPE
    SYNTAX      Integer32
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The type of the mapping rule. A value of 0 means it
         is a BMR; a non-zero value means it is a FMR."
 ::= { mapRuleEntry 12 }

mapSecurityCheckTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF MapSecurityCheckEntry
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "The (conceptual) table containing information on
         MAP security checks. This table can be used to statistic
         the number of invalid packets that been identified"
 ::= { mapSecurityCheck 1 }

mapSecurityCheckEntry OBJECT-TYPE
    SYNTAX      mapSecurityCheckEntry
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "Each entry in this table contains the information on a
```



```
particular MAP SecurityCheck."
INDEX      { mapSecurityCheckInvalidv4,
              mapSecurityCheckInvalidv6}
::= { mapSecurityCheckTable 1 }

mapSecurityCheckEntry ::=
SEQUENCE {
    mapSecurityCheckInvalidv4      Counter64,
    mapSecurityCheckInvalidv6      Counter64,
    mapSecurityCheckStatus         RowStatus,
    mapSecurityCheckStorageType    StorageType
}

mapSecurityCheckInvalidv4 OBJECT-TYPE
SYNTAX      Counter64
MAX-ACCESS  read-create
STATUS      current
DESCRIPTION
    "The CE SHOULD check that MAP received packets'
    transport-layer destination port number is in the range
    configured by MAP for the CE"
::= { mapSecurityCheckEntry 1 }

mapSecurityCheckInvalidv6 OBJECT-TYPE
SYNTAX      Counter64
MAX-ACCESS  read-create
STATUS      current
DESCRIPTION
    "The BR MUST perform a validation of the consistency of
    the source IPv6 address and source port number for the
    packet using BMR."
::= { mapSecurityCheckEntry 2 }

mapSecurityCheckStatus OBJECT-TYPE
SYNTAX      RowStatus
MAX-ACCESS  read-create
STATUS      current
DESCRIPTION
    "The status of this row, by which new entries may be
    created, or old entries deleted from this table."
::= { mapSecurityCheckEntry 3 }

mapSecurityCheckStorageType OBJECT-TYPE
SYNTAX      StorageType
MAX-ACCESS  read-create
STATUS      current
DESCRIPTION
```

```
        "The storage type of this row. If the row is
        permanent(4), no objects in the row need be
        writable."
    ::= { mapSecurityCheckEntry 4 }

-- Conformance Information

mapMIBConformance OBJECT IDENTIFIER ::= {mapMIB 2}

mapMIBCompliances OBJECT IDENTIFIER ::= { mapMIBConformance 1 }

mapMIBGroups OBJECT IDENTIFIER ::= { mapMIBConformance 2 }

-- compliance statements

mapMIBCompliance MODULE-COMPLIANCE
    STATUS current
    DESCRIPTION
        " Describes the minimal requirements for conformance
        to the MAP-E MIB."
    MODULE -- this module
        MANDATORY-GROUPS { mapMIBRuleGroup }
    ::= { mapMIBCompliances 1 }

-- Units of Conformance

mapMIBRuleGroup OBJECT-GROUP
    OBJECTS { mapRuleBAddress, mapMapRuleID,
        mapRuleIPv6Prefix,
        mapRuleIPv6PrefixLen,
        mapRuleIPv4Prefix,
        mapRuleIPv4PrefixLen,
        mapRuleStartPort,
        mapRuleEndPort mapRuleEALen,
        mapRuleStorageType }
    STATUS current
    DESCRIPTION
        " The collection of this objects are used to give the
        information of mapping rules in MAP-E."
    ::= { mapMIBGroups 1 }

    END
```

6. IANA Considerations

The MIB module in this document uses the following IANA-assigned OBJECT IDENTIFIER values recorded in the SMI Numbers registry:

Descriptor	OBJECT IDENTIFIER value
-----	-----
MAP-E-MIB	{ transmission XXX }

7. Security Considerations

The MAP-E MIB module can be used for configuration of certain objects, and anything that can be configured can be incorrectly configured, with potentially disastrous results. Because this MIB module reuses the IP tunnel MIB, the security considerations for these MIBs are also applicable to the MAP-E MIB.

SNMP versions prior to SNMPv3 did not include adequate security. Even if the network itself is secure (for example by using IPsec), even then, there is no control as to who on the secure network is allowed to access and GET/SET (read/change/create/delete) the objects in this MIB module.

It is RECOMMENDED that implementers consider the security features as provided by the SNMPv3 framework (see [RFC3410], section 8), including full support for the SNMPv3 cryptographic mechanisms (for authentication and privacy).

Further, deployment of SNMP versions prior to SNMPv3 is NOT RECOMMENDED. Instead, it is RECOMMENDED to deploy SNMPv3 and to enable cryptographic security. It is then a customer/operator responsibility to ensure that the SNMP entity giving access to an instance of this MIB module is properly configured to give access to the objects only to those principles (users) that have legitimate rights to indeed GET or SET (change/create/delete) them.

8. Acknowledgments

The authors would like to thank for valuable comments from David Harrington, Mark Townsley, and Shishio Tsuchiya.

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.

- [RFC2578] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Structure of Management Information Version 2 (SMIv2)", RFC 2578, April 1999.
- [RFC2579] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Textual Conventions for SMIv2", RFC 2579, April 1999.
- [RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Conformance Statements for SMIv2", RFC 2580, April 1999.
- [RFC2863] McCloghrie, K. and F. Kastenholz. "The Interfaces Group MIB", RFC 2863, June 2000.
- [RFC3411] Harrington, D., Presuhn, R., and B. Wijnen, "An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks", RFC 3411, December 2002.
- [RFC4001] Daniele, M., Haberman, B., Routhier, S., and J. Schoenwaelder, "Textual Conventions for Internet Network Addresses", RFC 4001, February 2005.
- [RFC4087] Thaler, D., "IP Tunnel MIB", RFC 4087, June 2005.
- [I-D.ietf-softwire-map]
Troan, O., etc., "Mapping of Address and Port (MAP)",
draft-ietf-softwire-map, working in progress.
- [I-D.mdt-softwire-map-dhcp-option]
Mrugalski, T., etc., "DHCPv6 Options for Mapping of Address
and Port", draft-mdt-softwire-map-dhcp-option, working in
progress.

9.2. Informative References

- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.

10. Change Log [RFC Editor please remove]

draft-fu-softwire-map-mib-00, original version, 2012-03-01
draft-fu-softwire-map-mib-01, 01 version, 2012-07-16
draft-fu-softwire-map-mib-03, deleted tunnel object according to the
discussion in IETF85, 2013-02-04
draft-fu-softwire-map-mib-04, added security check object according
to discussion in IETF86

draft-fu-softwire-map-mib-05, distinguishing FMR and BMR in mapRule object definition; added some description in section 4; modifying a little bit to the mapRuleEntry definition

Author's Addresses

Yu Fu
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd.
Hai-Dian District, Beijing 100095
P.R. China
Email: eleven.fuyu@huawei.com

Sheng Jiang
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd.
Hai-Dian District, Beijing 100095
P.R. China
Email: jiangsheng@huawei.com

Bing Liu
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd.,
Hai-Dian District, Beijing 100095
P.R. China
Email: leo.liubing@huawei.com

Jiang Dong
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China
Email: dongjiang@csnet1.cs.tsinghua.edu.cn

Peng Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China
Email: weapon@csnet1.cs.tsinghua.edu.cn

Softwire WG
Internet-Draft
Intended status: Standards Track
Expires: December 5, 2015

Q. Wang
China Telecom
W. Meng
C. Wang
ZTE Corporation
M. Boucadair
France Telecom
June 3, 2015

RADIUS Extensions for IPv4-Embedded Multicast and Unicast IPv6 Prefixes
draft-hu-softwire-multicast-radius-ext-08

Abstract

This document specifies a new Remote Authentication Dial-In User Service (RADIUS) attribute to carry the Multicast-Prefixes-64 information, aiming to delivery the Multicast and Unicast IPv6 Prefixes to be used to build multicast and unicast IPv4-Embedded IPv6 addresses. this RADIUS attribute is defined based on the equivalent DHCPv6 OPTION_v6_PREFIX64 option.

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 December 5, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Convention and Terminology	4
3. Multicast-Prefixes-64 Configuration with RADIUS and DHCPv6	5
4. RADIUS Attribute	8
4.1. Multicast-Prefixes-64	8
5. Table of Attributes	11
6. Security Considerations	12
7. IANA Considerations	13
8. Acknowledgments	14
9. Normative References	15
Authors' Addresses	16

1. Introduction

The solution specified in [I-D.ietf-softwire-dslite-multicast] relies on stateless functions to graft part of the IPv6 multicast distribution tree and IPv4 multicast distribution tree, also uses IPv4-in-IPv6 encapsulation scheme to deliver IPv4 multicast traffic over an IPv6 multicast-enabled network to IPv4 receivers.

To inform the mB4 element of the PREFIX64, a PREFIX64 option may be used. [I-D.ietf-softwire-multicast-prefix-option] defines a DHCPv6 PREFIX64 option to convey the IPv6 prefixes to be used for constructing IPv4-embedded IPv6 addresses.

In broadband environments, a customer profile may be managed by Authentication, Authorization, and Accounting (AAA) servers, together with AAA for users. The Remote Authentication Dial-In User Service (RADIUS) protocol [RFC2865] is usually used by AAA servers to communicate with network elements. Since the Multicast-Prefixes-64 information can be stored in AAA servers and the client configuration is mainly provided through DHCP running between the NAS and the requesting clients, a new RADIUS attribute is needed to send Multicast-Prefixes-64 information from the AAA server to the NAS.

This document defines a new RADIUS attribute to be used for carrying the Multicast-Prefixes-64, based on the equivalent DHCPv6 option already specified in [I-D.ietf-softwire-multicast-prefix-option].

This document makes use of the same terminology defined in [I-D.ietf-softwire-dslite-multicast].

This attribute can be in particular used in the context of DS-Lite Multicast, MAP-E Multicast and other IPv4-IPv6 Multicast techniques. However it is not limited to DS-Lite Multicast.

DS-Lite unicast RADIUS extensions are defined in [RFC6519] .

2. Convention and Terminology

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].

The terms DS-Lite multicast Basic Bridging BroadBand element (mB4) and the DS-Lite multicast Address Family Transition Router element (mAFTR) are defined in [I-D.ietf-softwire-dslite-multicast]

3. Multicast-Prefixes-64 Configuration with RADIUS and DHCPv6

Figure 1 illustrates in DS-Lite scenario how the RADIUS protocol and DHCPv6 work together to accomplish Multicast-Prefixes-64 configuration on the mB4 element for multicast service when an IP session is used to provide connectivity to the user.

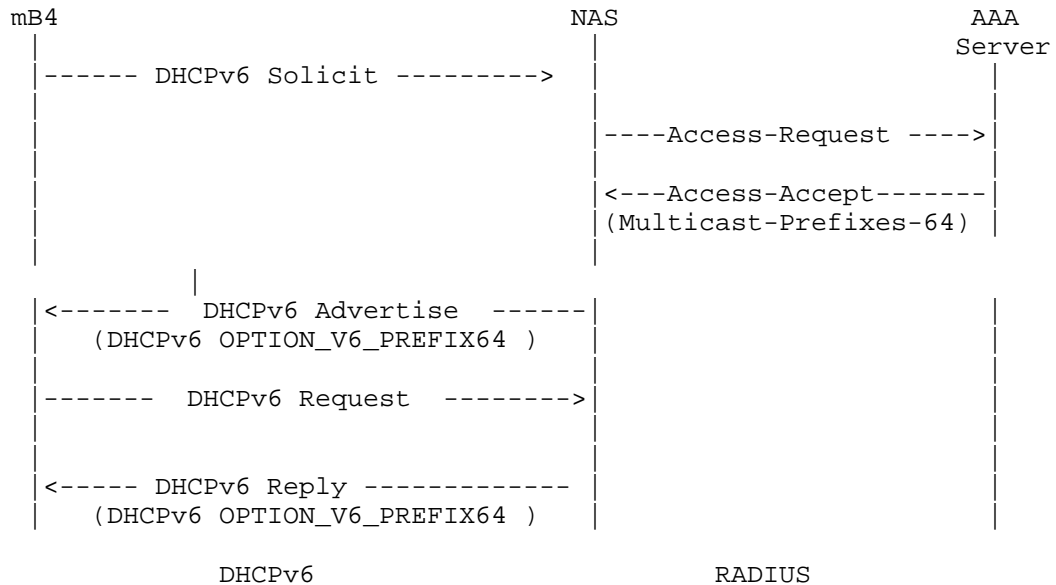


Figure 1: RADIUS and DHCPv6 Message Flow for an IP Session

The NAS operates as a client of RADIUS and as a DHCP Server/Relay for mB4. When the mB4 sends a DHCPv6 Solicit message to NAS (DHCP Server/Relay). The NAS sends a RADIUS Access-Request message to the RADIUS server, requesting authentication. Once the RADIUS server receives the request, it validates the sending client, and if the request is approved, the AAA server replies with an Access-Accept message including a list of attribute-value pairs that describe the parameters to be used for this session. This list MAY contain the Multicast-Prefixes-64 attribute (asm-length, ASM_PREFIX64, ssm-length, SSM_PREFIX64, unicast-length, U_PREFIX64). Then, when the NAS receives the DHCPv6 Request message containing the OPTION_V6_PREFIX64 option in its Option Request option, the NAS SHALL use the prefixes returned in the RADIUS Multicast-Prefixes-64 attribute to populate the DHCPv6 OPTION_V6_PREFIX64 option in the DHCPv6 reply message.

NAS MAY be configured to return the configured Multicast-Prefixes-64 by the AAA Server to any requesting client without relaying each received request to the AAA Server.

Figure 2 describes another scenario, which accomplish DS-Lite Multicast-Prefixes-64 configuration on the mB4 element for multicast service when a PPP session is used to provide connectivity to the user. Once the NAS obtains the Multicast-Prefixes-64 attribute from the AAA server through the RADIUS protocol, the NAS MUST store the received Multicast-Prefixes-64 locally. When a user is online and sends a DHCPv6 Request message containing the OPTION_V6_PREFIX64 option in its Option Request option, the NAS retrieves the previously stored Multicast-Prefixes-64 and uses it as OPTION_V6_PREFIX64 option in DHCPv6 Reply message.

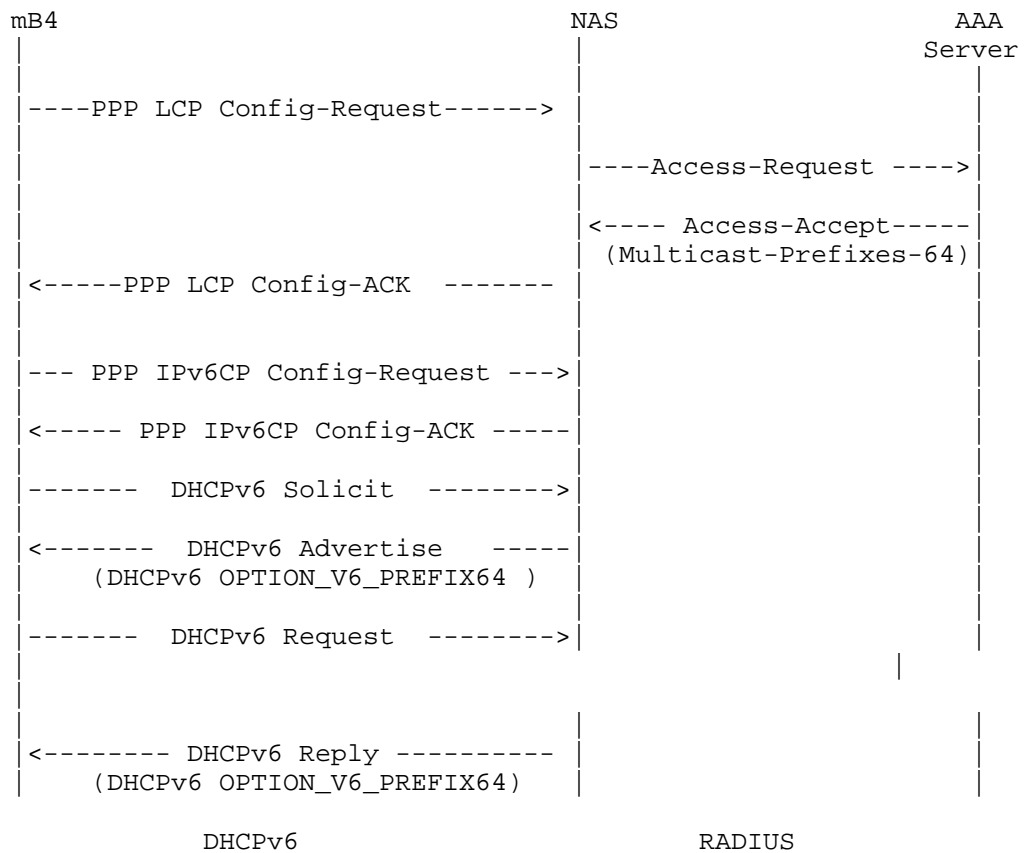


Figure 2: RADIUS and DHCPv6 Message Flow for a PPP Session

According to [RFC3315], after receiving the Multicast-Prefixes-64 attribute in the initial Access-Accept packet, the NAS MUST store the received V6_PREFIX64 locally. When the mB4 sends a DHCPv6 Renew message to request an extension of the lifetimes for the assigned address or prefix, the NAS does not have to initiate a new Access-

Request packet towards the AAA server to request the Multicast-Prefixes-64. The NAS retrieves the previously stored Multicast-Prefixes-64 and uses it in its reply.

Also, if the DHCPv6 server to which the DHCPv6 Renew message was sent at time T1 has not responded, the DHCPv6 client initiates a Rebind/Reply message exchange with any available server. In this scenario, the NAS receiving the DHCPv6 Rebind message MUST initiate a new Access-Request message towards the AAA server. The NAS MAY include the Multicast-Prefixes-64 attribute in its Access-Request message.

4. RADIUS Attribute

This section specifies the format of the new RADIUS attribute.

4.1. Multicast-Prefixes-64

The Multicast-Prefixes-64 attribute conveys the IPv6 prefixes to be used in [I-D.ietf-softwire-dslite-multicast] to synthesize IPv4-embedded IPv6 addresses. The NAS SHALL use the IPv6 prefixes returned in the RADIUS Multicast-Prefixes-64 attribute to populate the DHCPv6 PREFIX64 Option [I-D.ietf-softwire-multicast-prefix-option] .

This attribute MAY be used in Access-Request packets as a hint to the RADIUS server, for example, if the NAS is pre-configured with Multicast-Prefixes-64, these prefixes MAY be inserted in the attribute. The RADIUS server MAY ignore the hint sent by the NAS, and it MAY assign a different Multicast-Prefixes-64 attribute.

If the NAS includes the Multicast-Prefixes-64 attribute, but the AAA server does not recognize this attribute, this attribute MUST be ignored by the AAA server.

NAS MAY be configured with both ASM_PREFIX64 and SSM_PREFIX64 or only one of them. Concretely, AAA server MAY return ASM_PREFIX64 or SSM_PREFIX64 based on the user profile and service policies. AAA MAY return both ASM_PREFIX64 and SSM_PREFIX64. When SSM_PREFIX64 is returned by the AAA server, U_PREFIX64 MUST also be returned by the AAA server.

If the NAS does not receive the Multicast-Prefixes-64 attribute in the Access-Accept message, it MAY fall back to a pre-configured default Multicast-Prefixes-64, if any. If the NAS does not have any pre-configured, the delivery of multicast traffic is not supported.

If the NAS is pre-provisioned with a default Multicast-Prefixes-64 and the Multicast-Prefixes-64 received in the Access-Accept message are different from the configured default, then the Multicast-Prefixes-64 attribute received in the Access-Accept message MUST be used for the session.

A summary of the Multicast-Prefixes-64 RADIUS attribute format is shown Figure 3. The fields are transmitted from left to right.

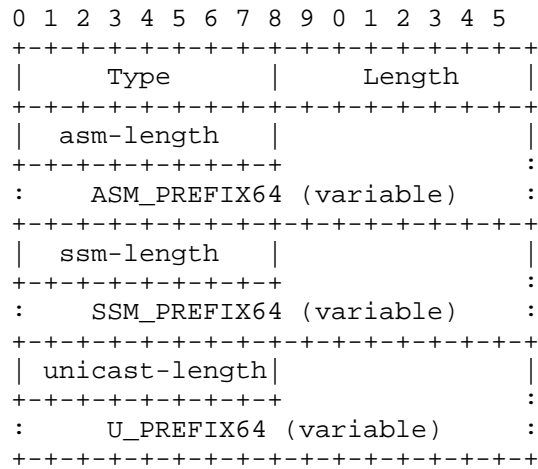


Figure 3: RADIUS attribute format for Multicast-Prefixes-64

Type:

145 for Multicast-Prefixes-64

Length:

This field indicates the total length in octets of this attribute including the Type and Length fields, and the length in octets of all PREFIX fields.

asm-length:

the prefix-length for the ASM IPv4-embedded prefix, as an 8-bit unsigned integer (0 to 128). This field represents the number of valid leading bits in the prefix.

ASM_PREFIX64:

this field identifies the IPv6 multicast prefix to be used to synthesize the IPv4-embedded IPv6 addresses of the multicast groups in the ASM mode. It is a variable size field with the length of the field defined by the asm-length field and is rounded up to the nearest octet boundary. In such case any additional padding bits must be zeroed. The conveyed multicast IPv6 prefix MUST belong to the ASM range. This prefix is likely to be a /96.

ssm-length:

the prefix-length for the SSM IPv4-embedded prefix, as an 8-bit unsigned integer (0 to 128). This field represents the number of valid leading bits in the prefix.

SSM_PREFIX64:

this field identifies the IPv6 multicast prefix to be used to synthesize the IPv4-embedded IPv6 addresses of the multicast groups in the SSM mode. It is a variable size field with the length of the field defined by the ssm-length field and is rounded up to the nearest octet boundary. In such case any additional padding bits must be zeroed. The conveyed multicast IPv6 prefix MUST belong to the SSM range. This prefix is likely to be a /96.

unicast-length:

the prefix-length for the IPv6 unicast prefix to be used to synthesize the IPv4-embedded IPv6 addresses of the multicast sources, as an 8-bit unsigned integer (0 to 128). This field represents the number of valid leading bits in the prefix.

U_PREFIX64:

this field identifies the IPv6 unicast prefix to be used in SSM mode for constructing the IPv4-embedded IPv6 addresses representing the IPv4 multicast sources in the IPv6 domain. U_PREFIX64 may also be used to extract the IPv4 address from the received multicast data flows. It is a variable size field with the length of the field defined by the unicast-length field and is rounded up to the nearest octet boundary. In such case any additional padding bits must be zeroed. The address mapping MUST follow the guidelines documented in [RFC6052].

5. Table of Attributes

The following tables provide a guide to which attributes may be found in which kinds of packets, and in what quantity.

The following table defines the meaning of the above table entries.

Access-Request	Access-Accept	Access-Reject	Challenge	Accounting-Request	#	Attribute
0-1	0-1	0	0	0-1	145	Multicast-Prefixes-64

CoA-Request	CoA-ACK	CoA-NACK	#	Attribute
0-1	0	0	145	Multicast-Prefixes-64

0 This attribute MUST NOT be present in the packet.

0+ Zero or more instances of this attribute MAY be present in the packet.

0-1 Zero or one instances of this attribute MAY be present in the packet.

1 Exactly one instances of this attribute MAY be present in the packet.

6. Security Considerations

This document has no additional security considerations beyond those already identified in [RFC2865] for the RADIUS protocol and in [RFC5176] for CoA messages.

The security considerations documented in [RFC3315] and [RFC6052] are to be considered.

7. IANA Considerations

Per this document, IANA has allocated a new RADIUS attribute type from the IANA registry "Radius Attribute Types" located at <http://www.iana.org/assignments/radius-types>.

Multicast-Prefixes-64 - 145

8. Acknowledgments

The authors would like to thank Ian Farrer, Chongfen Xie, Qi Sun, Linhui Sun and Hao Wang for their contributions to this work.

9. Normative References

- [I-D.ietf-softwire-dslite-multicast]
Qin, J., Boucadair, M., Jacquenet, C., Lee, Y., and Q. Wang, "Delivery of IPv4 Multicast Services to IPv4 Clients over an IPv6 Multicast Network", draft-ietf-softwire-dslite-multicast-09 (work in progress), March 2015.
- [I-D.ietf-softwire-multicast-prefix-option]
Boucadair, M., Qin, J., Tsou, T., and X. Deng, "DHCPv6 Option for IPv4-Embedded Multicast and Unicast IPv6 Prefixes", draft-ietf-softwire-multicast-prefix-option-08 (work in progress), March 2015.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
- [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.
- [RFC5176] Chiba, M., Dommety, G., Eklund, M., Mitton, D., and B. Aboba, "Dynamic Authorization Extensions to Remote Authentication Dial In User Service (RADIUS)", RFC 5176, January 2008.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6519] Maglione, R. and A. Durand, "RADIUS Extensions for Dual-Stack Lite", RFC 6519, February 2012.

Authors' Addresses

Qian Wang
China Telecom
No.118, Xizhimennei
Beijing 100035
China

Email: wangqian@ctbri.com.cn

Wei Meng
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China

Email: meng.wei2@zte.com.cn, vally.meng@gmail.com

Cui Wang
ZTE Corporation
No.50 Software Avenue, Yuhuatai District
Nanjing
China

Email: wang.cuil@zte.com.cn

Mohamed Boucadair
France Telecom
Rennes, 35000
France

Email: mohamed.boucadair@orange.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 10, 2015

O. Troan, Ed.
W. Dec
Cisco Systems
X. Li
C. Bao
CERNET Center/Tsinghua University
S. Matsushima
SoftBank Telecom
T. Murakami
IP Infusion
T. Taylor, Ed.
Huawei Technologies
March 09, 2015

Mapping of Address and Port with Encapsulation (MAP)
draft-ietf-softwire-map-13

Abstract

This document describes a mechanism for transporting IPv4 packets across an IPv6 network using IP encapsulation, and a generic mechanism for mapping between IPv6 addresses and IPv4 addresses and transport layer ports.

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 10, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions	4
3. Terminology	4
4. Architecture	5
5. Mapping Algorithm	7
5.1. Port mapping algorithm	8
5.2. Basic mapping rule (BMR)	10
5.3. Forwarding mapping rule (FMR)	12
5.4. Destinations outside the MAP domain	13
6. The IPv6 Interface Identifier	13
7. MAP Configuration	14
7.1. MAP CE	14
7.2. MAP BR	15
8. Forwarding Considerations	15
8.1. Receiving Rules	15
8.2. ICMP	16
8.3. Fragmentation and Path MTU Discovery	17
8.3.1. Fragmentation in the MAP domain	17
8.3.2. Receiving IPv4 Fragments on the MAP domain borders	17
8.3.3. Sending IPv4 fragments to the outside	18
9. NAT44 Considerations	18
10. IANA Considerations	18
11. Security Considerations	18
12. Contributors	19
13. Acknowledgments	20
14. References	20
14.1. Normative References	20
14.2. Informative References	21
Appendix A. Examples	23
Appendix B. A More Detailed Description of the Derivation of the Port Mapping Algorithm	27
B.1. Bit Representation of the Algorithm	29
B.2. GMA examples	30
Authors' Addresses	30

1. Introduction

Mapping of IPv4 addresses in IPv6 addresses has been described in numerous mechanisms dating back to 1995 [RFC1933]. The Automatic tunneling mechanism described in RFC1933 assigned a globally unique IPv6 address to a host by combining the host's IPv4 address with a well-known IPv6 prefix. Given an IPv6 packet with a destination address with an embedded IPv4 address, a node could automatically tunnel this packet by extracting the IPv4 tunnel end-point address from the IPv6 destination address.

There are numerous variations of this idea, described in 6over4 [RFC2529], 6to4 [RFC3056], ISATAP [RFC5214], and 6rd [RFC5969].

The commonalities of all these IPv6 over IPv4 mechanisms are:

- o Automatically provisions an IPv6 address for a host or an IPv6 prefix for a site
- o Algorithmic or implicit address resolution of tunnel end point addresses. Given an IPv6 destination address, an IPv4 tunnel endpoint address can be calculated.
- o Embedding of an IPv4 address or part thereof into an IPv6 address.

In later phases of IPv4 to IPv6 migration, it is expected that IPv6-only networks will be common, while there will still be a need for residual IPv4 deployment. This document describes a generic mapping of IPv4 to IPv6, and a mechanism for encapsulating IPv4 over IPv6.

Just as the IPv6 over IPv4 mechanisms referred to above, the residual IPv4 over IPv6 mechanism must be capable of:

- o Provisioning an IPv4 prefix, an IPv4 address or a shared IPv4 address.
- o Algorithmically map between either an IPv4 prefix, an IPv4 address or a shared IPv4 address and an IPv6 address.

The mapping scheme described here supports encapsulation of IPv4 packets in IPv6 in both mesh and hub-and-spoke topologies, including address mappings with full independence between IPv6 and IPv4 addresses.

This document describes delivery of IPv4 unicast service across an IPv6 infrastructure. IPv4 multicast is not considered further in this document.

The A+P (Address and Port) architecture of sharing an IPv4 address by distributing the port space is described in [RFC6346]. Specifically section 4 of [RFC6346] covers stateless mapping. The corresponding stateful solution DS-lite is described in [RFC6333]. The motivation for this work is described in [I-D.ietf-softwire-stateless-4v6-motivation].

A companion document defines a DHCPv6 option for provisioning of MAP [I-D.ietf-softwire-map-dhcp]. Other means of provisioning are possible. Deployment considerations are described in [I-D.ietf-softwire-map-deployment].

MAP relies on IPv6 and is designed to deliver dual-stack service while allowing IPv4 to be phased out within the service provider's (SP) network. The phasing out of IPv4 within the SP network is independent of whether the end user disables IPv4 service or not. Further, "greenfield"; IPv6-only networks may use MAP in order to deliver IPv4 to sites via the IPv6 network.

2. Conventions

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].

3. Terminology

MAP domain:	One or more MAP CEs and BRs connected to the same virtual link. A service provider may deploy a single MAP domain, or may utilize multiple MAP domains.
MAP rule	A set of parameters describing the mapping between an IPv4 prefix, IPv4 address or shared IPv4 address and an IPv6 prefix or address. Each domain uses a different mapping rule set.
MAP node	A device that implements MAP.
MAP Border Relay (BR):	A MAP enabled router managed by the service provider at the edge of a MAP domain. A Border Relay router has at least an IPv6-enabled interface and an IPv4 interface connected to the native IPv4 network. A MAP BR may also be referred to simply as a "BR" within the context of MAP.

MAP Customer Edge (CE):	A device functioning as a Customer Edge router in a MAP deployment. A typical MAP CE adopting MAP rules will serve a residential site with one WAN side interface, and one or more LAN side interfaces. A MAP CE may also be referred to simply as a "CE" within the context of MAP.
Port-set:	The separate part of the transport layer port space; denoted as a port-set.
Port-set ID (PSID):	Algorithmically identifies a set of ports exclusively assigned to a CE.
Shared IPv4 address:	An IPv4 address that is shared among multiple CEs. Only ports that belong to the assigned port-set can be used for communication. Also known as a Port-Restricted IPv4 address.
End-user IPv6 prefix:	The IPv6 prefix assigned to an End-user CE by other means than MAP itself. E.g., Provisioned using DHCPv6 PD [RFC3633], assigned via SLAAC [RFC4862], or configured manually. It is unique for each CE.
MAP IPv6 address:	The IPv6 address used to reach the MAP function of a CE from other CEs and from BRs.
Rule IPv6 prefix:	An IPv6 prefix assigned by a Service Provider for a mapping rule.
Rule IPv4 prefix:	An IPv4 prefix assigned by a Service Provider for a mapping rule.
Embedded Address (EA) bits:	The IPv4 EA-bits in the IPv6 address identify an IPv4 prefix/address (or part thereof) or a shared IPv4 address (or part thereof) and a port-set identifier.

4. Architecture

In accordance with the requirements stated above, the MAP mechanism can operate with shared IPv4 addresses, full IPv4 addresses or IPv4 prefixes. Operation with shared IPv4 addresses is described here, and the differences for full IPv4 addresses and prefixes are described below.

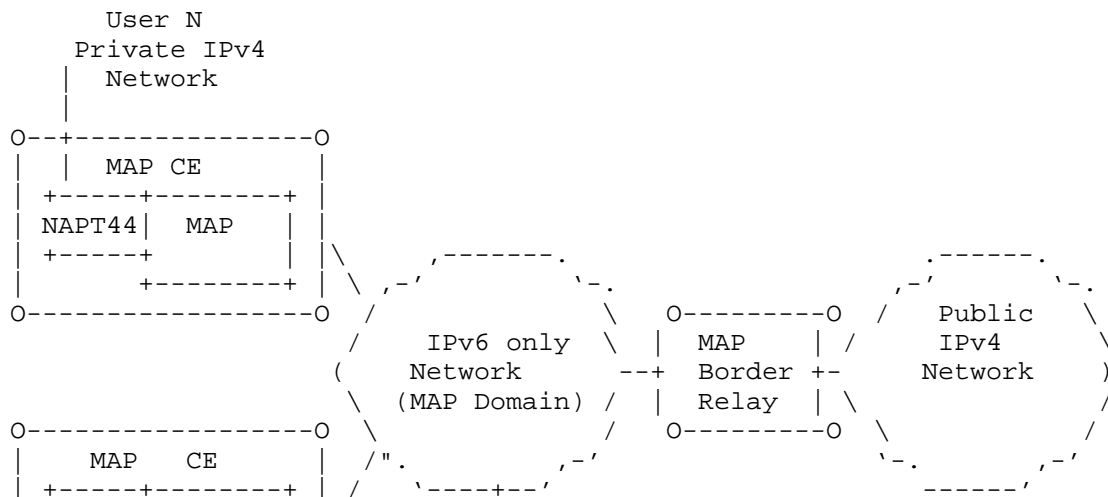
The MAP mechanism uses existing standard building blocks. The existing NAT [RFC2663] on the CE is used with additional support for restricting transport protocol ports, ICMP identifiers and fragment identifiers to the configured port-set. For packets outbound from the private IPv4 network, the CE NAT MUST translate transport identifiers (e.g., TCP and UDP port numbers) so that they fall within the CE's assigned port-range.

The NAT MUST in turn be connected to a MAP-aware forwarding function, that does encapsulation / decapsulation of IPv4 packets in IPv6. MAP supports the encapsulation mode specified in [RFC2473]. In addition MAP specifies an algorithm to do "address resolution" from an IPv4 address and port to an IPv6 address. This algorithmic mapping is specified in Section 5.

The MAP architecture described here restricts the use of the shared IPv4 address to only be used as the global address (outside) of the NAT running on the CE. A shared IPv4 address MUST NOT be used to identify an interface. While it is theoretically possible to make host stacks and applications port-aware, it would be a drastic change to the IP model [RFC6250].

For full IPv4 addresses and IPv4 prefixes, the architecture just described applies with two differences. First, a full IPv4 address or IPv4 prefix can be used as it is today, e.g., for identifying an interface or as a DHCP pool, respectively. Secondly, the NAT is not required to restrict the ports used on outgoing packets.

This architecture is illustrated in Figure 1.



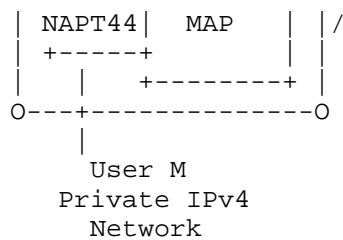


Figure 1: Network Topology

The MAP BR connects one or more MAP domains to external IPv4 networks.

5. Mapping Algorithm

A MAP node is provisioned with one or more mapping rules.

Mapping rules are used differently depending on their function. Every MAP node must be provisioned with a Basic mapping rule. This is used by the node to configure its IPv4 address, IPv4 prefix or shared IPv4 address. This same basic rule can also be used for forwarding, where an IPv4 destination address and optionally a destination port are mapped into an IPv6 address. Additional mapping rules are specified to allow for multiple different IPv4 sub-nets to exist within the domain and optimize forwarding between them.

Traffic outside of the domain (i.e., when the destination IPv4 address does not match (using longest matching prefix) any Rule IPv4 prefix in the Rules database) is forwarded to the BR.

There are two types of mapping rules:

1. Basic Mapping Rule (BMR) - mandatory. A CE can be provisioned with multiple End-user IPv6 prefixes. There can only be one Basic Mapping Rule per End-user IPv6 prefix. However all CE's having End-user IPv6 prefixes within (aggregated by) the same Rule IPv6 prefix may share the same Basic Mapping Rule. In combination with the End-user IPv6 prefix, the Basic Mapping Rule is used to derive the IPv4 prefix, address, or shared address and the PSID assigned to the CE.
2. Forwarding Mapping Rule (FMR) - optional, used for forwarding. The Basic Mapping Rule may also be a Forwarding Mapping Rule. Each Forwarding Mapping Rule will result in an entry in the Rules table for the Rule IPv4 prefix. Given a destination IPv4 address and port within the MAP domain, a MAP node can use the matching

FMR to derive the End-user IPv6 address of the interface through which that IPv4 destination address and port combination can be reached. In hub and spoke mode there are no FMRs.

Both mapping rules share the same parameters:

- o Rule IPv6 prefix (including prefix length)
- o Rule IPv4 prefix (including prefix length)
- o Rule EA-bits length (in bits)

A MAP node finds its BMR by doing a longest match between the End-user IPv6 prefix and the Rule IPv6 prefix in the Mapping Rules table. The rule is then used for IPv4 prefix, address or shared address assignment.

A MAP IPv6 address is formed from the BMR Rule IPv6 prefix. This address MUST be assigned to an interface of the MAP node and is used to terminate all MAP traffic being sent or received to the node.

Port-restricted IPv4 routes are installed in the Rules table for all the Forwarding Mapping Rules, and a default route is installed to the MAP BR (see Section 5.4).

Forwarding Mapping Rules are used to allow direct communication between MAP CEs, known as mesh mode. In hub and spoke mode, there are no forwarding mapping rules, all traffic MUST be forwarded directly to the BR.

While an FMR is optional in the sense that a MAP CE MAY be configured with zero or more FMRs depending on the deployment, all MAP CEs MUST implement support for both rule types.

5.1. Port mapping algorithm

The port mapping algorithm is used in domains whose rules allow IPv4 address sharing.

The simplest way to represent a port range is using a notation similar to CIDR [RFC4632]. For example the first 256 ports are represented as port prefix 0.0/8. The last 256 ports as 255.0/8. In hexadecimal, 0x0000/8 (PSID = 0) and 0xFF00/8 (PSID = 0xFF). Using this technique, but wishing to avoid allocating the system ports [RFC6335] to the user, one would have to exclude the use of one or more PSIDs (e.g., PSIDs 0 to 3 in the example just given).

When the PSID is embedded in the End-user IPv6 prefix, then to minimize dependencies between the End-user IPv6 prefix and the assigned port-set, it is desirable to minimize the restrictions of possible PSID values. This is achieved by using an infix representation of the port value. Using such a representation, the well-known ports are excluded by restrictions on the value of the high-order bitfield (A) rather than the PSID.

The infix algorithm allocates ports to a given CE as a series of contiguous ranges spaced at regular intervals throughout the complete range of possible port-set values.

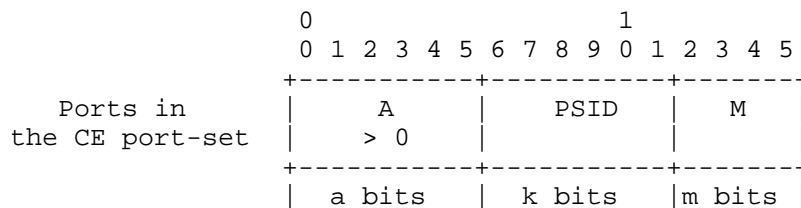


Figure 2: Structure of a port-restricted port field

a bits: The number of offset bits. 6 by default as this excludes the system ports (0-1023). To guarantee non-overlapping port sets, the offset 'a' MUST be the same for every MAP CE sharing the same address.

A: Selects the range of the port number. For 'a' > 0, A MUST be larger than 0. This ensures that the algorithm excludes the system ports. For the default value of 'a' (6), the system ports, are excluded by requiring that A be greater than 0. Smaller values of 'a' excludes a larger initial range. E.g., 'a' = 4, will exclude ports 0 - 4095. The interval between initial port numbers of successive contiguous ranges assigned to the same user is $2^{(16-a)}$.

k bits: The length in bits of the PSID field. To guarantee non-overlapping port sets, the length 'k' MUST be the same for every MAP CE sharing the same address. The sharing ratio is 2^k . The number of ports assigned to the user is $2^{(16-k)} - 2^m$ (excluded ports)

PSID: The Port-Set Identifier (PSID). Different PSID values guarantee non-overlapping port-sets thanks to the restrictions on 'a' and 'k' stated above, because the PSID always occupies the same bit positions in the port number.

m bits: The number of contiguous ports is given by 2^m .

M: Selects the specific port within a particular range specified by the concatenation of A and the PSID.

5.2. Basic mapping rule (BMR)

The Basic Mapping Rule is mandatory, used by the CE to provision itself with an IPv4 prefix, IPv4 address or shared IPv4 address. Recall from Section 5 that the BMR consists of the following parameters:

- o Rule IPv6 prefix (including prefix length)
- o Rule IPv4 prefix (including prefix length)
- o Rule EA-bits length (in bits)

Figure 3 shows the structure of the complete MAP IPv6 address as specified in this document.

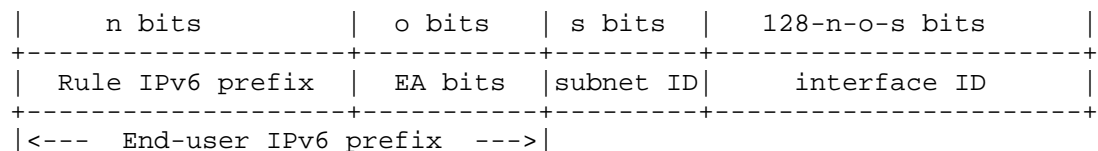


Figure 3: MAP IPv6 Address Format

The Rule IPv6 prefix (which is part of the End-user IPv6 prefix) that is common among all CEs using the same Basic Mapping Rule within the MAP domain. The EA bits encode the CE specific IPv4 address and port information. The EA bits, which are unique for a given Rule IPv6 prefix, can contain a full or part of an IPv4 address and, in the shared IPv4 address case, a Port-Set Identifier (PSID). An EA-bit length of 0 signifies that all relevant MAP IPv4 addressing information is passed directly in the BMR, and not derived from the End-user IPv6 prefix.

The MAP IPv6 address is created by concatenating the End-user IPv6 prefix with the MAP subnet identifier (if the End-user IPv6 prefix is shorter than 64 bits) and the interface identifier as specified in Section 6.

The MAP subnet identifier is defined to be the first subnet (s bits set to zero).

Define:

r = length of the IPv4 prefix given by the BMR;

o = length of the EA bit field as given by the BMR;

p = length of the IPv4 suffix contained in the EA bit field.

The length r MAY be zero, in which case the complete IPv4 address or prefix is encoded in the EA bits. If only a part of the IPv4 address / prefix is encoded in the EA bits, the Rule IPv4 prefix is provisioned to the CE by other means (e.g., a DHCPv6 option). To create a complete IPv4 address (or prefix), the IPv4 address suffix (p) from the EA bits, is concatenated with the Rule IPv4 prefix (r bits).

The offset of the EA bits field in the IPv6 address is equal to the BMR Rule IPv6 prefix length. The length of the EA bits field (o) is given by the BMR Rule EA-bits length, and can be between 0 and 48. A length of 48 means that the complete IPv4 address and port is embedded in the End-user IPv6 prefix (a single port is assigned). A length of 0 means that no part of the IPv4 address or port is embedded in the address. The sum of the Rule IPv6 Prefix length and the Rule EA-bits length MUST be less or equal than the End-user IPv6 prefix length.

If $o + r < 32$ (length of the IPv4 address in bits), then an IPv4 prefix is assigned. This case is shown in Figure 4.

IPv4 prefix:

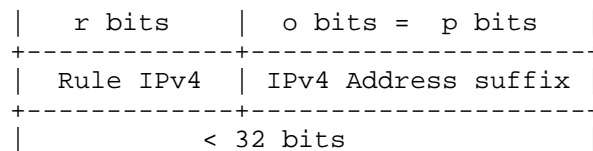


Figure 4: IPv4 prefix

If $o + r$ is equal to 32, then a full IPv4 address is to be assigned. The address is created by concatenating the Rule IPv4 prefix and the EA-bits. This case is shown in Figure 5.

Complete IPv4 address:

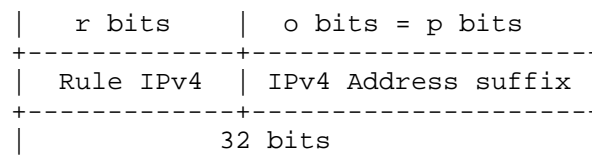


Figure 5: Complete IPv4 address

If $o + r$ is > 32 , then a shared IPv4 address is to be assigned. The number of IPv4 address suffix bits (p) in the EA bits is given by $32 - r$ bits. The PSID bits are used to create a port set. The length of the PSID bit field within EA bits is: $q = o - p$.

Shared IPv4 address:

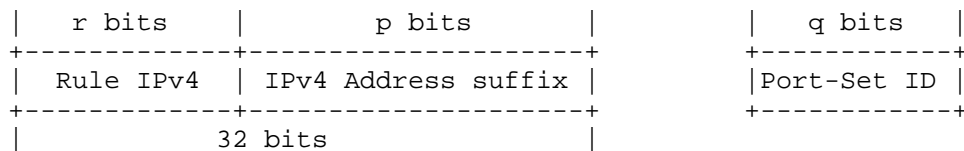


Figure 6: Shared IPv4 address

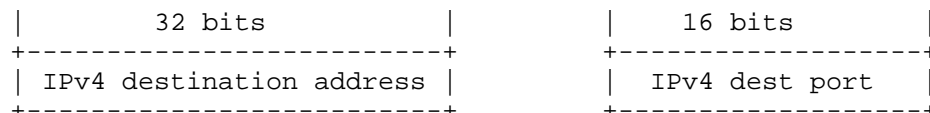
The length of r MAY be 32, with no part of the IPv4 address embedded in the EA bits. This results in a mapping with no dependence between the IPv4 address and the IPv6 address. In addition the length of o MAY be zero (no EA bits embedded in the End-User IPv6 prefix), meaning that also the PSID is provisioned using e.g., the DHCP option.

See Appendix A for an example of the Basic Mapping Rule.

5.3. Forwarding mapping rule (FMR)

The Forwarding Mapping Rule is optional, and used in mesh mode to enable direct CE to CE connectivity.

On adding an FMR rule, an IPv4 route is installed in the Rules table for the Rule IPv4 prefix.



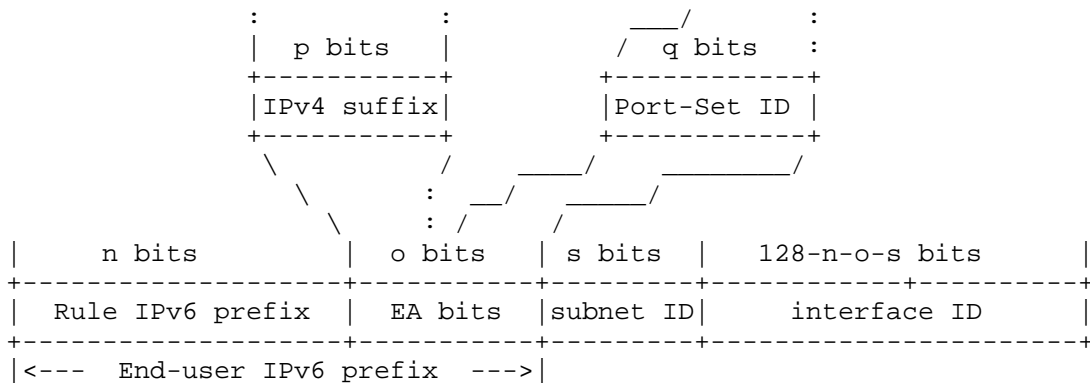


Figure 7: Derivation of MAP IPv6 address

See Appendix A for an example of the Forwarding Mapping Rule.

5.4. Destinations outside the MAP domain

IPv4 traffic between MAP nodes that are all within one MAP domain is encapsulated in IPv6, with the sender's MAP IPv6 address as the IPv6 source address and the receiving MAP node's MAP IPv6 address as the IPv6 destination address. To reach IPv4 destinations outside of the MAP domain, traffic is also encapsulated in IPv6, but the destination IPv6 address is set to the configured IPv6 address of the MAP BR.

On the CE, the path to the BR can be represented as a point to point IPv4 over IPv6 tunnel [RFC2473] with the source address of the tunnel being the CE's MAP IPv6 address and the BR IPv6 address as the remote tunnel address. When MAP is enabled, a typical CE router will install a default IPv4 route to the BR.

The BR forwards traffic received from the outside to CE's using the normal MAP forwarding rules.

6. The IPv6 Interface Identifier

The Interface identifier format of a MAP node is described below.

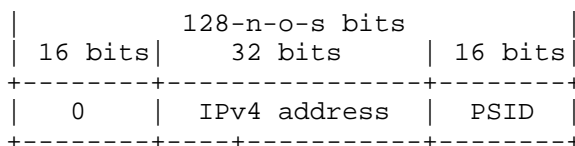


Figure 8

In the case of an IPv4 prefix, the IPv4 address field is right-padded with zeroes up to 32 bits. The PSID field is left-padded to create a 16 bit field. For an IPv4 prefix or a complete IPv4 address, the PSID field is zero.

If the End-user IPv6 prefix length is larger than 64, the most significant parts of the interface identifier is overwritten by the prefix.

7. MAP Configuration

For a given MAP domain, the BR and CE MUST be configured with the following MAP elements. The configured values for these elements are identical for all CEs and BRs within a given MAP domain.

- o The Basic Mapping Rule and optionally the Forwarding Mapping Rules, including the Rule IPv6 prefix, Rule IPv4 prefix, and Length of EA bits
- o Hub and spoke mode or Mesh mode. (If all traffic should be sent to the BR, or if direct CE to CE traffic should be supported).

In addition the MAP CE MUST be configured with the IPv6 address(es) of the MAP BR (Section 5.4).

7.1. MAP CE

The MAP elements are set to values that are the same across all CEs within a MAP domain. The values may be configured in a variety of manners, including provisioning methods such as the Broadband Forum's "TR-69" Residential Gateway management interface, an XML-based object retrieved after IPv6 connectivity is established, or manual configuration by an administrator. IPv6 DHCP options for MAP configuration is defined in [I-D.ietf-softwire-map-dhcp]. Other configuration and management methods may use the format described by this option for consistency and convenience of implementation on CEs that support multiple configuration methods.

The only remaining provisioning information the CE requires in order to calculate the MAP IPv4 address and enable IPv4 connectivity is the IPv6 prefix for the CE. The End-user IPv6 prefix is configured as part of obtaining IPv6 Internet access.

The MAP provisioning parameters, and hence the IPv4 service itself, are tied to the associated End-user IPv6 prefix lifetime; thus, the MAP service is also tied to this in terms of authorization, accounting, etc.

A single MAP CE MAY be connected to more than one MAP domain, just as any router may have more than one IPv4-enabled service provider facing interface and more than one set of associated addresses assigned by DHCP. Each domain a given CE operates within would require its own set of MAP configuration elements and would generate its own IPv4 address. Each MAP domain requires a distinct End-user IPv6 prefix.

The MAP DHCP option is specified in [I-D.ietf-softwire-map-dhcp].

7.2. MAP BR

The MAP BR MUST be configured with corresponding mapping rules for each MAP domain which it is acting as BR for.

For increased reliability and load balancing, the BR IPv6 address MAY be an anycast address shared across a given MAP domain. As MAP is stateless, any BR may be used at any time. If the BR IPv6 address is anycast the relay MUST use this anycast IPv6 address as the source address in packets relayed to CEs.

Since MAP uses provider address space, no specific routes need to be advertised externally for MAP to operate, neither in IPv6 nor IPv4 BGP. However, if anycast is used for the MAP IPv6 relays, the anycast addresses must be advertised in the service provider's IGP.

8. Forwarding Considerations

Figure 1 depicts the overall MAP architecture with IPv4 users (N and M) networks connected to a routed IPv6 network.

MAP uses Encapsulation mode as specified in [RFC2473].

For a shared IPv4 address, a MAP CE forwarding IPv4 packets from the LAN performs NAT44 functions first and creates appropriate NAT44 bindings. The resulting IPv4 packets MUST contain the source IPv4 address and source transport identifiers specified by the MAP provisioning parameters. The IPv4 packet is forwarded using the CE's MAP forwarding function. The IPv6 source and destination addresses MUST then be derived as per Section 5 of this draft.

8.1. Receiving Rules

A MAP CE receiving an IPv6 packet to its MAP IPv6 address sends this packet to the CE's MAP function where it is decapsulated. The resulting IPv4 packet is then forwarded to the CE's NAT44 function where it is handled according to the NAT's translation table.

A MAP BR receiving IPv6 packets selects a best matching MAP domain rule (Rule IPv6 prefix) based on a longest address match of the packet's IPv6 source address, as well as a match of the packet destination address against the configured BR IPv6 address(es). The selected MAP rule allows the BR to determine the EA-bits from the source IPv6 address.

To prevent spoofing of IPv4 addresses, any MAP node (CE and BR) MUST perform the following validation upon reception of a packet. First, the embedded IPv4 address or prefix, as well as PSID (if any), are extracted from the source IPv6 address using the matching MAP rule. These represent the range of what is acceptable as source IPv4 address and port. Secondly, the node extracts the source IPv4 address and port from the IPv4 packet encapsulated inside the IPv6 packet. If they are found to be outside the acceptable range, the packet MUST be silently discarded and a counter incremented to indicate that a potential spoofing attack may be underway. The source validation checks just described are not done for packets whose source IPv6 address is that of the BR (BR IPv6 address).

By default, the CE router MUST drop packets received on the MAP virtual interface (i.e., after decapsulation of IPv6) for IPv4 destinations not for its own IPv4 shared address, full IPv4 address or IPv4 prefix.

8.2. ICMP

ICMP message should be supported in MAP domain. Hence, the NAT44 in MAP CE MUST implement the behavior for ICMP message conforming to the best current practice documented in [RFC5508].

If a MAP CE receives an ICMP message having ICMP identifier field in ICMP header, NAT44 in the MAP CE MUST rewrite this field to a specific value assigned from the port set. BR and other CEs must handle this field similar to the port number in the TCP/UDP header upon receiving the ICMP message with ICMP identifier field.

If a MAP node receives an ICMP error message without the ICMP identifier field for errors that is detected inside a IPv6 tunnel, a node should relay the ICMP error message to the original source. This behavior SHOULD be implemented conforming to the section 8 of [RFC2473].

8.3. Fragmentation and Path MTU Discovery

Due to the different sizes of the IPv4 and IPv6 header, handling the maximum packet size is relevant for the operation of any system connecting the two address families. There are three mechanisms to handle this issue: Path MTU discovery (PMTUD), fragmentation, and transport-layer negotiation such as the TCP Maximum Segment Size (MSS) option [RFC0897]. MAP uses all three mechanisms to deal with different cases.

8.3.1. Fragmentation in the MAP domain

Encapsulating an IPv4 packet to carry it across the MAP domain will increase its size (typically by 40 bytes). It is strongly recommended that the MTU in the MAP domain be well managed and that the IPv6 MTU on the CE WAN side interface be set so that no fragmentation occurs within the boundary of the MAP domain.

Fragmentation on MAP domain entry is described in section 7.2 of [RFC2473].

The use of an anycast source address could lead to an ICMP error message generated on the path being sent to a different BR. Therefore, using dynamic tunnel MTU Section 6.7 of [RFC2473] is subject to IPv6 Path MTU black-holes. A MAP BR using an anycast source address SHOULD NOT by default use Path MTU discovery across the MAP domain.

Multiple BRs using the same anycast source address could send fragmented packets to the same CE at the same time. If the fragmented packets from different BRs happen to use the same fragment ID, incorrect reassembly might occur. See [RFC4459] for an analysis of the problem. Section 3.4 suggests solving the problem by fragmenting the inner packet.

8.3.2. Receiving IPv4 Fragments on the MAP domain borders

Forwarding of an IPv4 packet received from the outside of the MAP domain requires the IPv4 destination address and the transport protocol destination port. The transport protocol information is only available in the first fragment received. As described in section 5.3.3 of [RFC6346] a MAP node receiving an IPv4 fragmented packet from outside has to reassemble the packet before sending the packet onto the MAP link. If the first packet received contains the transport protocol information, it is possible to optimize this behavior by using a cache and forwarding the fragments unchanged. Implementers of MAP should be aware that there are a number of well-known attacks against IP fragmentation; see [RFC1858] and [RFC3128].

Implementers should also be aware of additional issues with reassembling packets at high rates, as described in [RFC4963].

8.3.3. Sending IPv4 fragments to the outside

If two IPv4 hosts behind two different MAP CEs with the same IPv4 address send fragments to an IPv4 destination host outside the domain, those hosts may use the same IPv4 fragmentation identifier, resulting in incorrect reassembly of the fragments at the destination host. Given that the IPv4 fragmentation identifier is a 16 bit field, it could be used similarly to port ranges. A MAP CE could rewrite the IPv4 fragmentation identifier to be within its allocated port-set, if the resulting fragment identifier space was large enough related to the rate fragments were sent. However, splitting the identifier space in this fashion would increase the probability of reassembly collision for all connections through the CPE. See also [RFC6864]

9. NAT44 Considerations

The NAT44 implemented in the MAP CE SHOULD conform with the behavior and best current practice documented in [RFC4787], [RFC5508], and [RFC5382]. In MAP address sharing mode (determined by the MAP domain/rule configuration parameters) the operation of the NAT44 MUST be restricted to the available port numbers derived via the basic mapping rule.

10. IANA Considerations

This specification does not require any IANA actions.

11. Security Considerations

Spoofing attacks: With consistency checks between IPv4 and IPv6 sources that are performed on IPv4/IPv6 packets received by MAP nodes, MAP does not introduce any new opportunity for spoofing attacks that would not already exist in IPv6.

Denial-of-service attacks: In MAP domains where IPv4 addresses are shared, the fact that IPv4 datagram reassembly may be necessary introduces an opportunity for DOS attacks. This is inherent to address sharing, and is common with other address sharing approaches such as DS-Lite and NAT64/DNS64. The best protection against such attacks is to accelerate IPv6 deployment, so that, where MAP is supported, it is less and less used.

Routing-loop attacks: This attack may exist in some automatic tunneling scenarios are documented in [RFC6324]. They cannot

exist with MAP because each BRs checks that the IPv6 source address of a received IPv6 packet is a CE address based on Forwarding Mapping Rule.

Attacks facilitated by restricted port set: From hosts that are not subject to ingress filtering of [RFC2827], some attacks are possible by an attacker injecting spoofed packets during ongoing transport connections ([RFC4953], [RFC5961], [RFC6056]. The attacks depend on guessing which ports are currently used by target hosts, and using an unrestricted port-set is preferable, i.e., using native IPv6 connections that are not subject to MAP port range restrictions. To minimize this type of attacks when using a restricted port-set, the MAP CE's NAT44 filtering behavior SHOULD be "Address-Dependent Filtering" [RFC4787], Section 5. Furthermore, the MAP CEs SHOULD use a DNS transport proxy [RFC5625] function to handle DNS traffic, and source such traffic from IPv6 interfaces not assigned to MAP.

[RFC6269] outlines general issues with IPv4 address sharing.

12. Contributors

This document is the result of the IETF Softwire MAP design team effort and numerous previous individual contributions in this area:

Chongfeng Xie (China Telecom)
Room 708, No.118, Xizhimennei Street Beijing 100035
People's Republic of China
Phone: +86-10-58552116
Email: xiechf@ctbri.com.cn

Qiong Sun (China Telecom)
Room 708, No.118, Xizhimennei Street Beijing 100035
People's Republic of China
Phone: +86-10-58552936
Email: sunqiong@ctbri.com.cn

Gang Chen (China Mobile)
53A, Xibianmennei Ave. Beijing 100053
People's Republic of China
Email: chengang@chinamobile.com

Yu Zhai
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University

Beijing 100084
People's Republic of China
Email: jacky.zhai@gmail.com

Wentao Shang (CERNET Center/Tsinghua University)
Room 225, Main Building, Tsinghua University Beijing 100084
People's Republic of China
Email: wentaoshang@gmail.com

Guoliang Han (CERNET Center/Tsinghua University)
Room 225, Main Building, Tsinghua University Beijing 100084
People's Republic of China
Email: bupthgl@gmail.com

Rajiv Asati (Cisco Systems)
7025-6 Kit Creek Road Research Triangle Park NC 27709 USA
Email: rajiva@cisco.com

13. Acknowledgments

This document is based on the ideas of many, including Masakazu Asama, Mohamed Boucadair, Gang Chen, Maoke Chen, Wojciech Dec, Xiaohong Deng, Jouni Korhonen, Tomasz Mrugalski, Jacni Qin, Chunfa Sun, Qiong Sun, and Leaf Yeh. The authors want in particular to recognize Remi Despres, who has tirelessly worked on generalized mechanisms for stateless address mapping.

The authors would like to thank Lichun Bao, Guillaume Gottard, Dan Wing, Jan Zorz, Necj Scoberne, Tina Tsou, Kristian Poscic, and especially Tom Taylor and Simon Perreault for the thorough review and comments of this document. Useful IETF Last Call comments were received from Brian Weis and Lei Yan.

14. References

14.1. Normative References

- [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.

- [RFC5625] Bellis, R., "DNS Proxy Implementation Guidelines", BCP 152, RFC 5625, August 2009.

14.2. Informative References

- [I-D.ietf-softwire-map-deployment]
Qiong, Q., Chen, M., Chen, G., Tsou, T., and S. Perreault,
"Mapping of Address and Port (MAP) - Deployment
Considerations", draft-ietf-softwire-map-deployment-03
(work in progress), October 2013.
- [I-D.ietf-softwire-map-dhcp]
Mrugalski, T., Troan, O., Dec, W., Bao, C.,
leaf.yeh.sdo@gmail.com, l., and X. Deng, "DHCPv6 Options
for configuration of Softwire Address and Port Mapped
Clients", draft-ietf-softwire-map-dhcp-06 (work in
progress), November 2013.
- [I-D.ietf-softwire-stateless-4v6-motivation]
Boucadair, M., Matsushima, S., Lee, Y., Bonness, O.,
Borges, I., and G. Chen, "Motivations for Carrier-side
Stateless IPv4 over IPv6 Migration Solutions", draft-ietf-
softwire-stateless-4v6-motivation-05 (work in progress),
November 2012.
- [RFC0897] Postel, J., "Domain name system implementation schedule",
RFC 897, February 1984.
- [RFC1858] Ziemba, G., Reed, D., and P. Traina, "Security
Considerations for IP Fragment Filtering", RFC 1858,
October 1995.
- [RFC1933] Gilligan, R. and E. Nordmark, "Transition Mechanisms for
IPv6 Hosts and Routers", RFC 1933, April 1996.
- [RFC2529] Carpenter, B. and C. Jung, "Transmission of IPv6 over IPv4
Domains without Explicit Tunnels", RFC 2529, March 1999.
- [RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address
Translator (NAT) Terminology and Considerations", RFC
2663, August 1999.
- [RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering:
Defeating Denial of Service Attacks which employ IP Source
Address Spoofing", BCP 38, RFC 2827, May 2000.
- [RFC3056] Carpenter, B. and K. Moore, "Connection of IPv6 Domains
via IPv4 Clouds", RFC 3056, February 2001.

- [RFC3128] Miller, I., "Protection Against a Variant of the Tiny Fragment Attack (RFC 1858)", RFC 3128, June 2001.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC4459] Savola, P., "MTU and Fragmentation Issues with In-the-Network Tunneling", RFC 4459, April 2006.
- [RFC4632] Fuller, V. and T. Li, "Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan", BCP 122, RFC 4632, August 2006.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC4953] Touch, J., "Defending TCP Against Spoofing Attacks", RFC 4953, July 2007.
- [RFC4963] Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly Errors at High Data Rates", RFC 4963, July 2007.
- [RFC5214] Templin, F., Gleeson, T., and D. Thaler, "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)", RFC 5214, March 2008.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC5508] Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT Behavioral Requirements for ICMP", BCP 148, RFC 5508, April 2009.
- [RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's Robustness to Blind In-Window Attacks", RFC 5961, August 2010.
- [RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.

- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-Protocol Port Randomization", BCP 156, RFC 6056, January 2011.
- [RFC6250] Thaler, D., "Evolution of the IP Model", RFC 6250, May 2011.
- [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.
- [RFC6324] Nakibly, G. and F. Templin, "Routing Loop Attack Using IPv6 Automatic Tunnels: Problem Statement and Proposed Mitigations", RFC 6324, August 2011.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.
- [RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", BCP 165, RFC 6335, August 2011.
- [RFC6346] Bush, R., "The Address plus Port (A+P) Approach to the IPv4 Address Shortage", RFC 6346, August 2011.
- [RFC6864] Touch, J., "Updated Specification of the IPv4 ID Field", RFC 6864, February 2013.

Appendix A. Examples

Example 1 - Basic Mapping Rule

Given the MAP domain information and an IPv6 address of an endpoint:

```
End-user IPv6 prefix: 2001:db8:0012:3400::/56
Basic Mapping Rule:  {2001:db8:0000::/40 (Rule IPv6 prefix),
                      192.0.2.0/24 (Rule IPv4 prefix),
                      16 (Rule EA-bits length)}
PSID length:         (16 - (32 - 24) = 8. (Sharing ratio of 256)
PSID offset:         6 (default)
```

A MAP node (CE or BR) can via the BMR, or equivalent FMR, determine the IPv4 address and port-set as shown below:

```
EA bits offset:      40
IPv4 suffix bits (p) Length of IPv4 address (32) -
                    IPv4 prefix length (24) = 8
IPv4 address:        192.0.2.18 (0xc0000212)
PSID start:          40 + p = 40 + 8 = 48
PSID length:         o - p = (56 - 40) - 8 = 8
PSID:                0x34
```

```
Available ports (63 ranges) : 1232-1235, 2256-2259, ..... ,
                             63696-63699, 64720-64723
```

The BMR information allows a MAP CE to determine (complete) its IPv6 address within the indicated IPv6 prefix.

IPv6 address of MAP CE: 2001:db8:0012:3400:0000:c000:0212:0034

Example 2 - BR:

Another example can be made of a MAP BR, configured with the following FMR when receiving a packet with the following characteristics:

IPv4 source address: 1.2.3.4 (0x01020304)
IPv4 source port: 80
IPv4 destination address: 192.0.2.18 (0xc0000212)
IPv4 destination port: 1232

Forwarding Mapping Rule: {2001:db8::/40 (Rule IPv6 prefix),
192.0.2.0/24 (Rule IPv4 prefix),
16 (Rule EA-bits length)}

IPv6 address of MAP BR: 2001:db8:ffff::1

The above information allows the BR to derive as follows the mapped destination IPv6 address for the corresponding MAP CE, and also the mapped source IPv6 address for the IPv4 source address.

IPv4 suffix bits (p): $32 - 24 = 8$ (18 (0x12))
PSID length: 8
PSID: 0x34 (1232)

The resulting IPv6 packet will have the following key fields:

IPv6 source address: 2001:db8:ffff::1
IPv6 destination address: 2001:db8:0012:3400:0000:c000:0212:0034

Example 3 - Forwarding Mapping Rule:

An IPv4 host behind the MAP CE (addressed as per the previous examples) corresponding with IPv4 host 1.2.3.4 will have its packets encapsulated by IPv6 using the IPv6 address of the BR configured on the MAP CE as follows:

IPv6 address of BR: 2001:db8:ffff::1
IPv4 source address: 192.0.2.18
IPv4 destination address: 1.2.3.4
IPv4 source port: 1232
IPv4 destination port: 80
MAP CE IPv6 source address: 2001:db8:0012:3400:0000:c000:0212:0034
IPv6 destination address: 2001:db8:ffff::1

Example 4 - Rule with no embedded address bits and no address sharing

End-User IPv6 prefix: 2001:db8:0012:3400::/56
Basic Mapping Rule: {2001:db8:0012:3400::/56 (Rule IPv6 prefix),
192.0.2.18/32 (Rule IPv4 prefix),
0 (Rule EA-bits length)}
PSID length: 0 (Sharing ratio is 1)
PSID offset: n/a

A MAP node (CE or BR) can via the BMR or equivalent FMR, determine the IPv4 address and port-set as shown below:

EA bits offset: 0
IPv4 suffix bits (p): Length of IPv4 address (32) -
IPv4 prefix length (32) = 0
IPv4 address: 192.0.2.18 (0xc0000212)
PSID start: 0
PSID length: 0
PSID: null

The BMR information allows a MAP CE also to determine (complete) its full IPv6 address by combining the IPv6 prefix with the MAP interface identifier (that embeds the IPv4 address).

IPv6 address of MAP CE: 2001:db8:0012:3400:0000:c000:0212:0000

Example 5 - Rule with no embedded address bits and address sharing (sharing ratio 256)


```

End-User IPv6 prefix: 2001:db8:0012:3400::/56
Basic Mapping Rule:  {2001:db8:0012:3400::/56 (Rule IPv6 prefix),
                      192.0.2.18/32 (Rule IPv4 prefix),
                      0 (Rule EA-bits length)}
PSID length:         8. (From DHCP. Sharing ratio of 256)
PSID offset:         6 (Default)
PSID :               0x34 (From DHCP.)

```

A MAP node can via the Basic Mapping Rule determine the IPv4 address and port-set as shown below:

```

EA bits offset:      0
IPv4 suffix bits (p): Length of IPv4 address (32) -
                      IPv4 prefix length (32) = 0
IPv4 address:        192.0.2.18 (0xc0000212)
PSID offset:         6
PSID length:         8
PSID:                0x34

```

```

Available ports (63 ranges) : 1232-1235, 2256-2259, ..... ,
                              63696-63699, 64720-64723

```

The Basic Mapping Rule information allows a MAP CE also to determine (complete) its full IPv6 address by combining the IPv6 prefix with the MAP interface identifier (that embeds the IPv4 address and PSID).

IPv6 address of MAP CE: 2001:db8:0012:3400:0000:c000:0212:0034

Note that the IPv4 address and PSID is not derived from the IPv6 prefix assigned to the CE, but provisioned separately using e.g., DHCP.

Appendix B. A More Detailed Description of the Derivation of the Port Mapping Algorithm

This Appendix describes how the port mapping algorithm described in Section 5.1 was derived. The algorithm is used in domains whose rules allow IPv4 address sharing.

The basic requirement for a port mapping algorithm is that the port-sets it assigns to different MAP CEs MUST be non-overlapping. A number of other requirements guided the choice of the algorithm:

- o In keeping with the general MAP algorithm the port-set MUST be derivable from a port-set identifier (PSID) that can be embedded in the End-user IPv6 prefix.
- o The mapping MUST be reversible, such that, given the port number, the PSID of the port-set to which it belongs can be quickly derived.
- o The algorithm MUST allow a broad range of address sharing ratios.
- o It SHOULD be possible to exclude subsets of the complete port numbering space from assignment. Most operators would exclude the system ports (0-1023). A conservative operator might exclude all but the transient ports (49152-65535).
- o The effect of port exclusion on the possible values of the End-user IPv6 prefix (i.e., due to restrictions on the PSID value) SHOULD be minimized.
- o For administrative simplicity, the algorithm SHOULD allocate the the same or almost the same number of ports to each CE sharing a given IPv4 address.

The two extreme cases that an algorithm satisfying those conditions might support are: (1) the port numbers are not contiguous for each PSID, but uniformly distributed across the allowed port range; (2) the port numbers are contiguous in a single range for each PSID. The port mapping algorithm proposed here is called the Generalized Modulus Algorithm (GMA) and supports both these cases.

For a given IPv4 address sharing ratio (R) and the maximum number of contiguous ports (M) in a port-set, the GMA is defined as:

- a. The port numbers (P) corresponding to a given PSID are generated by:

$$(1) \dots P = (R * M) * i + M * PSID + j$$

where i and j are indices and the ranges of i, j, and the PSID are discussed in a moment.

- b. For any given port number P, the PSID is calculated as:

$$(2) \dots PSID = \text{trunc}((P \text{ modulo } (R * M)) / M)$$

where `trunc()` is the operation of rounding down to the nearest integer.

Formula (1) can be interpreted as follows. First, the available port space is divided into blocks of size $R * M$. Each block is divided into R individual ranges of length M . The index i in formula (1) selects a block, $PSID$ selects a range within that block, and the index j selects a specific port value within the range. On the basis of this interpretation:

- o i ranges from $\text{ceil}(N / (R * M))$ to $\text{trunc}(65536 / (R * M)) - 1$, where ceil is the operation of rounding up to the nearest integer and N is the number of ports (e.g., 1024) excluded from the lower end of the range. That is, any block containing excluded values is discarded at the lower end, and if the final block has fewer than $R * M$ values it is discarded. This ensures that the same number of ports is assigned to every PSID.
- o PSID ranges from 0 to $R - 1$;
- o j ranges from 0 to $M - 1$.

B.1. Bit Representation of the Algorithm

If R and M are powers of 2 ($R = 2^k$, $M = 2^m$), formula (1) translates to a computationally convenient structure for any port number represented as a 16-bit binary number. This structure is shown in Figure 9.

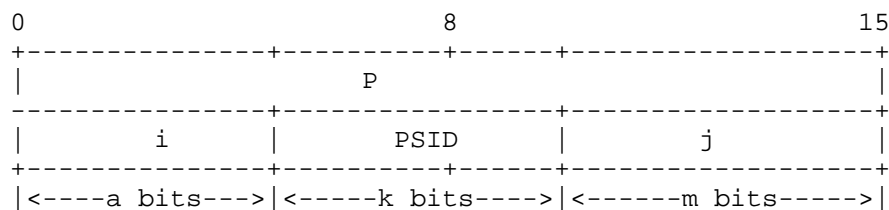


Figure 9: Bit Representation of a Port Number

As shown in the figure, the index value i of formula (1) is given by the first $a = 16 - k - m$ bits of the port number. The PSID value is given by the next k bits, and the index value j is given by the last m bits.

Because the PSID is always in the same position in the port number and always the same length, different PSID values are guaranteed to generate different sets of port numbers. In the reverse direction,

the generating PSID can be extracted from any port number by a bit mask operation.

Note that when M and R are powers of 2, 65536 divides evenly by $R * M$. Hence the final block is complete and the upper bound on i is exactly $65536 / (R * M) - 1$. The lower bound on i is still the minimum required to ensure that the required set of ports is excluded. No port numbers are wasted through discarding of blocks at the lower end if block size $R * M$ is a factor of N , the number of ports to be excluded.

As a final note, the number of blocks into which the range 0-65535 is being divided in the above representation is given by 2^a . Hence the case where $a = 0$ can be interpreted as one where the complete range has been divided into a single block, and individual port-sets are contained in contiguous ranges in that block. We cannot throw away the whole block in that case, so port exclusion has to be achieved by putting a lower bound equal to $\text{ceil}(N / M)$ on the allowed set of PSID values instead.

B.2. GMA examples

For example, for $R = 256$, $\text{PSID} = 0$, offset: $a = 6$ and PSID length: $k = 8$ bits

Available ports (63 ranges) : 1024-1027, 2048-2051, ,
63488-63491, 64512-64515

Example 1: with offset = 6 ($a = 6$)

For example, for $R = 64$, $\text{PSID} = 0$, $a = 0$ (PSID offset = 0 and PSID length = 6 bits), no port exclusion:

Available ports (1 range) : 0-1023

Example 2: with offset = 0 ($a = 0$) and $N = 0$

Authors' Addresses

Ole Troan (editor)
Cisco Systems
Philip Pedersens vei 1
Lysaker 1366
Norway

Email: ot@cisco.com

Wojciech Dec
Cisco Systems
Haarlerbergpark Haarlerbergweg 13-19
Amsterdam, NOORD-HOLLAND 1101 CH
Netherlands

Email: wdec@cisco.com

Xing Li
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing 100084
People's Republic of China

Email: xing@cernet.edu.cn

Congxiao Bao
CERNET Center/Tsinghua University
Room 225, Main Building, Tsinghua University
Beijing 100084
People's Republic of China

Email: congxiao@cernet.edu.cn

Satoru Matsushima
SoftBank Telecom
1-9-1 Higashi-Shinbashi, Munato-ku
Tokyo
Japan

Email: satoru.matsushima@g.softbank.co.jp

Tetsuya Murakami
IP Infusion
1188 East Arques Avenue
Sunnyvale
USA

Email: tetsuya@ipinfusion.com

Tom Taylor (editor)
Huawei Technologies
Ottawa
Canada

Email: tom.taylor.stds@gmail.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: December 11, 2015

Q. Sun
China Telecom
M. Chen
BBIX
G. Chen
China Mobile
T. Tsou
Huawei Technologies
S. Perreault
Jive Communications
June 9, 2015

Mapping of Address and Port (MAP) - Deployment Considerations
draft-ietf-softwire-map-deployment-06

Abstract

This document describes when and how an operator uses the technique of Mapping of Address and Port (MAP) for the IPv4 residual deployment in the IPv6-dominant domain.

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 December 11, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions	4
3. Case Studies	5
4. Deployment Consideration	7
4.1. Building the MAP Domain	7
4.1.1. MAP Deployment Model Planning	7
4.1.2. MAP Domain Planning	8
4.1.3. MAP Rule Provisioning	8
4.1.4. MAP DHCPv6 server deployment consideration	9
4.1.5. PSID Consideration	10
4.1.6. Addressing and Routing	10
4.1.7. MAP vs. MAP-T vs. 4rd	11
4.2. BR Settings	12
4.3. CE Settings	15
4.4. Supporting System	15
5. MAP Address Planning	17
5.1. Planning for Residual Deployment, a Step-by-step Guide . .	17
5.2. Remarks on Deployment Paradigms	19
6. Migration Methodology	21
6.1. Roadmap for MAP-based Solution	21
6.1.1. Start from Scratch	21
6.1.2. Coexisting Phases	21
6.1.3. Exit Strategy	21
6.2. Migration Mode	22
6.2.1. Passive Transition	22
6.2.2. Active Transition	22
7. IANA Considerations	23
8. Security Considerations	24
9. Contributors	25
10. Acknowledgements	26
11. References	27
11.1. Normative References	27
11.2. Informative References	27
Authors' Addresses	29

1. Introduction

IPv4 address exhaustion has become world-wide reality and the primary solution in the industry is to deploy IPv6-only networking. Meanwhile, having access to legacy IPv4 contents and services is a long-term requirement, will be so until the completion of the IPv6 transition. It demands sharing residual IPv4 address pools for IPv4 communications across the IPv6-only domain(s).

Mapping of Address and Port (MAP) [I-D.ietf-softwire-map] is designed in response to the requirement of stateless residual deployment. The term "residual deployment" refers to utilizing IPv4 addresses for IPv4 communications going across the IPv6 domain backbone. MAP assumes the IPv6-only backbone as the prerequisite of deployment so that native IPv6 services and applications are fully supported and encouraged. The statelessness of MAP ensures only moderate overhead is added to part of the network devices.

Residual deployment with MAP is new to most operators. This document is motivated to provide basic understanding on the usage of MAP, i.e., when and how an operator can do with MAP to meet its own operational requirements of IPv6 transition and its facility conditions, in the phase of IPv4 residual deployment. Potential readers of this document are those who want to know:

1. What are the requirements of MAP deployment ?
2. What technical options needs to be considered when deploying MAP, and how?
3. How does MAP impact on the address planning for both IPv6 and IPv4 pools?
4. How does MAP impact on daily network operations and administrations?
5. How do we migrate to IPv6-only network with the help of MAP?

Terminology of this document, unless it is intentionally specified, follows the definitions and abbreviations of [I-D.ietf-softwire-map].

Unless it is specifically specified, the deployment considerations and guidance proposed in this document are also applied to MAP-T [I-D.ietf-softwire-map-t], the translation variation of MAP, and 4rd [I-D.ietf-softwire-4rd], the reversible translation approach that aims to improve end-to-end consistency of double translation.

2. Conventions

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].

3. Case Studies

MAP can be deployed for large-scale carrier networks. There are typically two network models for broadband access service: one is to use PPPoE/PPPoA authentication method while the other is to use IPoE. The first one is usually applied to Residential network and SOHO networks. Subscribers in CPNs can access broadband network by PPP dial-up authentication. BRAS is the key network element which takes full responsibility of IP address assignment, user authentication, traffic aggregation, PPP session termination, etc. Then IP traffic is forwarded to Core Routers through Metro Area Network, and finally transited to Internet via Backbone network. The second network scenario is usually applied to large enterprise networks. Subscribers in CPNs can access broadband network by IPoE authentication. IP address is normally assigned by DHCP server, or static configuration.

In either case, a Customer Edge Router(CER) could obtain a prefix via prefix delegation procedure, and the hosts behind CER would get its own IPv6 addresses within the prefix through SLAAC or DHCPv6 statefully. A MAP CE would also obtain a set of MAP rules from DHCPv6 server.

Figure 1 depicts a generic model of stateless IPv4-over-IPv6 communication for broadband access services.

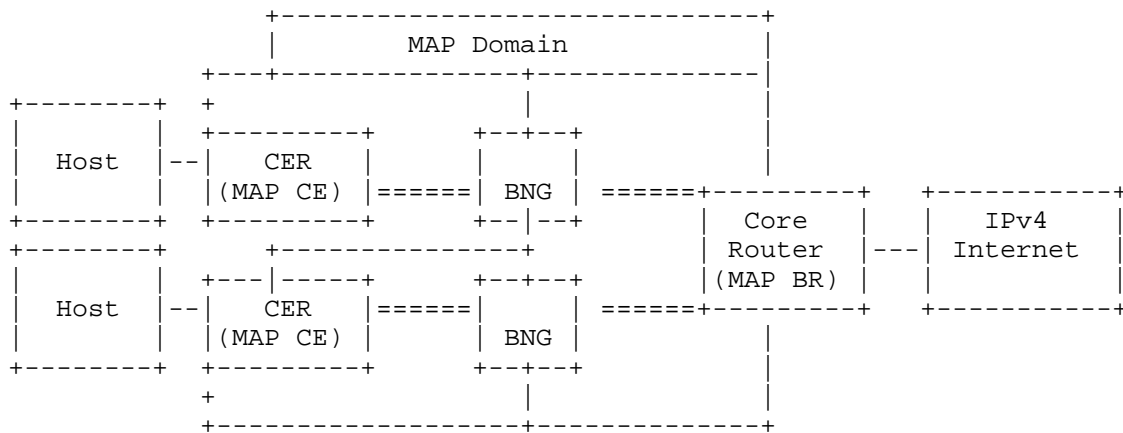


Figure 1: Stateless IPv4-over-IPv6 broadband access network architecture

When deploying MAP in home network, there can be two architecture: A. single ISP B. multihoming with two or more ISPs, sharing one CE. In the single ISP model, CE needs to communicate with only one MAP BR,

while in multihoming model CE has to communicate with multiple MAP BRs. Figure 2 [RFC7368] illustrates a typical case, where the home network has multiple connections to multiple providers or multiple logical connections to the same provider. In the multihoming model, a CE will be provisioned with multiple BMRs. Routing information will also be configured for multihoming; but detail of the routing configuration is out of the scope of this memo.

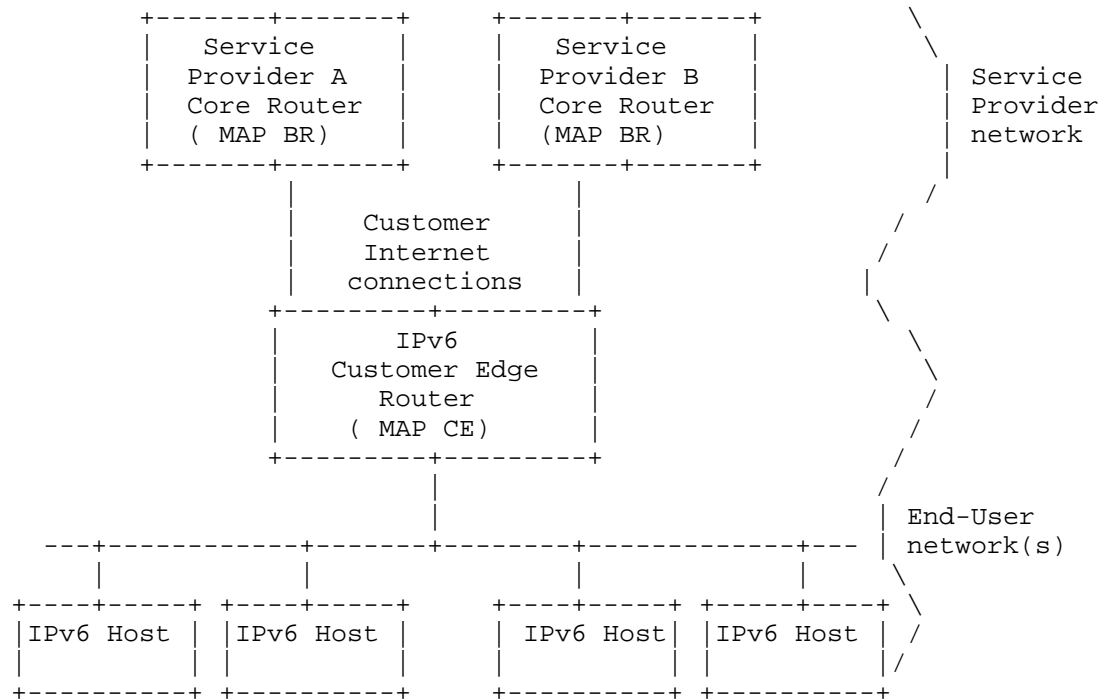


Figure 2: MAP multihoming

4. Deployment Consideration

4.1. Building the MAP Domain

When deploying stateless MAP in an operational network, a provider should firstly do MAP domain planning based on that existing network. According to the definition of [I-D.ietf-softwire-map], a MAP domain is a set of MAP CEs and BRs connected to the same virtual link. All CEs in the MAP domain are provisioned with a same set of MAP rules by MAP DHCPv6 server [I-D.ietf-softwire-map-dhcp]. There might be multiple BMRs in one MAP domain, e.g. in case of multi-ISP. A CE may be provisioned with multiple IPv6 prefix, which can be used to find the corresponding BMR via longest prefix match. As defined in [I-D.ietf-softwire-map-dhcp], a BMR should be provisioned together with a BR IPv6 address; the CE should maintain this binding, so that the mapping between BMR and BR is achieved which is useful in multi-ISP scenario. In in mesh mode, a longest-matching prefix lookup is done in the IPv4 routing table and the correct FMR is chosen.

Basically, operator should firstly determine its own deployment topology for MAP domain as described in Section 4.1.1, as different considerations apply for different deployment models. Next, MAP domain planning, MAP rule provision, addressing and routing, etc., for a MAP domain should be taken into consideration, as discussed in the sections following Section 4.1.1.

For the scenario where one CE is corresponding with multiple MAP border relays, it is possible that those MAP BRs belong to different MAP domains. The CE must pick up its own MAP rules and domain parameters in each domain. This is a typical case of multihoming. The MAP rules must have the information about BR(s) and information about the service types and the ISP.

4.1.1. MAP Deployment Model Planning

In order to do MAP domain planning, an operator should firstly make the decision to choose mesh or hub and spoke topology according to the operator's network policy. In the hub and spoke topology, all traffic within the same MAP domain has to go through the BR, result in less optimal traffic flow; however, it simplifies CE processing since there is no need to do FMR lookup for each incoming packet. Moreover, it provides enhanced manageability as the BR can take full control of all the traffic. As a result, it is reasonable to deploy hub and spoke topology for a network with a relatively flat architecture.

In mesh topology, CE to CE traffic flows are optimized since they pass directly between the two nodes. Mesh topology is recommended

when CE to CE traffic is high and there are not too many MAP rules, say fewer than 10 MAP rules, in the given domain.

4.1.2. MAP Domain Planning

Stateless MAP offers advantages in terms of scalability, high reliability, etc. As a result, it is reasonable to plan for a larger MAP domain to accommodate more subscribers with fewer BRs. Moreover, a larger MAP domain will also be easier for management and maintenance. However, a larger MAP domain may also result in less optimized traffic in the hub and spoke case, where all traffic has to go through a remote BR. In addition, it may result in an increased number of MAP rules and highly centralized address management. Choosing appropriate domain coverage requires the evaluation of tradeoffs.

When multiple IPv4 subnets are deployed in one MAP domain, it is recommended to further divide the MAP domain into multiple subdomains, each with only one IPv4 subnet. This can simplify the MAP domain planning. But there can be a side effect that it will increase the traffic between BRs. Different subdomains could be distinguished by different Rule IPv4 prefixes. As stated previously, all CEs within the same MAP subdomain will have the same Rule IPv4 prefix, Rule IPv6 prefix and PSID parameters.

4.1.3. MAP Rule Provisioning

In stateless MAP, Mesh or Hub and Spoke communications can be achieved among CEs in one MAP domain in terms of assigning appropriate FMR(s) to CEs. We recommend ISP deploy the full Hub and Spoke topology or full mesh topology describe below to simplify the configuration of the DHCPv6 server.

4.1.3.1. Full Hub and Spoke Communication among CEs

In order to achieve the full communication in the Hub and Spoke topology, no FMR is assigned to CEs. In this topology, when a CE sends packets to another CE in the same MAP domain via BR, or using the DMR as FMR, the packets must go through BR before arriving at the destination. DMR is specific for MAP-T only.

4.1.3.2. Full Mesh Communication among CEs

By assigning all BMRs in MAP domain to each CE as FMRs, Mesh communications can be achieved among all CEs. In this case, when CE receives an IPv4 packet, it looks up for an appropriate FMR with a specific Rule IPv4 prefix which has the longest match with the IPv4 destination address.

4.1.3.3. Mesh or Hub/Spoke communication among some CEs

Mesh communications among some CEs along with Hub/Spoke communications among some other CEs can be achieved by which differentiated FMRs are assigned to CEs. For instance, as shown in Figure 3, since both CE1 and CE2 has rule 1 and rule2, the communication between CE1 and CE2 can go directly without going through associated BR (Mesh topology). However, for CE1 and CE3, since there are no rule for each other, the communication between CE1 and CE3 must go through BR before reaching peer each other (Hub/Spoke topology).

	CE1	CE2	CE3
BMR	rule 1	rule 2	rule 3
FMRs	rule 1 rule 2	rule 1 rule 2 rule 3	rule 2 rule 3

Figure 3:

4.1.4. MAP DHCPv6 server deployment consideration

All the CEs within a MAP domain will get a set of MAP rules by DHCPv6 server. Each Mapping Rule keeps a record of Rule IPv6 prefix, Rule IPv4 prefix and Rule EA-bits length. Section 5 would give a step by step example of how to calculate these parameters.

As the MAP is stateless, the deployment of DHCPv6 server is independent of MAP domain planning. So there are three possible cases:

MAP domain : DHCPv6 server = 1:1 This is the ideal solution that each MAP domain would have its own MAP DHCPv6 server. In this case, MAP DHCPv6 server only needs to configure parameters for the specific MAP domain. In this model, it is easy to achieve the configuration in MAP and no extra configuration requirement is needed.

MAP domain : DHCPv6 server = 1:N This might happen when DHCPv6 servers are deployed in a large MAP domain in a distributed manner. In this case, all these DHCPv6 servers should be configured with the same set of MAP rules for the MAP domain, including multiple BMRs, FMRs and DMRs.

MAP domain : DHCPv6 server = N:1 This might happen when MAP domain is relatively small and a single MAP DHCPv6 server is deployed in the network. In this case, multiple MAP domains should be distinguished based on CE's IPv6 prefix in different MAP domains.

4.1.5. PSID Consideration

If a provider would like to introduce differentiated address sharing ratios for different CEs, it is better to define multiple MAP sub-domains with different Rule IPv4 prefixes. In this way, MAP domain division is only a logical method, rather than a geographical one.

The default PSID offset(a) is chosen as 6 in [I-D.ietf-softwire-map] and this excludes the system ports (0-1023). For MAP, the initial part of the port number (the a-bits) cannot be zero (see Appendix B of [I-D.ietf-softwire-map].) As is shown in the section 3.2.4 of [I-D.tsou-softwire-port-set-algorithms-analysis], it is possible that a lower value of 'a' will give a higher sharing ratio and more than 1024 ports are excluded as a result, e.g. 'a' = 4 will exclude ports 0 - 4095. The value of 'a' should be made explicitly configurable by operators.

With regard to PSID format, both continuous and non-continuous port set can be supported in GMA algorithm. Non-continuous port set has the advantage of better UPnP friendly, while continuous port set is the simplest way to implement. Since PSID format should be supported not only in CPEs, BRs and DHCPv6 server, but also in other sustaining systems as well, e.g. traffic logging system, user management system, a provider should make the decision based on a comprehensive investigation on its demand and the capabilities of existing equipments.

Note that some ISPs may need to offer services in a MAP domain with a shared address, e.g. there are hosts FTP server under CEs. The service provisioning may require well-know port range (i.e. port range belong to 0-1023). MAP would provide operators with an option to generate a port range including those in 0-1023. Afterwards, operators could decide to assign it to any requesting user. However, if the port-set is too small, it is not suggested to assign one with only the port set 0~1023 or even less. Considerable non-well-known ports are surely needed. Another easier approach is assigning a dedicated IPv4 address to such a CE if the demand really exists.

4.1.6. Addressing and Routing

In MAP addressing, it should follow the MAP rule planning in the MAP domain.

For IPv4 addressing, since the number of scattered IPv4 address prefixes would be equal to the number of FMR rules within a MAP domain, one should choose as large IPv4 address pool as possible to reduce the number of FMR rules. For IPv6 address, the Rule IPv6 prefixes should be equal to the end user IPv6 prefix in MAP domain.

If ISP has a /24 rule IPv4 prefix with sharing ratio of 64 gives 16000 customers, and a /16 rule IPv4 prefix supports 4 million customer. If up the sharing ratio to 256, 64000 and 16 million customers can be supports respectively. For the ISP who has scattered IPv4 address prefixes, in order to reduce the number of FMRs, according to needs of ports they can divide different classes. For instance, for the enterprise customers class which need many ports to use, provision them the BMR with low sharing ratio while for the private customers class which don't need so many ports provision them the BMR with high sharing ratio.

For MAP routing, there are no IPv4 routes exported to IPv6 networks.

4.1.7. MAP vs. MAP-T vs. 4rd

Basically, encapsulation provides an architectural building block of virtual link where the underlay behavior is fully hidden, while translation does a delivery participating into the end-to-end transferring path where behaviors are exposed. It is reflected in the following aspects.

1. Option header

If translation or 4rd 'reversible translation' is applied, IPv4 options at the IP layer are not translated according to [RFC791][RFC2460], and packets with those options MUST be dropped by Domain-entry nodes, and return ICMPv4 error messages to signal IPv4-option incompatibility. This limitation is acceptable because there are a lot firewalls in current IPv4 Internet also filter IPv4 packets

2. ICMP

Some IPv4 ICMP codes do not have a corresponding codes in ICMPv6, a detailed analysis on the double translation behavior suggest that some ICMPv4 messages, when they are translated to ICMPv6 and back to ICMPv4 across the IPv6 domain, the accuracy might be sacrificed to some extent. Encapsulation keeps the full transparency of ICMPv4 messages.

Reversible translation approach of 4rd, however, does not translate ICMPv4 messages into ICMPv6 version. Instead, it treats ICMP as same as a transport layer protocol data unit. This behavior is similar to

the encapsulation and keeps ICMP end-to-end transparency as well.

In either the encapsulation or translation mode, if an intermediate node generates an ICMPv6 error message, it should be converted into ICMPv4 version and returned to the source with a special source address and following the behavior specified in [RFC6791]. However, the behavior and semantics of the translation from ICMPv6 to ICMPv4 is different among encapsulation, translation and 4rd reversible translation approaches. Encapsulation treats routing error in the IPv6 domain as an (virtual)link error between the tunnel end points, while translation translate IPv6 routing error into corresponding IPv4 version, and 4rd, however, behaves according to whether the Tunnel Traffic Class option is set. The TTL behavior also reflect the differences among different approaches, which is worth paying attention to for the operating engineers. MAP-T translator is compatible with single translation approach.

3. PMTU and fragmentation

Both translation mode and encapsulation mode have PMTU and fragmentation problem. [RFC6145] discusses the problem in details for the translation, while [RFC2473] could be a reference on the issue in encapsulation.

4.2. BR Settings

1. BR placement

BR placement has important impacts on the operation of a MAP domain.

A first concern should be the avoidance of "triangle routing". In hub and spoke mode, all traffic will be routed through BR which may increase the path from the CE to an IPv4 peer. This can be accomplished easily by placing the BR close to the CE, such that the length of the path from the CE to the BR is minimized.

However, minimizing the CE-BR path would ignore a second concern, that of minimizing IPv4 operations. An ISP deploying MAP will probably want to focus on IPv6 operations, while keeping IPv4 operational expenditures to a minimum. This would imply that the size of the IPv4 network that the ISP has to administer would be kept to a minimum. Placing the BR near the CE means that the length of the IPv4 network between the BR and the IPv4 Internet would be longer.

Moreover, in case where the set of CEs is geographically dispersed, multiple BRs would be needed, which would further enlarge the IPv4 network that the ISP has to maintain.

Therefore, we offer the following guideline: BRs should be placed as close to the border with the IPv4 Internet as possible while keeping triangle routing to a minimum. Regional POPs should probably be considered as potential candidates.

Note also that MAP being stateless, asymmetric routing to/from the IPv4 Internet is natively supported and therefore no path-pinning mechanisms have to be additionally implemented.

Anycast can be used to let the network pick BR closest to a CE for traffic exiting the MAP domain. This is accomplished by provisioning a Default Mapping Rule containing an anycast IPv6 address or prefix. Operationally, this allows incremental deployment of BRs in strategic locations without modifying the provisioning system's configuration. CE's close to a newly-deployed BR will automatically start using it. The BR MUST participate in a dynamic IGP so that this can work automatically.

2. Reliability Considerations

Reliability of MAP is derived in major part from its statelessness. This means that MAP can benefit from the usual methods of Internet reliability.

Anycast, already mentioned in section 4.2.1, can be used to ensure reliability of traffic from CE to BR. Since there can be only one Default Mapping Rule per MAP domain, traffic from CE to BR will always use the same destination address. When this address is anycast, reliability is greatly increased. If a BR goes down, it stops advertising the IPv6 anycast address, and traffic is automatically re-routed to other BRs; the BR should also withdraw the routes for traffic from BR to CE, or the upstream routers connected to the BR should dynamically change the routes when it detects the failure of a BR, otherwise there will be a routing blackhole. For this mechanism to work correctly, it is crucial that the anycast route announcement be very closely tied to BR availability. See [RFC4786] for best current practices on the operation of anycast services. In practice, Equal-cost multi-path (ECMP) can be used to achieve active/active configuration. Operator can also increase the metric for one BR to have active/standby.

For reliability within a single link can be achieved with the help of a redundancy protocol such as VRRP [RFC5798]. This allows operation of a pair of BRs in active/standby configuration. No state needs to be shared for the operation of MAP, so there is no need to keep the standby node in a "warm" state: as long as it is up and ready to take over the virtual IPv6 address, quick failover can be achieved. This makes the pair behave as a single, much more reliable node, with less

reliance on quick routing protocol convergence for reliability.

It is expected that production-quality MAP deployments will make use of both anycast and a redundancy protocol such as VRRP.

3. MTU/Fragmentation

If the MTU is well-managed such that the IPv6 MTU on the CE WAN side interface is set so that no fragmentation occurs within the boundary of the MAP domain, then the Tunnel MTU can be set to the known IPv6 MTU minus the size of the encapsulating IPv6 header (40 bytes). For example, if the IPv6 MTU is known to be 1500 bytes, the Tunnel MTU might be set to 1460 bytes. Without more specific information, the Tunnel MTU SHOULD default to 1240 bytes.

BRs using an anycast address as source can cause problems. If traffic sent by a BR with a source anycast address causes an ICMP error to be returned, that error packet's destination address will be an anycast address, meaning that a different BR might receive it. In the case of a Too Big ICMP error, this could cause a path MTU discovery black hole. Another possible problem could occur if fragmented packets from different BRs using the same anycast address as source happen to contain the same fragment ID. This would break fragment reassembly. Since there is still no simple way to solve it completely, it is recommended to increase the MTU of the IPv6 network so that no fragmentation and Too Big ICMP error occurs.

In MAP domains where IPv4 addresses are not shared, IPv6 destinations are derived from IPv4 addresses alone. Thus, each IPv4 packet can be encapsulated and decapsulated independently of each other. The processing is completely stateless.

On the other hand, in MAP domains where IPv4 addresses are shared, BRs and CEs may have to encapsulate or translate IPv4 packets whose IPv6 destinations depend on destination ports. Precautions are needed, due to the fact that the destination port of a fragmented datagram is available only in its first fragment. A sufficient precaution consists in reassembling each datagram received in multiple packets, and to treat it as though it would have been received in single packet. This function is such that MAP is in this case stateful at the IP layer. (This is common with DS-lite and NAT64/DNS64 which, in addition, are stateful at the transport layer.) At domain entrance, this ensures that all pieces of all received IPv4 datagrams go to the right IPv6 destinations.

4.3. CE Settings

1. bridging vs. routing

In routing manner, the CE runs a standard NAT44 [RFC3022] using the allocated public address as external IP and ports via DHCPv6 option. When receiving an IPv4 packet with private source address from its end hosts, it performs NAT44 function by translating the source address into public and selecting a port from the allocated port-set. Then it encapsulates/translate (depending on whether MAP-E or MAP-T is in use) the packet with the concentrator's IPv6 address as destination IPv6 address, and forwards it to the concentrator. When receiving an IPv6 packet from the concentrator, the initiator decapsulates/translate the IPv6 packet to get the IPv4 packet with public destination IPv4 address. Then it performs NAT44 function and translates the destination address into private one based on the entry in NAT state table in the CE.

The CE is responsible for performing ALG functions (e.g., SIP, FTP), as well as supporting NAT Traversal mechanisms (e.g., UPnP, NAT-PMP, manual mapping configuration). This is no different from the standard IPv4 NAT today.

For the bridging manner, end host would run a software performing CE functionalities. In this case, end host gets public address directly. It is also suggested that the host run a local NAT to map randomly generated ports into the restricted, valid port-set. Another solution is to have the IP stack to only assign ports within the restricted, valid range to applications. Either way the host guarantees that every source port number in the outgoing packets falls into the allocated port-set.

2. CE-initiated application

CE-initiated case is applied for situations where applications run on CE directly. If the application in CE use the public address directly, it might conflict with other CEs. So it is highly suggested that CE should also run a local NAT to map a private address to public address in CE. In this way, the CE IPv4 address passed to local applications would be conflict with other CEs.

4.4. Supporting System

1. Lawful Intercept

Sharing IPv4 addresses among multiple CEs is susceptible to issues related to lawful intercept. For details, see [RFC6269] section 12.

2. Traffic Logging

It is always possible for a service provider that operates a MAP domain to determine the IPv6 prefix associated with a MAP IPv4 address (and port number in case of a shared address). This mapping is static, and it is therefore unnecessary to log every IPv4 address assignment. However, changes in that static mapping, such as rule changes in the provisioning system, need to be logged in order to be able to know the mapping at any point in time.

Sharing IPv4 addresses among multiple CEs is susceptible to issues related to traffic logging. For details, see [RFC6269] sections 8 and 13.1.

3. Geo-location aware service

Sharing IPv4 addresses among multiple CEs is susceptible to issues related to geo-location. For details, see [RFC6269] section 7.

4. User Management

MAP IPv4 address assignment, and hence the IPv4 service itself, is tied to the IPv6 prefix lease; thus, the MAP service is also tied to this in terms of authorization, accounting, etc. For example, the MAP address has the same lifetime as its associated IPv6 prefix.

5. MAP Address Planning

This section is purposed to provide a referential guidance to operators, illustrating a common method of address planning with MAP in IPv4 residual deployment.

5.1. Planning for Residual Deployment, a Step-by-step Guide

Residual deployment starts from IPv6 address planning.

(A) IPv6 considerations

- (A1) Determine the maximum number N of CEs to be supported, and, for generality, suppose $N = 2^n$.

For example, we suppose $n = 20$. It means there will be up to about one million CEs.

- (A2) Choose the length x of IPv6 prefixes to be assigned to ordinary customers.

Consider we have a /32 IPv6 block, it is not a problem for the IPv6 deployment with the given number of CEs. Let $x = 60$, allowing subnets inside in each CE delegated networks.

- (A3) Multiply N by a margin coefficient K , a power of two ($K = 2^k$), to take into account that:

- Some privileged customers may be assigned IPv6 prefixes of length x' , shorter than x , to have larger addressing spaces than ordinary customers, both in IPv6 and IPv4;
- Due to the hierarchy of routable prefixes, many theoretically delegable prefixes may not be actually delegable (ref: host density ratio of [RFC3194]).

In our example, let's take $k = 0$ for simplicity.

(B) IPv4 considerations

- (B1) List all (non overlapping, not yet assigned to any in-running networks) IPv4 prefixes $\{Hi\}$ that are available for IPv4 residual deployment.

Suppose that we hold two blocks and not yet assigned to any fixed network: 192.0.2.0/24 and 198.51.100.0/24.

- (B2) Take enough of them, among the shortest ones, to get a total whose size M is a power of two ($M = 2^m$), and includes a good proportion of the available IPv4 space.

If we use both blocks, $M = 2^{24} + 2^{24}$, and therefore $m = 25$. Suppose the intended sharing ratio is 8 subscribers per address, resulting in $(65536 - 1024)/8 = 8064$ ports per subscriber assuming that the well-known ports are excluded. Then the PSID length to achieve this will be $\log_2(8) = 3$ bits. Bearing in mind the IPv4 24 bit prefix length for each of our two prefixes, the EA-bit length is $(32 - 24) + 3 = 11$ bits.

- (B3) For each IPv4 prefix, H_i , of length h_i , choose an prefix extension, say R_i of length $r_i = m - (32 - h_i)$.

All these indexes must be non overlapping prefixes (e.g. 0, 10, 110, 111 for one /10, one /11, and two /12). In our example, we pick 0 for a contiguous address block while 1 for another.

Then we have:

```
H1 = 192.0.2.0/24, h1 = 24, r1 = 17 => R1 = bin(0);  
H2 = 198.51.100.0/24, h2 = 24, r2 = 17 => R2 = bin(1);
```

Sometimes the IPv4 residual pool is not well aggregated and the contiguous address blocks may have different sizes. For example, in (B1), if we have $H1 = 59.112.0.0/13$ and $H2 = 219.120.0.0/16$ as the IPv4 residual pool, then $M = 2^{19} + 2^{16}$, and in such a case, we must pick m so that $m = \text{ceil}(\log_2(M))$, where "ceil(x)" means the minimum integer not less than x , i.e., $m = 20$ in this case. Therefore $r1 = 20 - (32 - 13) = 1$, while $r2 = 20 - (32 - 16) = 4$. Several combinations are available for the $R1$ and $R2$ and one only needs to pay attention to avoiding overlapping when picking up the values.

- (C) After (A) and (B), derive the rule(s)

- (C1) Derive the length c of the MAP domain IPv6 prefix, C , that will appear at the beginning of all delegated prefixes ($c = x - (n + k)$).
- (C2) Take any prefix for this C of length c that starts with a RIR-allocated IPv6 prefix.
- (C3) For each IPv4 prefix H_i , make the rule, in which the key is H_i and the value is the domain IPv6 prefix C followed by the rule index R_i . Then this i -th rule's Rule IPv6 Prefix will have the length of $(c + r_i)$.

Then we can do that:

```
c = 40 => C = 2001:0db8:ff00::/40
Rule 1: Rule IPv6 Prefix = 2001:0db8:ff00::/41
Rule 2: Rule IPv6 Prefix = 2001:0db8:ff80::/41
```

If we have different lengths for the Rule IPv4 prefix (as the extra example discussed at the end of (B)), their Rule IPv6 prefixes should not have the same length, as their rule index length is different.

As a result, for a certain CE delegating 2001:0db8:ff98:7650::/60, its parameters are:

```
Rule IPv6 Prefix = 2001:0db8:ff80::/41 => Rule 2
IPv4 Suffix = bin(111 0110 0)
                        PSID = bin(101) = 0x5
Rule IPv4 Prefix = 198.51.100.0/24
CE IPv4 Address = 198.51.100.236
```

If different sharing ratio is demanded, we may partition CEs into groups and do (A) and (B) for each group, determining the PSID length for them separately.

5.2. Remarks on Deployment Paradigms

1. IPv6 address planning in residual deployment is independent of the usage of the residual IPv4 addresses. The IPv4 address pool for "residual deployment" contains IPv4 addresses not yet allocated to customers/subscribers and/or those already recalled from ex-customers, re-programmed into relatively well-aggregated blocks.
2. It is recommended to have the number of rule entries as less as possible so that the merit of stateless deployment is reflected in practical performances. However, this effort is often constrained by the condition of an operator whether (a): it holds large-enough contiguous IPv4 address block(s) for the residual deployment, and (b): a short-enough IPv6 domain prefix so that the /64 delegation is easily satisfied even the EA-bits is quite long. When condition (a) is not satisfied, sub-domains have to be defined for each relatively small but contiguous aggregated block; when condition (b) is not satisfied, one has to divide the IPv4 aggregates into smaller blocks artificially in order to reduce the length of EA-bits. When we have good conditions fitting (a) and (b), it is NOT recommended to define short EA-bits with small length of IPv4 suffix (the value p) nor to increase the number of rule entries (also the number of sub-

domains) unless it really has to.

3. An extreme case is, when EA-bits contain the full IPv4 address while a full IPv4 address is assigned to a CE, i.e., $o = p = 32$, and $q = 0$, the MAP address format becomes almost equivalent to RFC6052-format [RFC6052] except the off-domain IPv4 peer's mapped IPv6 address. This frees the domain to distribute rules but the DMR. In such a case, IPv6 addressing is fully dependent of IPv4, which defers from the typical residual deployment case. MAP is mainly designed for residual deployment but also applied for the case of legacy IPv4 networks keeping communication with the IPv4 world over the IPv6 domain without renumbering, as long as the address planning doesn't matter.
4. Another extreme case is, when EA-bits' length becomes to zero, i.e., $o = p = q = 0$, a rule actually defines a correspondence between an IPv6 address and an IPv4 address (or a prefix), without any algorithmic correlation to each other. Using such a case in practice is not prohibited by the specification, but it is not recommended to deploy null EA-bits in large scale as the concern discussed in the above Remark 2, and as it has the limitation that the PSID must be null ($q = 0$) and therefore multiple CEs sharing a same IPv4 address is not supported here. It is recommended to apply Lightweight 4over6 [I-D.ietf-softwire-lw4over6], if a full de-correlation between IPv6 address and IPv4 address as well as port range is demanded.
5. A not-so-extreme case, $p = 0$, $o = q$, i.e., only PSID is applied for the EA-bits, is also a case possibly happening in practice. It also potentially generates a huge number of rules and therefore large-scale deployment of this case is not recommended either.
6. For operators who would like to utilize "some bits" of IPv6 address to do service identification, QoS differentiation, etc., it is recommended that these special-purpose bits should be embedded before the EA-bits so as to reduce the possibility of bit-conflict. However, it requires quite shorter IPv6 aggregate prefix of the operator. The bit-conflict is more likely to happen in this case if different domains have different Rule prefix lengths. Operators with this demand should pay attention to the impact on the domain rule planning.

6. Migration Methodology

6.1. Roadmap for MAP-based Solution

6.1.1. Start from Scratch

IPv6 deployment normally involves a step-wise approach where parts of the network should properly updated gradually. As IPv6 deployment progresses it may be simpler for operators to employ a single-version network, since deploying both IPv4 and IPv6 in parallel would cost more than IPv6-only network. Therefore switching to an IPv6-only network in relatively small scale will become more prevalent. Meanwhile, a significant part of network will still stay in IPv4 for long time, especially at early stage of IPv6 transition. There may not be enough public or private IPv4 addresses to support end-to-end network communication, without segmenting the network into small parts with sharing one IPv4 address space. That is a time to introduce MAP to bridge these IPv4 islands through IPv6 network.

6.1.2. Coexisting Phases

SP has various deployment strategy in the middle of transition. It's foreseeable that IPv6 would likely coexist with IPv4 in a long period. The MAP deployment would also fit into the coexisting mode. To be specific, dual-stack technology is recommended in RFC6180 as the simplest deployment model to advance IPv6 deployment. MAP technology could get along well with native IPv6 connections and compatible with residual IPv4 networks. RFC6264 described a incremental transition approach in order to migrate networks to IPv6-only. DS-Lite is treated as a technology to accelerate the whole process. MAP can also take the same role to achieve a smooth transition.

6.1.3. Exit Strategy

The benefit of IPv6-only + MAP is that all IPv6 flows would go directly to the Internet, no need for encapsulation or translation. In this way, as more content providers and service are available over IPv6, the utilization on MAP CE and BR goes down since fewer destinations require MAP progressing. This way would advance IPv6, because it provides everyone incentives to use IPv6, and eventually the result is an pure IPv6 network with no need for IPv4. As more content providers and hosts equipped with IPv6 capabilities, the MAP utilization goes down until it is eventually not used at all when all content is IPv6. In this way, MAP has an "exit strategy". The corresponding solutions will leave the network in time.

6.2. Migration Mode

IPv4 Residual deployment is a interim phase during IPv6 migration. It would be beneficial to ISPs, if this phase is as short as possible since end-to-end IPv6 traversal is the really goals. When IPv6 is getting more and more mature, MAP would be retired in a natural way .

6.2.1. Passive Transition

Passive Transition is following IPv4 retirement law. In another word, MAP would always get along with IPv4, even all nodes is dual-stack capable. At a later stage of IPv6 migration, MAP can also be served for dual-stack hosts, which is sending traffic through the IPv4 stack. There is still a value for this approach because it could steer IPv4 traffic to IPv6 going through a MAP CE processing. When it comes the time ISP decide to turn off IPv4, MAP would be unnecessary due to IPv4 disappearance.

6.2.2. Active Transition

Active Transition is targeting to accelerate IPv4 exit and increase native IPv6 utilization. A desirable way deploying MAP is only providing IPv6 traversal ability to a IPv4-only host. However, MAP CE can not determine received traffic is send from a IPv4 node or a dual-stack node. In the latter case, IPv6 utilization is preferred for the most part . When a network evolves to a post-IPv6 era, it might be good for ISPs to consider to implement enforcement rules to help IPv6 migration.

- o ISP could install only IPv6 record (i.e. AAAA) in DNS server, which would provide users with IPv6 steering effects. When a host is IPv6-capable and gets IPv6 DNS reply in advance, MAP functionalities would be restricted by IPv6-only record response.
- o ISP could retrieve shared IPv4 address by increasing sharing ratio. In this case, number of concurrent IPv4 sessions on MAP CE would be suppressed. It would encourage native IPv6 growth in some extent.
- o ISP could allocate a dedicated IPv6 prefix for MAP deployment. The allocation could not only facilitate the differentiation between MAPed traffic and native IPv6 traffic, but also clearly observe the change of MAP traffic. When the traffic is reducing for a while, ISP could close the MAP functionalities in some specific area. It would result networks to native IPv6-only capable.

7. IANA Considerations

This specification does not require any IANA actions.

8. Security Considerations

There are no new security considerations pertaining to this document.

9. Contributors

The members of the MAP design team are:

Congxiao Bao, Mohamed Boucadair, Gang Chen, Maoke Chen, Wojciech Dec, Xiaohong Deng, Remi Despres, Jouni Korhonen, Xing Li, Satoru Matsushima, Tomasz Mrugalski, Tetsuya Murakami, Jacni Qin, Qiong Sun, Tina Tsou, Dan Wing, Leaf Yeh, and Jan Zorz.

Thanks to Chunfa Sun who was an active co-author of some earlier versions of this draft. Thanks to Shishio Tsuchiya's valueable suggestion for this document.

10. Acknowledgements

Remi Despres contributed the original example of step-by-step deployment guidance in discussion with the authors. Ole Troan, as the head of MAP Design Team, joined the discussion directly and contributed a lot of ideas and comments. We also thank other members of the MAP Design Team for their comments and suggestions.

Thanks to Tom Talyer, Qi Sun and Ian Farrer for their thorough review and helpful comments.

11. References

11.1. Normative References

- [I-D.ietf-softwire-4rd]
Despres, R., Jiang, S., Penno, R., Lee, Y., Chen, G., and M. Chen, "IPv4 Residual Deployment via IPv6 - a Stateless Solution (4rd)", draft-ietf-softwire-4rd-10 (work in progress), December 2014.
- [I-D.ietf-softwire-map]
Troan, O., Dec, W., Li, X., Bao, C., Matsushima, S., Murakami, T., and T. Taylor, "Mapping of Address and Port with Encapsulation (MAP)", draft-ietf-softwire-map-13 (work in progress), March 2015.
- [I-D.ietf-softwire-map-dhcp]
Mrugalski, T., Troan, O., Farrer, I., Perreault, S., Dec, W., Bao, C., Yeh, L., and X. Deng, "DHCPv6 Options for configuration of Softwire Address and Port Mapped Clients", draft-ietf-softwire-map-dhcp-12 (work in progress), March 2015.
- [I-D.ietf-softwire-map-t]
Li, X., Bao, C., Dec, W., Troan, O., Matsushima, S., and T. Murakami, "Mapping of Address and Port using Translation (MAP-T)", draft-ietf-softwire-map-t-08 (work in progress), December 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.
- [RFC6791] Li, X., Bao, C., Wing, D., Vaithianathan, R., and G. Huston, "Stateless Source Address Mapping for ICMPv6 Packets", RFC 6791, November 2012.

11.2. Informative References

- [RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, December 1998.
- [RFC3194] Durand, A. and C. Huitema, "The H-Density Ratio for Address Assignment Efficiency An Update on the H ratio", RFC 3194, November 2001.

- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC7368] Chown, T., Arkko, J., Brandt, A., Troan, O., and J. Weil, "IPv6 Home Networking Architecture Principles", RFC 7368, October 2014.

Authors' Addresses

Qiong Sun
China Telecom
Room 708 No.118, Xizhimenneidajie
Beijing, 100035
P.R.China

Phone: +86 10 5855 2923
Email: sunqiong@ctbri.com.cn

Maoke Chen
BBIX, Inc.
Tokyo Shiodome Building, Higashi-Shimbashi 1-9-1
Minato-ku, Tokyo 105-7310
Japan

Email: maoke@bbix.net

Gang Chen
China Mobile
28 Xuanwumenxi Ave; Xuanwu District
Beijing
P.R. China

Email: chengang@chinamobile.com

Tina Tsou
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1-408-330-4424
Email: tina.tsou.zouting@huawei.com

Simon Perreault
Jive Communications
Quebec, QC
Canada

Email: sperreault@jive.com

Softwires Working Group
Internet-Draft

Intended status: Standards Track
Expires: June 5, 2015

X. Li
C. Bao
CERNET Center/Tsinghua University
W. Dec, Ed.
O. Troan
Cisco Systems
S. Matsushima
SoftBank Telecom
T. Murakami
IP Infusion
December 2, 2014

Mapping of Address and Port using Translation (MAP-T)
draft-ietf-softwire-map-t-08

Abstract

This document specifies the "Mapping of Address and Port" stateless IPv6-IPv4 Network Address Translation (NAT64) based solution architecture for providing shared or non-shared IPv4 address connectivity to and across an IPv6 network.

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 June 5, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Conventions	3
3. Terminology	3
4. Architecture	5
5. Mapping Rules	7
5.1. Destinations outside the MAP domain	7
6. The IPv6 Interface Identifier	7
7. MAP-T Configuration	8
7.1. MAP CE	8
7.2. MAP BR	9
8. MAP-T Packet Forwarding	9
8.1. IPv4 to IPv6 at the CE	9
8.2. IPv6 to IPv4 at the CE	10
8.3. IPv6 to IPv4 at the BR	11
8.4. IPv4 to IPv6 at the BR	11
9. ICMP Handling	11
10. Fragmentation and Path MTU Discovery	12
10.1. Fragmentation in the MAP domain	12
10.2. Receiving IPv4 Fragments on the MAP domain borders	12
10.3. Sending IPv4 fragments to the outside	13
11. NAT44 Considerations	13
12. Usage Considerations	13
12.1. EA-bit length 0	13
12.2. Mesh and Hub and spoke modes	13
12.3. Communication with IPv6 servers in the MAP-T domain	14
12.4. Compatibility with other NAT64 solutions	14
13. IANA Considerations	14
14. Security Considerations	14
15. Contributors	15
16. Acknowledgements	16
17. References	16
17.1. Normative References	16
17.2. Informative References	16
Appendix A. Examples of MAP-T translation	19
Appendix B. Port mapping algorithm	22
Authors' Addresses	22

1. Introduction

Experiences from initial service provider IPv6 network deployments, such as [RFC6219], indicate that successful transition to IPv6 can happen while supporting legacy IPv4 users without a full end-to-end dual IP stack deployment. However, due to public IPv4 address exhaustion this requires an IPv6 technology that supports IPv4 users utilizing shared IPv4 addressing, while also allowing the network operator to optimize their operations around IPv6 network practices. The use of double NAT64 translation based solutions is an optimal way to address these requirements, especially in combination with stateless translation techniques that minimize operational challenges outlined in [I-D.ietf-software-stateless-4v6-motivation].

The Mapping of Address and Port - Translation (MAP-T) architecture specified in this draft is such a double stateless NAT64 based solution. It builds on existing stateless NAT64 techniques specified in [RFC6145], along with the stateless algorithmic address & transport layer port mapping scheme defined in MAP-E [I-D.ietf-softwire-map]. The MAP-T solution differs from MAP-E in the use of IPv4-IPv6 translation, rather than encapsulation, as the form of IPv6 domain transport. The translation mode is considered advantageous in scenarios where the encapsulation overhead, or IPv6 operational practices (e.g. Use of IPv6 only servers, or reliance on IPv6 + protocol headers for traffic classification) rule out encapsulation. These scenarios are presented in [I-D.maglione-softwire-map-t-scenarios]

2. Conventions

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].

3. Terminology

MAP-T Mapping of Address and Port by means of
 address Translation.

MAP Customer Edge (CE): A device functioning as a Customer Edge (CE) router in a MAP deployment. A typical MAP CE adopting MAP rules will serve a residential site with one WAN side IPv6 addressed interface, and one or more LAN side interfaces addressed using private IPv4 addressing.

MAP Border Relay (BR):	A MAP enabled router managed by the service provider at the edge of a MAP domain. A Border Relay (BR) router has at least an IPv6-enabled interface and an IPv4 interface connected to the native IPv4 network. A MAP BR may also be referred to simply as a "BR" within the context of MAP.
MAP domain:	One or more MAP CEs and BRs connected by means of an IPv6 network and sharing a common set of MAP Rules. A service provider may deploy a single MAP domain, or may utilize multiple MAP domains.
MAP Rule:	A set of parameters describing the mapping between an IPv4 prefix, IPv4 address or shared IPv4 address and an IPv6 prefix or address. Each MAP domain uses a different mapping rule set.
MAP Rule set:	A Rule set is composed out of all the MAP Rules communicated to a device, that are intended for determining the devices' IP+port mapping and forwarding operations. The MAP Rule set is interchangeably referred to in this document as a MAP Rule table or simply Rule table. Two specific types of rules, Basic Mapping Rule (BMR) and Forward Mapping Rule (FMR), are defined in Section 5 of [I-D.ietf-softwire-map]. The Default Mapping Rule (DMR) is defined in this document.
MAP Rule table:	See MAP Rule set.
MAP node:	A device that implements MAP.
Port-set:	Each node has a separate part of the transport layer port space; denoted as a port-set.
Port-set ID (PSID):	Algorithmically identifies a set of ports exclusively assigned to the CE.
Shared IPv4 address:	An IPv4 address that is shared among multiple CEs. Only ports that belong to the assigned port-set can be used for communication. Also known as a Port-Restricted IPv4 address.

End-user IPv6 prefix:	The IPv6 prefix assigned to an End-user CE by other means than MAP itself. E.g. Provisioned using DHCPv6 PD [RFC3633], assigned via SLAAC [RFC4862], or configured manually. It is unique for each CE.
MAP IPv6 address:	The IPv6 address used to reach the MAP function of a CE from other CEs and from BRs.
Rule IPv6 prefix:	An IPv6 prefix assigned by a Service Provider for a MAP rule.
Rule IPv4 prefix:	An IPv4 prefix assigned by a Service Provider for a MAP rule.
Embedded Address (EA) bits:	The IPv4 EA-bits in the IPv6 address identify an IPv4 prefix/address (or part thereof) or a shared IPv4 address (or part thereof) and a port-set identifier.

4. Architecture

Figure 1 depicts the overall MAP-T architecture, which sees any number of privately addressed IPv4 users (N and M) connected by means of MAP-T CEs to an IPv6 network that is equipped with one or more MAP-T BR. CEs and BRs that share MAP configuration parameters, referred to as MAP rules, form a MAP-T Domain.

Functionally the MAP-T CE and BR utilize and extend some well established technology building blocks to allow the IPv4 users to correspond with nodes on the Public IPv4 network, or IPv6 network as follows:

- o A (NAT44) NAT [RFC2663] function on a MAP CE is extended with support for restricting the allowable TCP/UDP ports for a given IPv4 address. The IPv4 address and port range used are determined by the MAP provisioning process and identical to MAP-E [I-D.ietf-softwire-map].
- o A stateless NAT64 function [RFC6145] is extended to allow stateless mapping of IPv4 and transport layer port ranges to IPv6 address space.

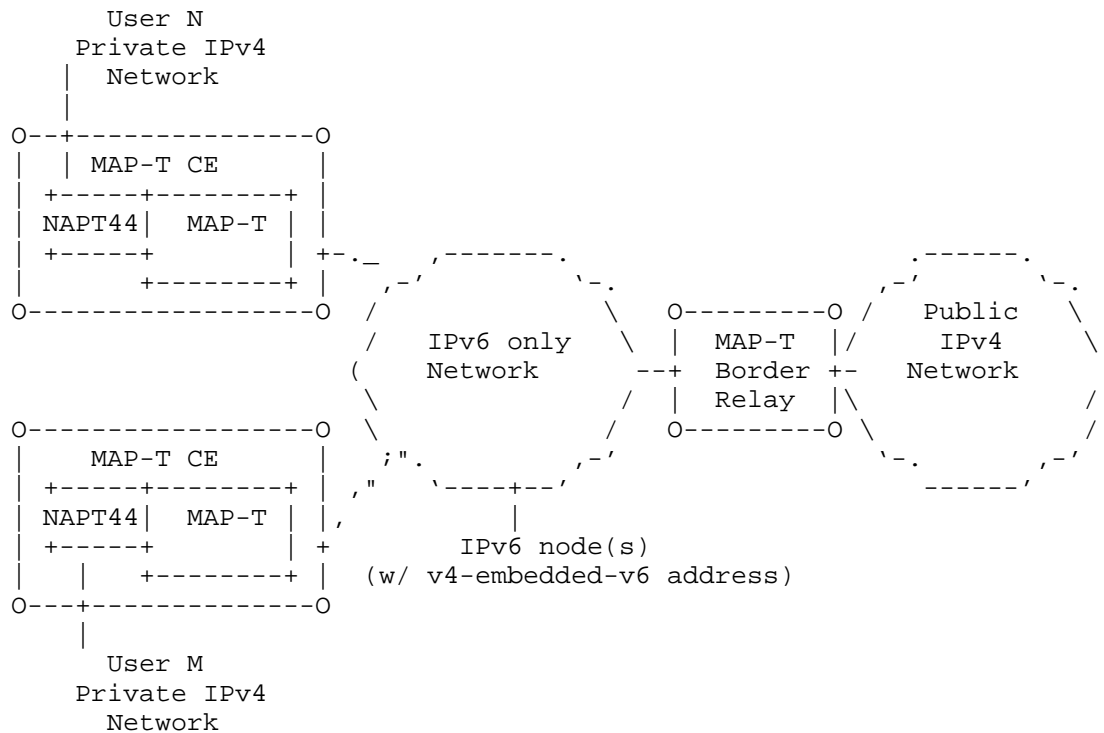


Figure 1: MAP-T Architecture

Each MAP-T CE is assigned with a regular IPv6 prefix from the operator's IPv6 network. This, in conjunction with MAP domain configuration settings and the use of the MAP procedures allows the computation of a MAP IPv6 address and a corresponding IPv4 address. To allow for IPv4 address sharing, the CE may also have be configured with a TCP/UDP port-range that is identified by means of a MAP Port Set Identifier (PSID) value. Each CE is responsible for forwarding traffic between a given user's private IPv4 address space and the MAP domain's IPv6 address space. The IPv4-IPv6 adaptation uses stateless NAT64, in conjunction with the MAP algorithm for address computation.

The MAP-T BR connects one or more MAP-T domains to external IPv4 networks using stateless NAT64 as extended by the MAP-T behaviour described in this document.

In contrast to MAP-E, NAT64 technology is used in the architecture for two purposes. Firstly, it is intended to diminish encapsulation overhead and allow IPv4 and IPv6 traffic to be treated as similarly as possible. Secondly, it is intended to allow IPv4-only nodes to

correspond directly with IPv6 nodes in the MAP-T domain that have IPv4 embedded IPv6 addresses as per [RFC6052]).

The MAP-T architecture is based on the following key properties i) algorithmic IPv4-IPv6 address mapping codified as MAP Rules covered in Section 5 ii) A MAP IPv6 address identifier, described in Section 6 iii) MAP-T IPv4-IPv6 forwarding behavior described in Section 8.

5. Mapping Rules

The MAP-T algorithmic mapping rules are identical to those in Section 5 of the MAP-E specification [I-D.ietf-softwire-map], with the following exception. The forwarding of traffic to and from IPv4 destinations outside a MAP-T domain is to be performed as described here under, instead of Section 5.4 of the MAP-E specification.

5.1. Destinations outside the MAP domain

IPv4 traffic sent by MAP nodes that are all within one MAP domain is translated to IPv6, with the sender's MAP IPv6 address, derived via the Basic Mapping Rule (BMR), as the IPv6 source address and the recipient's MAP IPv6 address, derived via the Forward Mapping Rule (FMR), as the IPv6 destination address.

IPv4 addressed destinations outside of the MAP domain are represented by means of IPv4-Embedded IPv6 address as per [RFC6052], using the BR's IPv6 prefix. For a CE sending traffic to any such destination, the source address of the IPv6 packet will be that of the CE's MAP IPv6 address, and the destination IPv6 address will be the destination IPv4-embedded-IPv6 address. This address mapping is termed as following the MAP-T Default Mapping Rule (DMR) and is defined in terms of the IPv6 prefix advertised by one or more BRs, which provide external connectivity. A typical MAP-T CE will install an IPv4 default route using this rule. A BR will use this rule when translating all outside IPv4 source addresses to the IPv6 MAP domain.

The DMR IPv6 prefix-length SHOULD be by default 64 bits long, and in any case MUST NOT exceed 96 bits. The mapping of the IPv4 destination behind the IPv6 prefix will by default follow the /64 rule as per [RFC6052]. Any trailing bits after the IPv4 address are set to 0x0.

6. The IPv6 Interface Identifier

The Interface identifier format of a MAP-T node is the same as described in section 6 of [I-D.ietf-softwire-map]. For convenience this is cited below:

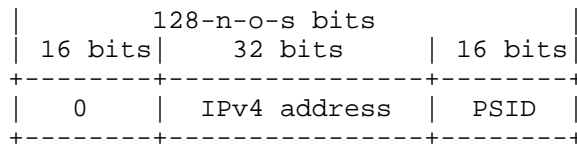


Figure 2

In the case of an IPv4 prefix, the IPv4 address field is right-padded with zeros up to 32 bits. The PSID is zero left-padded to create a 16 bit field. For an IPv4 prefix or a complete IPv4 address, the PSID field is zero.

If the End-user IPv6 prefix length is larger than 64, the most significant parts of the interface identifier is overwritten by the prefix.

7. MAP-T Configuration

For a given MAP domain, the BR and CE MUST be configured with the following MAP parameters. The values for these parameters are identical for all CEs and BRs within a given MAP-T domain.

- o The Basic Mapping Rule and optionally the Forwarding Mapping Rules, including the Rule IPv6 prefix, Rule IPv4 prefix, and Length of Embedded Address bits
- o Use of Hub and spoke mode or Mesh mode. (If all traffic should be sent to the BR, or if direct CE to CE correspondence should be supported).
- o Use of IPv4-IPv6 Translation (MAP-T)
- o The BR's IPv6 prefix used in the DMR

7.1. MAP CE

For a given MAP domain, the MAP configuration parameters are the same across all CEs within that domain. These values may be conveyed and configured on the CEs using a variety of methods, including; DHCPv6, Broadband Forum's "TR-69" Residential Gateway management interface, Netconf, or manual configuration. This document does not prescribe any of these methods, but recommends that a MAP CE SHOULD implement DHCPv6 options as per [I-D.ietf-softwire-map-dhcp]. Other configuration and management methods may use the data model described by this option for consistency and convenience of implementation on CEs that support multiple configuration methods.

Besides the MAP configuration parameters, a CE requires an IPv6 prefix to be assigned to the CE. This End-user IPv6 prefix is configured as part of obtaining IPv6 Internet access, and is acquired using standard IPv6 means applicable in the network where the CE is located.

The MAP provisioning parameters, and hence the IPv4 service itself, are tied to the End-user IPv6 prefix; thus, the MAP service is also tied to this in terms of authorization, accounting, etc.

A single MAP CE MAY be connected to more than one MAP domain, just as any router may have more than one IPv4-enabled service provider facing interface and more than one set of associated addresses assigned by DHCPv6. Each domain a given CE operates within would require its own set of MAP configuration elements and would generate its own IPv4 address. Each MAP domain requires a distinct End-user IPv6 prefix.

7.2. MAP BR

The MAP BR MUST be configured with the same MAP elements as the MAP CEs operating within the same domain.

For increased reliability and load balancing, the BR IPv6 prefix MAY be shared across a given MAP domain. As MAP is stateless, any BR may be used for forwarding to/from the domain at any time.

Since MAP uses provider address space, no specific IPv6 or IPv4 routes need to be advertised externally outside the service provider's network for MAP to operate. However, the BR prefix needs to be advertised in the service provider's IGP.

8. MAP-T Packet Forwarding

The end-to-end packet flow in MAP-T involves an IPv4 or IPv6 packet being forwarded by a CE or BR in one of two directions for each such case. This section presents a conceptual view of the operations involved in such forwarding.

8.1. IPv4 to IPv6 at the CE

A MAP-T CE receiving IPv4 packets SHOULD perform NAPT NAT44 processing, and create any necessary NAPT44 bindings. The source address and source port-range of packets resulting from the NAPT44 processing MUST correspond to the source IPv4 address and source transport port-range assigned to the CE by means of the MAP Basic Mapping Rule (BMR).

The IPv4 packet is subject to a longest IPv4 destination address + port match MAP rule selection, which then determines the parameters for the subsequent NAT64 operation. By default, all traffic is matched to the default mapping rule (DMR), and subject to the stateless NAT64 operation using the DMR parameters for NAT64 Section 5.1. Packets that are matched to (optional) Forward Mapping Rules (FMRs) are subject to the stateless NAT64 operation using the FMR parameters Section 5 for the MAP algorithm. In all cases the CE's MAP IPv6 address Section 6 is used as a source address.

A MAP-T CE MUST support a Default Mapping Rule and SHOULD support one or more Forward Mapping Rules.

8.2. IPv6 to IPv4 at the CE

A MAP-T CE receiving an IPv6 packet performs its regular IPv6 operations (filtering, pre-routing, etc). Only packets that are addressed to the CE's MAP-T IPv6 addresses, and with source addresses matching the IPv6 map-rule prefixes of a DMR or FMR, are processed by the MAP-T CE, with the DMR or FMR being selected based on a longest match. The CE MUST check that each MAP-T received packet's destination transport-layer destination port number is in the range allowed for by the CE's MAP BMR configuration. The CE MUST silently drop any non conforming packet and an appropriate counter incremented. When receiving a packet whose source IP address longest matches an FMR prefix, the CE MUST perform a check of consistency of the source address against the allowed values as per the derived allocated source port-range. If the source port number of a packet is found to be outside the allocated range, the CE MUST drop the packet and SHOULD respond with an ICMPv6 "Destination Unreachable, Source address failed ingress/egress policy" (Type 1, Code 5).

For each MAP-T processed packet, the CE's NAT64 function MUST compute an IPv4 source and destination addresses. The IPv4 destination address is computed by extracting relevant information from the IPv6 destination and the information stored in the BMR as per Section 5. The IPv4 source address is formed by classifying a packet's source as longest matching a DMR or FMR rule prefix, and then using the respective rule parameters for the NAT64 operation.

The resulting IPv4 packet is then forwarded to the CE's NAPT NAPT44 function, where the destination IPv4 address and port number MUST be mapped to their original value, before being forwarded according to the CE's regular IPv4 rules. When the NAPT44 function is not enabled, by virtue of MAP configuration, the traffic from the stateless NAT64 function is directly forwarded according to the CE's IPv4 rules.

8.3. IPv6 to IPv4 at the BR

A MAP-T BR receiving an IPv6 packet MUST select a matching MAP rule based on a longest address match of the packet's source address against the MAP Rules present on the BR. In combination with the Port-Set-Id derived from the packet's source IPv6 address, the selected MAP rule allows the BR to verify that the CE is using its allowed address and port range. Thus, the BR MUST perform a validation of the consistency of the source against the allowed values from the identified port-range. If the packet's source port number is found to be outside the range allowed, the BR MUST drop the packet and increment a counter to indicate the event. The BR SHOULD also respond with an ICMPv6 "Destination Unreachable, Source address failed ingress/egress policy" (Type 1, Code 5).

When constructing the IPv4 packet, the BR MUST derive the source and destination IPv4 addresses as per Section 5 of this document and translate the IPv6 to IPv4 headers as per [RFC6145]. The resulting IPv4 packet is then passed to regular IPv4 forwarding.

8.4. IPv4 to IPv6 at the BR

A MAP-T BR receiving IPv4 packets uses a longest match IPv4 + transport layer port lookup to identify the target MAP-T domain and select the FMR and DMR rules. The MAP-T BR MUST then compute and apply the IPv6 destination addresses from the IPv4 destination address and port as per the selected FMR. The MAP-T BR MUST also compute and apply the IPv6 source addresses from the IPv4 source address as per Section 5.1 (i.e. Using the IPv4 source and the BR's IPv6 prefix it forms an IPv6 embedded IPv4 address). Throughout the generic IPv4 to IPv6 header translation procedures following [RFC6145] apply. The resulting IPv6 packets are then passed to regular IPv6 forwarding.

Note that the operation of a BR when forwarding to/from MAP-T domains that are defined without IPv4 address sharing is the same as that of stateless NAT64 IPv4/IPv6 translation.

9. ICMP Handling

MAP-T CEs and BRs MUST follow ICMP/ICMPv6 translation as per [RFC6145], however additional behavior is also required due to the presence of NAPT44. Unlike TCP and UDP, which provide two transport protocol port fields to represent both source and destination, the ICMP/ICMPv6 [RFC0792], [RFC4443] Query message header has only one ID field which needs to be used to identify a sending IPv4 host. When receiving IPv4 ICMP messages, the MAP-T CE MUST rewrite the ID field to a port value derived from the CE's Port-Set-Id.

A MAP-T BR receiving an IPv4 ICMP packet , which contains an ID field that is bound for a shared address in the MAP-T domain, SHOULD use the ID value as a substitute for the destination port in determining the IPv6 destination address. In all other cases, the MAP-T BR MUST derive the destination IPv6 address by simply mapping the destination IPv4 address without additional port info.

10. Fragmentation and Path MTU Discovery

Due to the different sizes of the IPv4 and IPv6 header, handling the maximum packet size is relevant for the operation of any system connecting the two address families. There are three mechanisms to handle this issue: Path MTU discovery (PMTUD), fragmentation, and transport-layer negotiation such as the TCP Maximum Segment Size (MSS) option [RFC0897]. MAP can use all three mechanisms to deal with different cases.

Note: The NAT64 [RFC6145] mechanism is not lossless. When IPv4 originated communication traverses across a double NAT64 function (a.k.a. NAT464), any IPv4 originated ICMP-independent PathMTU Discovery, as specified in [RFC 4821], ceases to be entirely reliable. This is because the [RFC4821] defined DF=1/MF=1 combination, following a double NAT64 translation, results in DF=0/MF=1.

10.1. Fragmentation in the MAP domain

Translating an IPv4 packet to carry it across the MAP domain will increase its size typically by 20 bytes. The MTU in the MAP domain should be well managed and the IPv6 MTU on the CE WAN side interface SHOULD be configured so that no fragmentation occurs within the boundary of the MAP domain.

Fragmentation in MAP-T domain SHOULD be handled as described in section 4 and 5 of [RFC6145].

10.2. Receiving IPv4 Fragments on the MAP domain borders

Forwarding of an IPv4 packet received from the outside of the MAP domain requires the IPv4 destination address and the transport protocol destination port. The transport protocol information is only available in the first fragment received. As described in section 5.3.3 of [RFC6346] a MAP node receiving an IPv4 fragmented packet from outside SHOULD reassemble the packet before sending the packet onto the MAP domain. If the first packet received contains the transport protocol information, it is possible to optimize this behavior by using a cache and forwarding the fragments unchanged. A

description of such a caching algorithm is outside the scope of this document.

10.3. Sending IPv4 fragments to the outside

Two IPv4 hosts behind two different MAP CE's with the same IPv4 address sending fragments to an IPv4 destination host outside the domain may happen to use the same IPv4 fragmentation identifier, resulting in incorrect reassembly of the fragments at the destination host. Given that the IPv4 fragmentation identifier is a 16 bit field, it can be used similarly to port ranges. Thus, a MAP CE SHOULD rewrite the IPv4 fragmentation identifier to a value equivalent to a port of its allocated port-set.

11. NAT44 Considerations

The NAT44 implemented in the MAP CE SHOULD conform with the behavior and best current practice documented in [RFC4787], [RFC5508], and [RFC5382]. In MAP address sharing mode (determined by the MAP domain /rule configuration parameters) the operation of the NAT44 MUST be restricted to the available port numbers derived via the basic mapping rule.

12. Usage Considerations

12.1. EA-bit length 0

The MAP solution supports use and configuration of domains where a BMR expresses an EA-bit length of 0. This results in independence between the IPv6 prefix assigned to the CE and the IPv4 address and/or port-range used by MAP. The k-bits of PSID information may in this case be derived from the BMR.

The constraint imposed is that each such MAP domain be composed of just 1 MAP CE which has a predetermined IPv6 end-user prefix. The BR would be configured with an FMR for each such CPE, where the rule would uniquely associate the IPv4 address + optional PSID and the IPv6 prefix of that given CE.

12.2. Mesh and Hub and spoke modes

The hub and spoke mode of communication, whereby all traffic sent by a MAP-T CE is forwarded via a BR, and the mesh mode, whereby a CE is directly able to forward traffic to another CE, are governed by the activation of Forward Mapping Rule that cover the IPv4-prefix destination, and port-index range. By default, a MAP CE configured only with a BMR, as per this specification, will use it to configure its IPv4 parameters and IPv6 MAP address without enabling mesh mode.

12.3. Communication with IPv6 servers in the MAP-T domain

By default, MAP-T allows communication between both IPv4-only and any IPv6 enabled devices, as well as with native IPv6-only servers provided that the servers are configured with an IPv4-mapped IPv6 address. This address could be part of the IPv6 prefix used by the DMR in the MAP-T domain. Such IPv6 servers (e.g. An HTTP server, or a web content cache device) are thus able to serve both IPv6 users as well as IPv4-only users alike utilizing IPv6. Any such IPv6-only servers SHOULD have both A and AAAA records in DNS. DNS64 [RFC6147] become required only when IPv6 servers in the MAP-T domain are expected themselves to initiate communication to external IPv4-only hosts.

12.4. Compatibility with other NAT64 solutions

The MAP-T CEs NAT64 function is by default compatible for use with [RFC6146] stateful NAT64 devices that are placed in the operator's network. In such a case the MAP-T CE's DMR prefix is configured to correspond to the NAT64 device prefix. This in effect allows the use of MAP-T CEs in environments that need to perform statistical multiplexing of IPv4 addresses, while utilizing stateful NAT64 devices, and can take the role of a CLAT as defined in [RFC6877].

13. IANA Considerations

This specification does not require any IANA actions.

14. Security Considerations

Spoofing attacks: With consistency checks between IPv4 and IPv6 sources that are performed on IPv4/IPv6 packets received by MAP nodes, MAP does not introduce any new opportunity for spoofing attacks that would not already exist in IPv6.

Denial-of-service attacks: In MAP domains where IPv4 addresses are shared, the fact that IPv4 datagram reassembly may be necessary introduces an opportunity for DOS attacks. This is inherent to address sharing, and is common with other address sharing approaches such as DS-Lite and NAT64/DNS64. The best protection against such attacks is to accelerate IPv6 support in both clients and servers.

Routing-loop attacks: This attack may exist in some automatic tunneling scenarios are documented in [RFC6324]. They cannot exist with MAP because each BRs checks that the IPv6 source address of a received IPv6 packet is a CE address based on Forwarding Mapping Rule.

Attacks facilitated by restricted port-set: From hosts that are not subject to ingress filtering of [RFC2827], some attacks are possible by an attacker injecting spoofed packets during ongoing transport connections ([RFC4953], [RFC5961], [RFC6056]). The attacks depend on guessing which ports are currently used by target hosts, and using an unrestricted port-set is preferable, i.e. Using native IPv6 connections that are not subject to MAP port-range restrictions. To minimize this type of attacks when using a restricted port set, the MAP CE's NAT44 filtering behavior SHOULD be "Address-Dependent Filtering". Furthermore, the MAP CEs SHOULD use a DNS transport proxy function to handle DNS traffic, and source such traffic from IPv6 interfaces not assigned to MAP-T. Practicalities of these methods are discussed in Section 5.9 of [I-D.dec-stateless-4v6].

ICMP Flooding Given the necessity to process and translate ICMP and ICMPv6 messages by the BR and CE nodes, a foreseeable attack vector is that of a flood of such messages leading to a saturation of the node's ICMP computing resources. This attack vector is not specific to MAP, and its mitigation lies a combination of policing the rate of ICMP messages, policing the rate at which such messages can get processed by the MAP nodes, and of course identifying and blocking off the source(s) of such traffic.

[RFC6269] outlines general issues with IPv4 address sharing.

15. Contributors

The following individuals authored major contributions to this document, and made the document possible:

Chongfeng Xie (China Telecom) Room 708, No.118, Xizhimennei Street
Beijing 100035 CN Phone: +86-10-58552116 Email: xiechf@ctbri.com.cn

Qiong Sun (China Telecom) Room 708, No.118, Xizhimennei Street
Beijing 100035 CN Phone: +86-10-58552936 Email: sunqiong@ctbri.com.cn

Rajiv Asati (Cisco Systems) 7025-6 Kit Creek Road Research Triangle
Park NC 27709 USA Email: rajiva@cisco.com

Gang Chen (China Mobile) 53A,Xibianmennei Ave. Beijing 100053
P.R.China Email: chengang@chinamobile.com

Wentao Shang (CERNET Center/Tsinghua University) Room 225, Main
Building, Tsinghua University Beijing 100084 CN Email:
wentaoshang@gmail.com

Guoliang Han (CERNET Center/Tsinghua University) Room 225, Main Building, Tsinghua University Beijing 100084 CN Email: bupthgl@gmail.com

Yu Zhai CERNET Center/Tsinghua University Room 225, Main Building, Tsinghua University Beijing 100084 CN Email: jacky.zhai@gmail.com

16. Acknowledgements

This document is based on the ideas of many. In particular Remi Despres, who has tirelessly worked on generalized mechanisms for stateless address mapping.

The authors would also like to thank Mohamed Boucadair, Guillaume Gottard, Dan Wing, Jan Zorz, Nejc Scoberne, Tina Tsou, Gang Chen, Maoke Chen, Xiaohong Deng, Jouni Korhonen, Tomasz Mrugalski, Jacni Qin, Chunfa Sun, Qiong Sun, Leaf Yeh, Andrew Yourtchenko, Roberta Maglione and Hongyu Chen for their review and comments.

17. References

17.1. Normative References

- [I-D.ietf-softwire-map]
Troan, O., Dec, W., Li, X., Bao, C., Matsushima, S., Murakami, T., and T. Taylor, "Mapping of Address and Port with Encapsulation (MAP)", draft-ietf-softwire-map-12 (work in progress), November 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.
- [RFC6346] Bush, R., "The Address plus Port (A+P) Approach to the IPv4 Address Shortage", RFC 6346, August 2011.

17.2. Informative References

- [I-D.dec-stateless-4v6]
Dec, W., Asati, R., and H. Deng, "Stateless 4Via6 Address Sharing", draft-dec-stateless-4v6-04 (work in progress), October 2011.

- [I-D.ietf-software-map-dhcp]
Mrugalski, T., Troan, O., Farrer, I., Perreault, S., Dec, W., Bao, C., leaf.yeh.sdo@gmail.com, l., and X. Deng, "DHCPv6 Options for configuration of Software Address and Port Mapped Clients", draft-ietf-software-map-dhcp-11 (work in progress), November 2014.
- [I-D.ietf-software-stateless-4v6-motivation]
Boucadair, M., Matsushima, S., Lee, Y., Bonness, O., Borges, I., and G. Chen, "Motivations for Carrier-side Stateless IPv4 over IPv6 Migration Solutions", draft-ietf-software-stateless-4v6-motivation-05 (work in progress), November 2012.
- [I-D.maglione-software-map-t-scenarios]
Maglione, R., Dec, W., Leung, I., and E. Mallette, "Use cases for MAP-T", draft-maglione-software-map-t-scenarios-05 (work in progress), October 2014.
- [RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.
- [RFC0897] Postel, J., "Domain name system implementation schedule", RFC 897, February 1984.
- [RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.
- [RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, March 2007.

- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC4953] Touch, J., "Defending TCP Against Spoofing Attacks", RFC 4953, July 2007.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC5508] Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT Behavioral Requirements for ICMP", BCP 148, RFC 5508, April 2009.
- [RFC5961] Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's Robustness to Blind In-Window Attacks", RFC 5961, August 2010.
- [RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-Protocol Port Randomization", BCP 156, RFC 6056, January 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.
- [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P. Roberts, "Issues with IP Address Sharing", RFC 6269, June 2011.
- [RFC6324] Nakibly, G. and F. Templin, "Routing Loop Attack Using IPv6 Automatic Tunnels: Problem Statement and Proposed Mitigations", RFC 6324, August 2011.
- [RFC6877] Mawatari, M., Kawashima, M., and C. Byrne, "464XLAT: Combination of Stateful and Stateless Translation", RFC 6877, April 2013.

Appendix A. Examples of MAP-T translation

Example 1 - Basic Mapping Rule:

Given the following MAP domain information and IPv6 end-user prefix assigned to a MAP CE:

End-user IPv6 prefix: 2001:db8:0012:3400::/56
Basic Mapping Rule: {2001:db8:0000::/40 (Rule IPv6 prefix),
192.0.2.0/24 (Rule IPv4 prefix),
16 (Rule EA-bits length)}
PSID length: (16 - (32 - 24) = 8. (Sharing ratio of 256)
PSID offset: 6 (default)

A MAP node (CE or BR) can via the BMR, or equivalent FMR, determine the IPv4 address and port-set as shown below:

EA bits offset: 40
IPv4 suffix bits (p): Length of IPv4 address (32) - IPv4 prefix
length (24) = 8
IPv4 address: 192.0.2.18 (0xc0000212)
PSID start: 40 + p = 40 + 8 = 48
PSID length (q): o - p = (End-user prefix len -
rule IPv6 prefix len) - p
= (56 - 40) - 8 = 8
PSID: 0x34

Available ports (63 ranges): 1232-1235, 2256-2259, ,
63696-63699, 64720-64723

The BMR information allows a MAP CE to determine (complete) its IPv6 address within the indicated end-user IPv6 prefix.

IPv6 address of MAP CE: 2001:db8:0012:3400:0000:c000:0212:0034

Example 2 - BR:

Another example can be made of a MAP-T BR, configured with the following FMR when receiving a packet with the following characteristics:

IPv4 source address: 10.2.3.4 (0x0a020304)
TCP source port: 80
IPv4 destination address: 192.0.2.18 (0xc0000212)
TCP destination port: 1232

Forwarding Mapping Rule: {2001:db8::/40 (Rule IPv6 prefix),
192.0.2.0/24 (Rule IPv4 prefix),
16 (Rule EA-bits length)}

MAP-T BR Prefix (DMR): 2001:db8:ffff::/64

The above information allows the BR to derive as follows the mapped destination IPv6 address for the corresponding MAP-T CE, and also the source IPv6 address for the mapped IPv4 source address.

IPv4 suffix bits (p): $32 - 24 = 8$ (18 (0x12))
PSID length: 8
PSID: 0 x34 (1232)

The resulting IPv6 packet will have the following header fields:

IPv6 source address: 2001:db8:ffff:0:000a:0203:0400::
IPv6 destination address: 2001:db8:0012:3400:0000:c000:0212:0034
TCP source Port: 80
TCP destination Port: 1232

Example 3- FMR:

An IPv4 host behind a MAP-T CE (configured as per the previous examples) corresponding with an IPv4 host 10.2.3.4 will have its packets converted into IPv6 using the DMR configured on the MAP-T CE as follows:

```

Default Mapping Rule:      {2001:db8:ffff::/64 (Rule IPv6 prefix),
                           0.0.0.0/0 (Rule IPv4 prefix)}

IPv4 source address:       192.0.2.18
IPv4 destination address:  10.2.3.4
IPv4 source port:         1232
IPv4 destination port:    80
MAP-T CE IPv6 source address: 2001:db8:0012:3400:0000:c000:0212:0034
IPv6 destination address:  2001:db8:ffff:0:000a:0203:0400::

```

Example 4 - Rule with no embedded address bits and no address sharing

```

End-user IPv6 prefix:      2001:db8:0012:3400::/56
Basic Mapping Rule:        {2001:db8:0012:3400::/56 (Rule IPv6 prefix),
                           192.0.2.1/32 (Rule IPv4 prefix),
                           0 (Rule EA-bits length)}
PSID length:               0 (Sharing ratio is 1)
PSID offset:               n/a

```

A MAP node can via the BMR or equivalent FMR, determine the IPv4 address and port-set as shown below:

```

EA bits offset:            0
IPv4 suffix bits (p):      Length of IPv4 address - IPv4 prefix
                           length = 32 - 32 = 0
IPv4 address:              192.0.2.18 (0xc0000212)
PSID start:                0
PSID length:               0
PSID:                      null

```

The BMR information allows a MAP CE also to determine (complete) its full IPv6 address by combining the IPv6 prefix with the MAP interface identifier (that embeds the IPv4 address).

IPv6 address of MAP CE: 2001:db8:0012:3400:0000:c000:0201:0000

Example 5 - Rule with no embedded address bits and address sharing (sharing ratio 256)

```
End-user IPv6 prefix: 2001:db8:0012:3400::/56
Basic Mapping Rule:  {2001:db8:0012:3400::/56 (Rule IPv6 prefix),
                      192.0.2.18/32 (Rule IPv4 prefix),
                      0 (Rule EA-bits length)}
PSID length:         (16 - (32 - 24)) = 8. Sharing ratio of 256.
                      Provisioned with DHCPv6.
PSID offset:         6 (default)
PSID:                0x20 (Provisioned with DHCPv6)
```

A MAP node can via the BMR determine the IPv4 address and port-set as shown below:

```
EA bits offset:      0
IPv4 suffix bits (p): Length of IPv4 address - IPv4 prefix
                      length = 32 - 32 = 0
IPv4 address         192.0.2.18 (0xc0000212)
PSID start:          0
PSID length:         8
PSID:                0x34
```

Available ports (63 ranges) : 1232-1235, 2256-2259, ,
63696-63699, 64720-64723

The BMR information allows a MAP CE also to determine (complete) its full IPv6 address by combining the IPv6 prefix with the MAP interface identifier (that embeds the IPv4 address and PSID).

IPv6 address of MAP CE: 2001:db8:0012:3400:0000:c000:0212:0034

Note that the IPv4 address and PSID is not derived from the IPv6 prefix assigned to the CE, but provisioned separately using for example MAP options in DHCPv6.

Appendix B. Port mapping algorithm

The driving principles and the mathematical expression of the mapping algorithm used by MAP can be found in Appendix B of [I-D.ietf-softwire-map]

Authors' Addresses

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

Email: xing@cernet.edu.cn

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

Email: congxiao@cernet.edu.cn

Wojciech Dec (editor)
Cisco Systems
Haarlerbergpark Haarlerbergweg 13-19
Amsterdam, NOORD-HOLLAND 1101 CH
Netherlands

Email: wdec@cisco.com

Ole Troan
Cisco Systems
Oslo
Norway

Email: ot@cisco.com

Satoru Matsushima
SoftBank Telecom
1-9-1 Higashi-Shinbashi, Munato-ku
Tokyo
Japan

Email: satoru.matsushima@tm.softbank.co.jp

Tetsuya Murakami
IP Infusion
1188 East Arques Avenue
Sunnyvale
USA

Email: tetsuya@ipinfusion.com

Softwire
Internet-Draft
Intended status: Standards Track
Expires: June 21, 2016

Y. Cui
J. Dong
P. Wu
M. Xu
Tsinghua University
A. Yla-Jaaski
Aalto University
December 19, 2015

Softwire Mesh Management Information Base (MIB)
draft-ietf-softwire-mesh-mib-14

Abstract

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular it defines objects for managing a softwire mesh.

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 June 21, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. The Internet-Standard Management Framework	2
3. Terminology	3
4. Structure of the MIB Module	3
4.1. The swmSupportedTunnelTable Subtree	3
4.2. The swmEncapsTable Subtree	3
4.3. The swmBGPNeighborTable Subtree	4
4.4. The swmConformance Subtree	4
5. Relationship to Other MIB Modules	4
5.1. Relationship to the IF-MIB	4
5.2. Relationship to the IP Tunnel MIB	5
5.3. MIB modules required for IMPORTS	5
6. Definitions	5
7. Security Considerations	13
8. IANA Considerations	14
9. Acknowledgements	14
10. References	14
10.1. Normative References	14
10.2. Informative References	16
Authors' Addresses	16

1. Introduction

The Software mesh framework RFC 5565 [RFC5565] is a tunneling mechanism that enables connectivity between islands of IPv4 networks across a single IPv6 backbone and vice versa. In a software mesh, extended multiprotocol-BGP (MP-BGP) is used to set up tunnels and advertise prefixes among address family border routers (AFBRs).

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular it defines objects for managing a software mesh [RFC5565].

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC 3410 [RFC3410].

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). They

are defined using the mechanisms stated in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIV2 (Structure of Management Information Version 2), which is described in STD 58, RFC 2578 [RFC2578], STD 58, RFC 2579 [RFC2579] and STD 58, RFC 2580 [RFC2580].

3. Terminology

This document uses terminology from the software problem statement RFC 4925 [RFC4925], the BGP encapsulation subsequent address family identifier (SAFI) and the BGP tunnel encapsulation attribute RFC 5512 [RFC5512], the software mesh framework RFC 5565 [RFC5565] and the BGP IPsec tunnel encapsulation attribute and RFC 5566 [RFC5566].

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

4. Structure of the MIB Module

The software mesh MIB provides a method to monitor the software mesh objects through SNMP.

4.1. The swmSupportedTunnelTable Subtree

The swmSupportedTunnelTable subtree provides the information about what types of tunnels can be used for software mesh scenarios in the AFBR. The software mesh framework RFC 5565 [RFC5565] does not mandate the use of any particular tunneling technology. Based on the BGP tunnel encapsulation attribute tunnel types introduced by RFC 5512 [RFC5512] and RFC 5566 [RFC5566], the software mesh tunnel types include at least L2TPv3 (Layer Two Tunneling Protocol-Version 3) over IP, GRE (Generic Routing Encapsulation), Transmit tunnel endpoint, IPsec in Tunnel-mode, IP in IP tunnel with IPsec Transport Mode, MPLS-in-IP tunnel with IPsec Transport Mode and IP in IP. The detailed encapsulation information of different tunnel types (e.g., L2TPv3 Session ID, GRE Key, etc.) is not managed in the swmMIB.

4.2. The swmEncapTable Subtree

The swmEncapTable subtree provides software mesh NLRI-NH information (Network Layer Reachability Information-Next Hop) about the AFBR. It keeps the mapping between the External-IP (E-IP) prefix and the Internal-IP (I-IP) address of the next hop. The mappings determine which I-IP destination address will be used to encapsulate the received packet according to its E-IP destination address. The definitions of E-IP and I-IP are explained in section 4.1 of RFC

5565[RFC5565]. The number of entries in swmEncapsTable shows how many software mesh tunnels are maintained in this AFBR.

4.3. The swmBGPNeighborTable Subtree

The subtree provides the software mesh BGP neighbor information of an AFBR. It includes the address of the software mesh BGP peer, and the kind of tunnel that the AFBR would use to communicate with this BGP peer.

4.4. The swmConformance Subtree

The subtree provides the conformance information of MIB objects.

5. Relationship to Other MIB Modules

5.1. Relationship to the IF-MIB

The Interfaces MIB [RFC2863] defines generic managed objects for managing interfaces. Each logical interface (physical or virtual) has an ifEntry. Tunnels are handled by creating logical interfaces (ifEntry). Being a tunnel, software mesh interface has an entry in the Interface MIB, as well as an entry in IP Tunnel MIB. Those corresponding entries are indexed by ifIndex.

The ifOperStatus in the ifTable represents whether the mesh function of the AFBR has been triggered. If the software mesh capability is negotiated during the BGP OPEN phase, the mesh function is considered to be started, and the ifOperStatus is "up". Otherwise the ifOperStatus is "down".

In the case of an IPv4-over-IPv6 software mesh tunnel, ifInUcastPkts counts the number of IPv6 packets which are sent to the virtual interface for decapsulation into IPv4. The ifOutUcastPkts counts the number of IPv6 packets which are generated by encapsulating IPv4 packets sent to the virtual interface. Particularly, if these IPv4 packets need fragmentation, ifOutUcastPkts counts the number of packets after fragmentation.

In the case of an IPv6-over-IPv4 software mesh tunnel, ifInUcastPkts counts the number of IPv4 packets, which are delivered up to the virtual interface for decapsulation into IPv6. The ifOutUcastPkts counts the number of IPv4 packets, which are generated by encapsulating IPv6 packets sent down to the virtual interface. Particularly, if these IPv6 packets need to be fragmented, ifOutUcastPkts counts the number of packets after fragmentation. Similar definitions apply to other counter objects in the ifTable.

5.2. Relationship to the IP Tunnel MIB

The IP Tunnel MIB [RFC4087] contains objects applicable to all IP tunnels, including software mesh tunnels. Meanwhile, the Software Mesh MIB extends the IP Tunnel MIB to further describe encapsulation-specific information.

When running a point to multi-point tunnel, it is necessary for a software mesh AFBR to maintain an encapsulation table in order to perform correct "forwarding" among AFBRs. This forwarding function on an AFBR is performed by using the E-IP destination address to look up in the encapsulation table for the I-IP encapsulation destination address. An AFBR also needs to know the BGP peer information of the other AFBRs, so that it can negotiate the NLRI-NH information and the tunnel parameters with them.

The Software mesh MIB requires the implementation of the IP Tunnel MIB. The tunnelIfEncapsMethod in the tunnelIfEntry MUST be set to softwareMesh("xx"), and a corresponding entry in the software mesh MIB module will be presented for the tunnelIfEntry. The tunnelIfRemoteInetAddress MUST be set to "0.0.0.0" for IPv4 or "::" for IPv6 because it is a point to multi-point tunnel.

-- RFC Ed.: Please replace "xx" with IANA assigned number here.

The tunnelIfAddressType in the tunnelIfTable represents the type of address in the corresponding tunnelIfLocalInetAddress and tunnelIfRemoteInetAddress objects. The tunnelIfAddressType is identical to swmEncapsIIPDstType in software mesh, which can support either IPv4-over-IPv6 or IPv6-over-IPv4. When the swmEncapsEIPDstType is IPv6 and the swmEncapsIIPDstType is IPv4, the tunnel type is IPv6-over-IPv4; When the swmEncapsEIPDstType is IPv4 and the swmEncapsIIPDstType is IPv6, the encapsulation mode would be IPv4-over-IPv6.

5.3. MIB modules required for IMPORTS

The following MIB module IMPORTS objects from SNMPv2-SMI [RFC2578], SNMPv2-CONF [RFC2580], IF-MIB [RFC2863] and INET-ADDRESS-MIB [RFC4001].

6. Definitions

SOFTWARE-MESH-MIB DEFINITIONS ::= BEGIN

IMPORTS

MODULE-IDENTITY, OBJECT-TYPE, mib-2 FROM SNMPv2-SMI

OBJECT-GROUP, MODULE-COMPLIANCE FROM SNMPv2-CONF

InetAddress, InetAddressType, InetAddressPrefixLength

FROM INET-ADDRESS-MIB

ifIndex

FROM IF-MIB

IANAAtunnelType

FROM IANAifType-MIB;

swmMIB MODULE-IDENTITY

LAST-UPDATED "201512190000Z" -- December 19, 2015

ORGANIZATION "Softwire Working Group"

CONTACT-INFO "

Yong Cui

Email: yong@csnet1.cs.tsinghua.edu.cn

Jiang Dong

Email: knight.dongjiang@gmail.com

Peng Wu

Email: weapon9@gmail.com

Mingwei Xu

Email: xmw@cernet.edu.cn

Antti Yla-Jaaski

Email: antti.yla-jaaski@aalto.fi

Email comments directly to the softwire WG Mailing
List at softwires@ietf.org

"

DESCRIPTION

"This MIB module contains managed object definitions for
the softwire mesh framework.

Copyright (C) The Internet Society (2015). This
version of this MIB module is part of RFC 5565;
see the RFC itself for full legal notices."

REVISION "201512190000Z"

DESCRIPTION

"The MIB module is defined for management of object in
the Softwire mesh framework."

::= { mib-2 xxx }

```
--RFC Ed.: Please replace "xxx" with IANA assigned number here.

swmObjects OBJECT IDENTIFIER ::= { swmMIB 1 }

-- swmSupportedTunnelTable
swmSupportedTunnelTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF SwmSupportedTunnelEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "A table of objects that shows what kind of tunnels
        can be supported by the AFBR."
    ::= { swmObjects 1 }

swmSupportedTunnelEntry OBJECT-TYPE
    SYNTAX      SwmSupportedTunnelEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "A set of objects that show what kind of tunnels
        can be supported in the AFBR. If the AFBR supports
        multiple tunnel types, the swmSupportedTunnelTable
        would have several entries."
    INDEX { swmSupportedTunnelType }
    ::= { swmSupportedTunnelTable 1 }

SwmSupportedTunnelEntry ::= SEQUENCE {
    swmSupportedTunnelType      IANAtunnelType
}

swmSupportedTunnelType OBJECT-TYPE
    SYNTAX      IANAtunnelType
    MAX-ACCESS  read-only
    STATUS      current
    DESCRIPTION
        "Represents the tunnel type that can be used for software
        mesh scenarios, such as L2TPv3 over IP, GRE, Transmit
        tunnel endpoint, IPsec in Tunnel-mode, IP in IP tunnel with
        IPsec Transport Mode, MPLS-in-IP tunnel with IPsec Transport
        Mode and IP in IP. There is no restriction of tunnel type
        the Software mesh can use."
    REFERENCE
        "L2TPv3 over IP, GRE, IP in IP in RFC5512.
        Transmit tunnel endpoint, IPsec in Tunnel-mode, IP in IP
        tunnel with IPsec Transport Mode, MPLS-in-IP tunnel with
        IPsec Transport Mode in RFC5566."
    ::= { swmSupportedTunnelEntry 1 }
```

```
-- end of swmSupportedTunnelTable

--swmEncapsTable
swmEncapsTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF SwmEncapsEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "A table of objects that display the
        software mesh encapsulation information."
    ::= { swmObjects 2 }

swmEncapsEntry OBJECT-TYPE
    SYNTAX      SwmEncapsEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "A table of objects that manage the software mesh I-IP
        encapsulation destination based on the E-IP destination
        prefix."
    INDEX { ifIndex,
            swmEncapsEIPDstType,
            swmEncapsEIPDst,
            swmEncapsEIPPrefixLength
          }
    ::= { swmEncapsTable 1 }

SwmEncapsEntry ::= SEQUENCE {
    swmEncapsEIPDstType      InetAddressType,
    swmEncapsEIPDst          InetAddress,
    swmEncapsEIPPrefixLength InetAddressPrefixLength,
    swmEncapsIIPDstType      InetAddressType,
    swmEncapsIIPDst          InetAddress
}

swmEncapsEIPDstType OBJECT-TYPE
    SYNTAX      InetAddressType
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "This object specifies the address type used for
        swmEncapsEIPDst. It is different from the tunnelIfAddressType
        in the tunnelIfTable. The swmEncapsEIPDstType is IPv6 (2)
        if it is IPv6-over-IPv4 tunneling. The swmEncapsEIPDstType is
        IPv4 (1) if it is IPv4-over-IPv6 tunneling."
    REFERENCE
        "IPv4 and IPv6 in RFC 4001."
    ::= { swmEncapsEntry 1 }
```

```
swmEncapsEIPDst OBJECT-TYPE
    SYNTAX      InetAddress
    MAX-ACCESS   not-accessible
    STATUS       current
    DESCRIPTION
        "The E-IP destination prefix, which is
        used for I-IP encapsulation destination looking up.
        The type of this address is determined by the
        value of swmEncapsEIPDstType"
    REFERENCE
        "E-IP and I-IP in RFC 5565."
    ::= { swmEncapsEntry 2 }

swmEncapsEIPPrefixLength OBJECT-TYPE
    SYNTAX      InetAddressPrefixLength
    MAX-ACCESS   not-accessible
    STATUS       current
    DESCRIPTION
        "The prefix length of the E-IP destination prefix."
    ::= { swmEncapsEntry 3 }

swmEncapsIIPDstType OBJECT-TYPE
    SYNTAX      InetAddressType
    MAX-ACCESS   read-only
    STATUS       current
    DESCRIPTION
        "This object specifies the address type used for
        swmEncapsIIPDst. It is the same as the tunnelIfAddressType
        in the tunnelIfTable."
    REFERENCE
        "IPv4 and IPv6 in RFC 4001."
    ::= { swmEncapsEntry 4 }

swmEncapsIIPDst OBJECT-TYPE
    SYNTAX      InetAddress
    MAX-ACCESS   read-only
    STATUS       current
    DESCRIPTION
        "The I-IP destination address, which is used as the
        encapsulation destination for the corresponding E-IP
        prefix. Since the tunnelIfRemoteInetAddress in the
        tunnelIfTable should be 0.0.0.0 or ::, swmEncapsIIPDst
        should be the destination address used in the outer
        IP header."
    REFERENCE
        "E-IP and I-IP in RFC 5565."
    ::= { swmEncapsEntry 5 }
-- End of swmEncapsTable
```

```
-- swmBGPNeighborTable
swmBGPNeighborTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF SwmBGPNeighborEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "A table of objects that display the software mesh
        BGP neighbor information."
    ::= { swmObjects 3 }

swmBGPNeighborEntry OBJECT-TYPE
    SYNTAX      SwmBGPNeighborEntry
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "A set of objects that display the software mesh
        BGP neighbor information."
    INDEX {
        ifIndex,
        swmBGPNeighborInetAddressType,
        swmBGPNeighborInetAddress
    }
    ::= { swmBGPNeighborTable 1 }

SwmBGPNeighborEntry ::= SEQUENCE {
    swmBGPNeighborInetAddressType    InetAddressType,
    swmBGPNeighborInetAddress        InetAddress,
    swmBGPNeighborTunnelType         IANAtunnelType
}

swmBGPNeighborInetAddressType OBJECT-TYPE
    SYNTAX      InetAddressType
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "This object specifies the address type used for
        swmBGPNeighborInetAddress."
    ::= { swmBGPNeighborEntry 1 }

swmBGPNeighborInetAddress OBJECT-TYPE
    SYNTAX      InetAddress
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "The address of the AFBR's BGP neighbor. The
        address type is the same as the tunnelIfAddressType
        in the tunnelIfTable."
    ::= { swmBGPNeighborEntry 2 }
```

```
swmBGPNeighborTunnelType OBJECT-TYPE
    SYNTAX      IANAtunnelType
    MAX-ACCESS  read-only
    STATUS      current
    DESCRIPTION
        "Represents the type of tunnel that the AFBR
        chooses to transmit traffic with another AFBR/BGP
        neighbor."
    ::= { swmBGPNeighborEntry 3 }
-- End of swmBGPNeighborTable

-- conformance information
swmConformance
    OBJECT IDENTIFIER ::= { swmMIB 2 }
swmCompliances
    OBJECT IDENTIFIER ::= { swmConformance 1 }
swmGroups
    OBJECT IDENTIFIER ::= { swmConformance 2 }

-- compliance statements
swmCompliance MODULE-COMPLIANCE
    STATUS current
    DESCRIPTION
        "Describes the requirements for conformance to the software
        mesh MIB.

        The following index objects cannot be added as OBJECT
        clauses but nevertheless have compliance requirements:
        "
    -- OBJECT  swmEncapsEIPDstType
    -- SYNTAX  InetAddressType { ipv4(1), ipv6(2) }
    -- DESCRIPTION
    -- "An implementation is required to support
    -- global IPv4 and/or IPv6 addresses, depending
    -- on its support for IPv4 and IPv6."

    -- OBJECT  swmEncapsEIPDst
    -- SYNTAX  InetAddress (SIZE(4|16))
    -- DESCRIPTION
    -- "An implementation is required to support
    -- global IPv4 and/or IPv6 addresses, depending
    -- on its support for IPv4 and IPv6."

    -- OBJECT  swmEncapsEIPPrefixLength
    -- SYNTAX  InetAddressPrefixLength (Unsigned32 (0..128))
    -- DESCRIPTION
    -- "An implementation is required to support
```

```
-- global IPv4 and/or IPv6 addresses, depending
-- on its support for IPv4 and IPv6."

-- OBJECT swmBGPNeighborInetAddressType
-- SYNTAX InetAddressType { ipv4(1), ipv6(2) }
-- DESCRIPTION
-- "An implementation is required to support
-- global IPv4 and/or IPv6 addresses, depending
-- on its support for IPv4 and IPv6."

-- OBJECT swmBGPNeighborInetAddress
-- SYNTAX InetAddress (SIZE(4|16))
-- DESCRIPTION
-- "An implementation is required to support
-- global IPv4 and/or IPv6 addresses, depending
-- on its support for IPv4 and IPv6."

MODULE -- this module
MANDATORY-GROUPS {
    swmSupportedTunnelGroup,
    swmEncapsGroup,
    swmBGPNeighborGroup
}

 ::= { swmCompliances 1 }

swmSupportedTunnelGroup OBJECT-GROUP
OBJECTS {
    swmSupportedTunnelType
}
STATUS current
DESCRIPTION
    "The collection of objects which are used to show
    what kind of tunnel the AFBR supports."
 ::= { swmGroups 1 }

swmEncapsGroup OBJECT-GROUP
OBJECTS {
    swmEncapsIIPDst,
    swmEncapsIIPDstType
}
STATUS current
DESCRIPTION
    "The collection of objects which are used to display
    software mesh encapsulation information."
 ::= { swmGroups 2 }

swmBGPNeighborGroup OBJECT-GROUP
OBJECTS {
```

```
        swmBGPNeighborTunnelType
    }
    STATUS    current
    DESCRIPTION
        "The collection of objects which are used to display
        software mesh BGP neighbor information."
    ::= { swmGroups 3 }
```

END

7. Security Considerations

Because this MIB module reuses the IP tunnel MIB, the security considerations of the IP tunnel MIB is also applicable to the Software mesh MIB.

There are no management objects defined in this MIB module that have a MAX-ACCESS clause of read-write and/or read-create. So, if this MIB module is implemented correctly, then there is no risk that an intruder can alter or create any management objects of this MIB module via direct SNMP SET operations.

Some of the readable objects in this MIB module (i.e., objects with a MAX-ACCESS other than not-accessible) may be considered sensitive or vulnerable in some network environments. It is thus important to control even GET and/or NOTIFY access to these objects and possibly to even encrypt the values of these objects when sending them over the network via SNMP. These are objects and their sensitivity/vulnerability.

Particularly, `swmSupportedTunnelType`, `swmEncapsIIPDstType`, `swmEncapsIIPDst` and `swmBGPNeighborTunnelType` can expose the types of tunnels used within the internal network, and potentially reveal the topology of the internal network.

SNMP versions prior to SNMPv3 did not include adequate security. Even if the network itself is secure (for example by using IPsec), there is no control as to who on the secure network is allowed to access and GET/SET (read/change/create/delete) the objects in this MIB module.

Implementations SHOULD provide the security features described by the SNMPv3 framework (see [RFC3410]), and implementations claiming compliance to the SNMPv3 standard MUST include full support for authentication and privacy via the User-based Security Model (USM) [RFC3414] with the AES cipher algorithm [RFC3826]. Implementations MAY also provide support for the Transport Security Model

(TSM)[RFC5591] in combination with a secure transport such as SSH [RFC5592] or TLS/DTLS [RFC6353].

Further, deployment of SNMP versions prior to SNMPv3 is NOT RECOMMENDED. Instead, it is RECOMMENDED to deploy SNMPv3 and to enable cryptographic security. It is then a customer/operator responsibility to ensure that the SNMP entity giving access to an instance of this MIB module is properly configured to give access to the objects only to those principals (users) that have legitimate rights to indeed GET or SET (change/create/delete) them.

8. IANA Considerations

The MIB module in this document uses the following IANA-assigned OBJECT IDENTIFIER values recorded in the SMI Numbers registry, and the following IANA-assigned tunnelType values recorded in the IANAtunnelType-MIB registry:

Descriptor -----	OBJECT IDENTIFIER value -----
swmMIB	{ mib-2 xxx }

IANAtunnelType ::= TEXTUAL-CONVENTION
SYNTAX INTEGER {
 softwareMesh ("xx") -- software Mesh tunnel
}

9. Acknowledgements

The authors would like to thank Dave Thaler, Jean-Philippe Dionne, Qi Sun, Sheng Jiang, Yu Fu for their valuable comments.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.
- [RFC2578] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, DOI 10.17487/RFC2578, April 1999, <<http://www.rfc-editor.org/info/rfc2578>>.

- [RFC2579] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Textual Conventions for SMIV2", STD 58, RFC 2579, DOI 10.17487/RFC2579, April 1999, <<http://www.rfc-editor.org/info/rfc2579>>.
- [RFC2580] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Conformance Statements for SMIV2", STD 58, RFC 2580, DOI 10.17487/RFC2580, April 1999, <<http://www.rfc-editor.org/info/rfc2580>>.
- [RFC4001] Daniele, M., Haberman, B., Routhier, S., and J. Schoenwaelder, "Textual Conventions for Internet Network Addresses", RFC 4001, DOI 10.17487/RFC4001, February 2005, <<http://www.rfc-editor.org/info/rfc4001>>.
- [RFC3414] Blumenthal, U. and B. Wijnen, "User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)", STD 62, RFC 3414, DOI 10.17487/RFC3414, December 2002, <<http://www.rfc-editor.org/info/rfc3414>>.
- [RFC3826] Blumenthal, U., Maino, F., and K. McCloghrie, "The Advanced Encryption Standard (AES) Cipher Algorithm in the SNMP User-based Security Model", RFC 3826, DOI 10.17487/RFC3826, June 2004, <<http://www.rfc-editor.org/info/rfc3826>>.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, DOI 10.17487/RFC5512, April 2009, <<http://www.rfc-editor.org/info/rfc5512>>.
- [RFC5565] Wu, J., Cui, Y., Metz, C., and E. Rosen, "Softwire Mesh Framework", RFC 5565, DOI 10.17487/RFC5565, June 2009, <<http://www.rfc-editor.org/info/rfc5565>>.
- [RFC5566] Berger, L., White, R., and E. Rosen, "BGP IPsec Tunnel Encapsulation Attribute", RFC 5566, DOI 10.17487/RFC5566, June 2009, <<http://www.rfc-editor.org/info/rfc5566>>.
- [RFC5591] Harrington, D. and W. Hardaker, "Transport Security Model for the Simple Network Management Protocol (SNMP)", STD 78, RFC 5591, DOI 10.17487/RFC5591, June 2009, <<http://www.rfc-editor.org/info/rfc5591>>.

- [RFC5592] Harrington, D., Salowey, J., and W. Hardaker, "Secure Shell Transport Model for the Simple Network Management Protocol (SNMP)", RFC 5592, DOI 10.17487/RFC5592, June 2009, <<http://www.rfc-editor.org/info/rfc5592>>.
- [RFC6353] Hardaker, W., "Transport Layer Security (TLS) Transport Model for the Simple Network Management Protocol (SNMP)", STD 78, RFC 6353, DOI 10.17487/RFC6353, July 2011, <<http://www.rfc-editor.org/info/rfc6353>>.

10.2. Informative References

- [RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, DOI 10.17487/RFC2863, June 2000, <<http://www.rfc-editor.org/info/rfc2863>>.
- [RFC4925] Li, X., Ed., Dawkins, S., Ed., Ward, D., Ed., and A. Durand, Ed., "Softwire Problem Statement", RFC 4925, DOI 10.17487/RFC4925, July 2007, <<http://www.rfc-editor.org/info/rfc4925>>.
- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, DOI 10.17487/RFC3410, December 2002, <<http://www.rfc-editor.org/info/rfc3410>>.
- [RFC4087] Thaler, D., "IP Tunnel MIB", RFC 4087, DOI 10.17487/RFC4087, June 2005, <<http://www.rfc-editor.org/info/rfc4087>>.

Authors' Addresses

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

Phone: +86-10-6260-3059
EMail: yong@csnet1.cs.tsinghua.edu.cn

Jiang Dong
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5822
EMail: knight.dongjiang@gmail.com

Peng Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R.China

Phone: +86-10-6278-5822
EMail: weapon9@gmail.com

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

Phone: +86-10-6278-5822
EMail: xmw@cernet.edu.cn

Antti Yla-Jaaski
Aalto University
Konemiehentie 2
Espoo 02150
Finland

Phone: +358-40-5954222
EMail: antti.yla-jaaski@aalto.fi

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: May 5, 2013

S. Tsuchiya, Ed.
Cisco Systems
S. Ohkubo
Sakura Internet
Y. Kawakami
INTERNET MULTIFEED CO.
Nov 2012

Stateless IPv4 over IPv6 report
draft-janog-softwire-report-01

Abstract

Stateless IPv4 over IPv6 tunnel such as MAP(Mapping of Address and Port) designs to support IPv4 over IPv6 island and resolve IPv4 shortage problem by Address and Port Mapping technique.

This document describes supported vendor's implementation, ipv4 functionality over IPv6 and interoperability report.

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 May 5, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	4
2. Implementation Report	4
2.1. Participant List	4
2.1.1. MAP-E Border Relay (BR)	4
2.1.2. MAP-E Customer Edge (CE)	5
2.2. Security mechanism	5
2.2.1. Question	5
2.2.2. Typical implementation	5
2.3. Provisioning method	6
2.4. Reachability to BR address	6
3. Test Parameter	6
4. IPv4 functionality over IPv6	7
4.1. T-1:ICMP	7
4.2. T-2:IPSec VPN	9
4.3. T-3:SSL VPN	9
4.4. T-4:FTP	9
4.5. T-5:PPTP	9
4.6. T-6:L2TP	9
4.7. T-7:Instant Messaging and VoIP	10
4.7.1. Facebook on the web (http)	10
4.7.2. Facebook via a client (xmpp)	10
4.7.3. Jabber.org chat service (xmpp)	10
4.7.4. Gmail chat on the web (http)	10
4.7.5. Gmail chat via a client (xmpp)	10
4.7.6. Google Talk client	10
4.7.7. AIM (AOL)	10
4.7.8. ICQ (AOL)	10
4.7.9. Skype	10
4.7.10. MSN	10
4.7.11. Webex	11
4.7.12. Sametime	11
4.7.13. facetime	11
4.8. T-8:NAT verification tool	11
4.8.1. T-8-1:RFC4787	11
4.8.2. T-8-2:NAT-Analyzer	11
4.9. Result summary	11
5. BR redundancy	12
6. Interoperability	12
7. Conclusion	13
8. Contributor	13

9. Acknowledgements	13
10. IANA Considerations	14
11. Security Considerations	14
12. References	14
12.1. Normative References	14
12.2. Informative References	15
Appendix A. Additional Stuff	16
A.1. test network topology and parameters	17
A.2. Configuration	17
A.2.1. IP Infusion NetBSD 4.0.1:BR	17
A.2.2. IP Infusion Linux 2.6.18:BR	18
A.2.3. Furukawa Network Solution Corp.:BR	18
A.2.4. Vyatta ASAMAP:BR	18
A.2.5. Internet Initiative Japan Inc. SEIL:BR	19
A.2.6. Cisco IOS-XR:BR	19
A.2.7. IP Infusion NetBSD 4.0.1:CE	19
A.2.8. IP Infusion Linux 2.6.18:CE	20
A.2.9. Furukawa Network Solution Corp.:CE	20
A.2.10. Vyatta ASAMAP:CE	21
A.2.11. Internet Initiative Japan Inc. SEIL:CE	21
A.2.12. Yamaha :CE	21
A.2.13. CERNET OpenWRT :CE	21
Authors' Addresses	22

1. Introduction

Stateless IPv4 over IPv6 tunnel such as MAP(Mapping of Address and Port) designs to support IPv4 over IPv6 island and resolve IPv4 shortage problem by Address and Port Mapping technique.

Japan Network Operators Group [JANOG] made a Working Group to evaluate this feature.

7 vendors and 9 implementations attended to the interop events which hold at Nagaoka city of Niigata, Japan.

This document describes MAP-E [I-D.ietf-softwire-map] supported vendor's implementation, ipv4 functionality over IPv6 and interoperability report.

2. Implementation Report

MAP-E [I-D.ietf-softwire-map] is already supported by a lot of vendors. The total number was 7 vendors and 9 implementations at this point.

In this section, describes about interop event participant list, security mechanism and provisioning method and reachability to BR address.

2.1. Participant List

2.1.1. MAP-E Border Relay (BR)

Vendor	OS/Equipment
IP Infusion	Linux 2.6.18
IP Infusion	NetBSD 4.0.1
Furukawa Network Solution Corp.	FX5000
ASAMAP	Vyatta
Internet Initiative Japan Inc.	SEIL/X1
Cisco Systems	IOS-XR/ASR9000

2.1.2. MAP-E Customer Edge (CE)

Vendor	OS/Equipment
IP Infusion	Linux 2.6.18
IP Infusion	NetBSD 4.0.1
Furukawa Network Solution Corp.	F60W
ASAMAP	Vyatta
CERNET	OpenWRT
Internet Initiative Japan Inc.	SEIL/X1
Yamaha Corporation	RTX1200

2.2. Security mechanism

We took a survey to participant about security considerations which is described on Section 13 of [I-D.ietf-softwire-map].

2.2.1. Question

Q1. How to behave when CE/BR receives IPv4 packets which does not match MAP cpe domain?

Q2. How to behave when CE/BR receives IPv6 packets which has inconsistency between IPv6 src address and IPv4 src address?

Q3. How to behave when CE/BR receives IPv6 packets which has inconsistency between IPv6 dst address and IPv4 dst address?

Q4. How to behave when CE receives IPv6 packets which is not from BR address?

2.2.2. Typical implementation

A1. Most of vendor look up IP routing table, then routing to next hop or drops. But some vendors check routing table at first, then check consistency between the Rule IPv4 prefix and IPv4 destination packets. As the result, routing to next hop or drops.

A2. All of BRs checks inconsistency between IPv6 src address and IPv4 src address. Most of CE checks inconsistency between IPv6 src address and IPv4 src address. If IPv6 src address is included in the Rule IPv6 prefix then validates IPv4 src address. If the src address is BR address, then it does not validate.

A3. Most of BR only checks whether the IPv6 dst address is BR address. Most of CE validates inconsistency between IPv6 dst address

and IPv4 dst address before NAT process.

A4. Depends on configuration. If CE is configured as "Hub and Spoke mode" or permit only from BR address, then the packets will be dropped. If CE is configured as "Mesh mode" or no filter, then packets will transit.

2.3. Provisioning method

Section 7 of [I-D.ietf-softwire-map] describes configuration of MAP-E. All of CE and BR who attended the event are supported manual configuration only at this time.

Most of vendors directly configures "Length of EA bits", but some vendors configures "sharing ratio" and "contiguous-ports", "length of EA bits" would be calculated as the result.

The former configuration type would be useful for operation and trouble shooting, because "rule in the packets" is visible on configurations.

On the other hand, latter type configuration would be easy to understand design such as how many user will be shared one address.

2.4. Reachability to BR address

3 BR implementations are required to configure BR address as interface address.

The rest of 2 BR implementations must not configure BR address as interface address.

The former implementation, can confirm reachability of BR address from IPv6 network. But latter implementation, can not confirm reachability to BR address but of course can confirm reachability to BR themselves.

3. Test Parameter

MAP-E Parameter

Parameter	Value
Rule IPv6 prefix	2403:9200::/32
Rule IPv4 prefix	203.86.225.0/28
End-user IPv6 prefix	2403:9200:fff1::/48 - 2403:9200:fff7::/48
EA bits	16bit(48-32)
Port-Set ID	12bit
PSID offset	4
BR IPv6 address	2403:9200:fff0:0::2
Topology	Mesh

Each of MAP-E has only 15 TCP/UDP ports, 1 IP address is shared by 4096 users.

MAP Simulation Tool shows this rule. [1]

MTU Parameter

Parameter	IPv6 MTU	TCP MSS clamp	Tunnel IF IPv4 MTU
Value	1500byte	Enable	1460byte

4. IPv4 functionality over IPv6

MAP-E [I-D.ietf-softwire-map] uses A+P technologies with NAT44. Basic NAT requirement is defined as [RFC4787], [RFC5508] and [RFC5382]. But there is difference of implementation among vendors. ALG support also depends on vendors implementations.

4.1. T-1:ICMP

Section 9 of [I-D.ietf-softwire-map] describes about ICMP handling in MAP domain.

T-1-1: echo request/echo reply

Confirmed from MAP-E CE network to global internet.

Echo request (ICMP type 8 code 0) has identifier, so MAP-E CE has to rewrite this field from ports-set value. Echo reply (ICMP type 0 code 0) has same identifier as echo request. MAP-BR must handle from this identifier field.

Therefore the test could confirm capability of ICMP both CE and BR.

All of CE and BR are supported this feature.

T-1-2: Host Unreachable

Confirmed from MAP-E CE network to global internet.

Layer3 switch reply "Host unreachable" (ICMP type 3 code 1) message, the message does not has identifier field. So MAP-BR has to inspect ICMP payload.

The test could confirm capability to handling for null identifier ICMP of MAP-BR.

3 BR already supported this feature.2 BR does not support this feature at this time.

T-1-3: TTL equals 0 during transit

Confirmed traceroute from MAP-E CE network to global internet.

Layer3 switch reply "TTL equals 0 during transit" (ICMP type 11 code 0) message, the message does not has identifier field. So MAP-BR has to inspect ICMP payload.

The test could confirm capability to handling for null identifier ICMP of MAP-BR.

3 BR already supported this feature.2 BR does not support this feature at this time.

T-1-4: Fragmentation needed but no frag. bit set

Confirmed Echo with DF-bit from MAP-E CE network to global internet.

Layer3 switch reply "Fragmentation needed but no frag. bit set" (ICMP type 3 code 4) message, the message does not has identifier field. So MAP-BR has to inspect ICMP payload.

The test could confirm capability to handling for null identifier ICMP of MAP-BR.

3 BR already supported this feature.2 BR does not support this feature at this time.

4.2. T-2:IPSec VPN

IPSec VPN [RFC2401] uses ESP packets, therefore MAP-E CE should support NAT traversal [RFC3948].

T-2-1:IPSec

All of CE failed.This result is expected behavior.

T-2-2:IPSec VPN(UDP:NAT Traversal)

All of CE succeeded.

4.3. T-3:SSL VPN

It should be no problem, because SSL VPN[RFC4347] uses TCP sockets.

All of CE succeeded.

4.4. T-4:FTP

FTP[RFC0959] PORT(Active) and PASV(Passive) mode had sometimes problem in NAT44. [RFC2428] is enhancement FTP for IPv6/NAT. MAP-E devices may need support FTP ALG if the customer required FTP Active mode.

T-4-1:Passive(PASV) mode

All of CE succeeded.

T-4-2:Active(PORT) mode

Only 2 vendor's CE succeeded.

4.5. T-5:PPTP

PPTP[RFC2637] uses GRE and TCP port 1723. Unless configuring to pass GRE and TCP port 1723, can not use PPTP on MAP-E .NOTE:Microsoft is warning use of PPTP due to security reason. [2743314]

All of CE failed.This result is expected behavior.

4.6. T-6:L2TP

L2TP/IPsec[RFC3193] should support on MAP-E using with NAT Traversal[RFC3948].

All of CE succeeded.

4.7. T-7:Instant Messaging and VoIP

Verified functionality of Instant Messaging and VoIP tool that described on section 5.3 of [RFC6586] and facetime within same MAP-E CE, different MAP-E CEs and between MAP-E BR and CE.

4.7.1. Facebook on the web (http)

All of combination basically succeeded. Tested both chat and video.

4.7.2. Facebook via a client (xmpp)

All of combination succeeded.

4.7.3. Jabber.org chat service (xmpp)

Not tested

4.7.4. Gmail chat on the web (http)

All of combination basically succeeded. Tested chat,voice and video.

4.7.5. Gmail chat via a client (xmpp)

All of combination basically succeeded. Tested chat,voice and video.

4.7.6. Google Talk client

All of combination basically succeeded. Tested chat,voice and video.

4.7.7. AIM (AOL)

All of combination basically succeeded. Tested chat,voice and video.

4.7.8. ICQ (AOL)

All of combination basically succeeded. Tested chat,voice and video.

4.7.9. Skype

All of combination basically succeeded. Tested chat,voice and video.

4.7.10. MSN

All of combination basically succeeded. Tested chat,voice and video.

4.7.11. Webex

All of combination succeeded. Tested chat, voice and video in the meeting.

4.7.12. Sametime

Not tested

4.7.13. facetime

All of combination succeeded.

4.8. T-8:NAT verification tool

According to Section-11 of [I-D.ietf-softwire-map], MAP CE should support [RFC4787], [RFC5508] and [RFC5382]. This section describes the result of MAP CEs which were verified by test tool.

4.8.1. T-8-1:RFC4787

STUN [RFC5389], NAT Behavior Discovery [RFC5780] and UDP hole punching are used for online game. [RFC4787] is Best Current Practice of NAT behavior requirement for UDP. Konami Digital Entertainment Co., Ltd. provided [RFC4787] verification tool.

The test result will be update.

REQ-1: Endpoint-Independent Mapping

REQ-8: Filtering Behavior

REQ-9: Hairpinning

REQ-3: Port overloading

4.8.2. T-8-2:NAT-Analyzer

[NAT-Analyzer] is JAVA applet in the browser to verify NAT functionality.

The test result will be update.

4.9. Result summary

Even DNS queries could occupy to MAP-E CE NAT table in this port restricted (only 15 ports) environments.

There is two solution; one is specific short timer for DNS or UDP. Another is configured DNS transport proxy on MAP-E CE.

As section-3 of [I-D.draft-dec-stateless-4v6] describes, MAP-E CE expected to act as a DNS resolver proxy, using native DNS over IPv6 to the SP network.

IPv6 MTU expected as 1500byte, but some vendors had the implicit limitation which does not configured over 1280byte. It could not see a lot of site if the vendor which had limitation works as BR. Also there was complex issue even if the vendor works as CE.

As section 10.1 of [I-D.ietf-softwire-map], IPv6 MTU in the MAP domain should configured well managed value.

Most of modern applications and VPN protocols could use in multi vendor MAP-E[I-D.ietf-softwire-map].

But there is the difference of test result of [RFC4787] and FTP Active mode. It's depends on vendor's NAT implementation.

5. BR redundancy

As MAP-E is stateless,so specific technic is not required for redundancy.

Tested BR redundancy between 2 BRs by routing convergence. Skype session had been kept and could communicate after route convergence(about 2 sec).

6. Interoperability

MAP-E stateless technology that means does not need maintenance of state machine. So there are no problem of interoperability.

But there was one issue about "ipv6 interface identifier" from misunderstanding of section-6 of [I-D.ietf-softwire-map].

If it could add example to explain format in this section, then it would be more understandable.

The issue is already fixed.

7. Conclusion

A lot of vendors already implemented MAP-E.

There is no critical issue about interoperability between different vendor's CE and CE, CE and BR, BR and BR.

Most of issue were already discussed on IETF.

MAP-E [I-D.ietf-softwire-map] is mature.

8. Contributor

Test network Contributors

Chisato Kashiwagi Chisato.Kashiwagi@ipinfusion.com
Shuuichi Saito shuu@fnsc.co.jp
Takuya Iimura tiimura@cisco.com
Tomoki Murai murai@fnsc.co.jp
Tomoyuki Sahara tsahara@iij.ad.jp
Ryo Sato sr.10005@konami.com
Motohiko Sato sm.64846@konami.com
Congxiao Bao congxiao@cernet.edu.cn
Guoliang Han bupthgl@gmail.com
Kohei Ono Kohei.Ono@ipinfusion.com
Naohide Kamitani kamitani@fnsc.co.jp
Masakazu Asama m-asama@ginzado.ne.jp
Kazuhiko Satoyoshi satoyoshi@soundnet.yamaha.co.jp
Takehiro Sukizaki sukizaki@soundnet.yamaha.co.jp
Tetsuya Innami tinnami@cisco.com
Satoshi Kubota sa-kubota@jpne.co.jp
Yuji Yamazaki yuji.yamazaki@g.softbank.co.jp
Satoru Matsushima satoru.matsushima@gmail.com
Kaoru Oka kaoru.oka@g.softbank.co.jp
Miki Takata miki@baking.jp
Takayuki Osabe osabet@nscs.jp
Satoshi Ebe satoshie@nscs.jp
Yasuyuki Kaneko yasuyuki.kaneko@global-netcore.jp
Maoke Chen fibrib@gmail.com

9. Acknowledgements

The author would like to thanks NS Computer Service, ENOG and JANOG committee. The authors would like to thank you Satoru Matsushima, Seiichi Kawamura for their thorough review and comments.

10. IANA Considerations

This document has no actions for IANA.

11. Security Considerations

There is no additional security requirement.

12. References

12.1. Normative References

- [I-D.dec-stateless-4v6]
Dec, W., "Stateless 4via6 Address Sharing",
draft-dec-stateless-4v6-00 (work in progress), March 2011.
- [I-D.ietf-softwire-map]
Troan, O., Dec, W., Li, X., Bao, C., Zhai, Y., Matsushima,
S., and T. Murakami, "Mapping of Address and Port (MAP)",
draft-ietf-softwire-map-00 (work in progress), June 2012.
- [I-D.murakami-softwire-4rd]
Murakami, T. and O. Troan, "IPv4 Residual Deployment on
IPv6 infrastructure - protocol specification",
draft-murakami-softwire-4rd-00 (work in progress),
July 2011.
- [RFC0959] Postel, J. and J. Reynolds, "File Transfer Protocol",
STD 9, RFC 959, October 1985.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2401] Kent, S. and R. Atkinson, "Security Architecture for the
Internet Protocol", RFC 2401, November 1998.
- [RFC2637] Hamzeh, K., Pall, G., Verthein, W., Taarud, J., Little,
W., and G. Zorn, "Point-to-Point Tunneling Protocol",
RFC 2637, July 1999.
- [RFC3193] Patel, B., Aboba, B., Dixon, W., Zorn, G., and S. Booth,
"Securing L2TP using IPsec", RFC 3193, November 2001.
- [RFC3948] Huttunen, A., Swander, B., Volpe, V., DiBurro, L., and M.
Stenberg, "UDP Encapsulation of IPsec ESP Packets",
RFC 3948, January 2005.

- [RFC4347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security", RFC 4347, April 2006.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC5508] Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT Behavioral Requirements for ICMP", BCP 148, RFC 5508, April 2009.
- [RFC5780] MacDonald, D. and B. Lowekamp, "NAT Behavior Discovery Using Session Traversal Utilities for NAT (STUN)", RFC 5780, May 2010.
- [RFC6586] Arkko, J. and A. Keranen, "Experiences from an IPv6-Only Network", RFC 6586, April 2012.

12.2. Informative References

- [2743314] ""Microsoft Security Advisory (2743314) Unencapsulated MS-CHAP v2 Authentication Could Allow Information Disclosure "", <<http://technet.microsoft.com/en-us/security/advisory/2743314>>.
- [ENOG] ""Echigo Network Operators' Group"", <<http://enog.jp/>>.
- [I-D.ietf-softwire-stateless-4v6-motivation] Boucadair, M., Matsushima, S., Lee, Y., Bonness, O., Borges, I., and G. Chen, "Motivations for Stateless IPv4 over IPv6 Migration Solutions", draft-ietf-softwire-stateless-4v6-motivation-00 (work in progress), September 2011.
- [JANOG] ""Japan Network Operators Group"", <<http://www.janog.gr.jp/en>>.
- [NAT-Analyzer] ""Network Measurement Activities at TUM"", <<http://natatest.net.in.tum.de/>>.
- [RFC3849] Huston, G., Lord, A., and P. Smith, "IPv6 Address Prefix Reserved for Documentation", RFC 3849, July 2004.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.

- [RFC5387] Touch, J., Black, D., and Y. Wang, "Problem and Applicability Statement for Better-Than-Nothing Security (BTNS)", RFC 5387, November 2008.
- [RFC5569] Despres, R., "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd)", RFC 5569, January 2010.
- [RFC5737] Arkko, J., Cotton, M., and L. Vegoda, "IPv4 Address Blocks Reserved for Documentation", RFC 5737, January 2010.
- [RFC5952] Kawamura, S. and M. Kawashima, "A Recommendation for IPv6 Address Text Representation", RFC 5952, August 2010.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6104] Chown, T. and S. Venaas, "Rogue IPv6 Router Advertisement Problem Statement", RFC 6104, February 2011.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.
- [RFC6346] Bush, R., "The Address plus Port (A+P) Approach to the IPv4 Address Shortage", RFC 6346, August 2011.

URIs

- [1] <<http://6lab.cisco.com/map/MAP.php?WlsiUnVsZSAwIiwMjQwMzo5MjAwOjA6MC8zMlIsIjIwMy44Ni4yMjUuMC8yOCIsMTYsNCwxXV0=>>>

Appendix A. Additional Stuff

A.1. test network topology and parameters

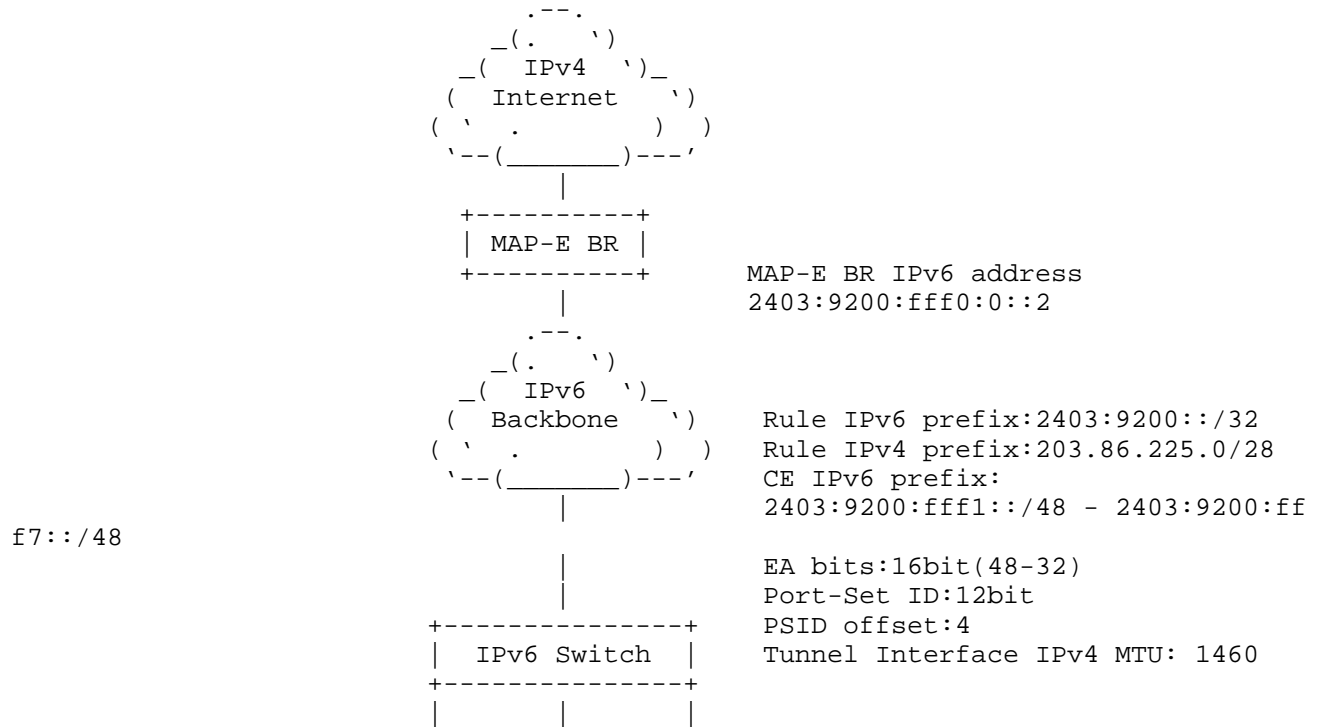


Figure 1

A.2. Configuration

A.2.1. IP Infusion NetBSD 4.0.1:BR

```

---- MAP tunnel I/F ----
# cat /etc/ifconfig.map0
up
mtu 1460
inet 10.99.99.1/24
rule_ipv6_prefix 2403:9200::/32
rule_ipv4_prefix 203.86.225.0/28
rule_eabits_length 16
psid_offset 4
encap_src_check 0
fmr 1

```

A.2.2. IP Infusion Linux 2.6.18:BR

```
---- MAP tunnel I/F ----
ip -6 tunnel change map1 rule_ipv6_prefix 2403:9200::/32
ip -6 tunnel change map1 rule_ipv4_prefix 203.86.225.0/28
ip -6 tunnel change map1 rule_eabits_length 16
ip -6 tunnel change map1 psid_offset 4
ip -6 tunnel change map1 fmr 1
ip -6 tunnel change map1 map_autosetaddr 1
ip -6 tunnel change map1 map_autosetgw 1
ip -4 addr add 203.86.225.18/30 dev map1
```

A.2.3. Furukawa Network Solution Corp.:BR

```
!
interface tunnel 1
 tunnel mode ipinip ipv4 ipv6-tunnel-profile 1
exit
!
ipv4 ipv6-tunnel-profile 1
 profile-mode map-encap
 rule-ipv4-prefix 203.86.225.0/28
 rule-ipv6-prefix 2403:9200::/32
 user-len 16
 source-address 2403:9200:fff0::2
exit
```

A.2.4. Vyatta ASAMAP:BR

```
interfaces {
  loopback lo {
  }
  map map0 {
    br-address 2403:9200:fff0::2/64
    default-forwarding-mode encapsulation
    default-forwarding-rule true
    role br
    rule 1 {
      ea-length 16
      ipv4-prefix 203.86.225.0/28
      ipv6-prefix 2403:9200::/32
    }
  }
}
```

A.2.5. Internet Initiative Japan Inc. SEIL:BR

```
interface frd0 mtu 1460
frd mode br
frd br-address 2403:9200:fff0::2
frd rule add R0 external-ipv4-prefix 203.86.225.0/28 internal-ipv6-prefix 2403:92
00::/32 index-length 16 psid-offset 4
```

A.2.6. Cisco IOS-XR:BR

```
!
interface ServiceApp5
  ipv4 address 203.0.113.5 255.255.255.252
  load-interval 30
  service cgn JANOG service-type map-e
  logging events link-status
!
interface ServiceApp6
  ipv6 address 2001:db8:1:1::1/64
  load-interval 30
  service cgn JANOG service-type map-e
  logging events link-status
!
interface ServiceInfral
  ipv4 address 203.0.113.1 255.255.255.252
  service-location 0/0/CPU0
!
service cgn JANOG
  service-location preferred-active 0/0/CPU0
  service-type map-e Software
  cpe-domain ipv4 prefix 203.86.225.0/28
  cpe-domain ipv6 prefix 2403:9200::/32
  sharing-ratio 12
  aftr-endpoint-address 2403:9200:fff0::2
  contiguous-ports 0
  address-family ipv4
    interface ServiceApp5
  !
  address-family ipv6
    interface ServiceApp6
  !
end
```

A.2.7. IP Infusion NetBSD 4.0.1:CE

```
-- ifconfig.map0 -- actual MAP-E parameter
up
mtu 1460
rule_ipv6_prefix 2403:9200::/32
rule_ipv4_prefix 203.86.225.0/28
lan_if_name any
wan_if_name lol
map_autosetaddr 1
map_autosetgw 1
rule_eabits_length 16
psid_offset 4
map_border_router 2403:9200:fff0:0::2
map_mss auto
fmr 1
```

A.2.8. IP Infusion Linux 2.6.18:CE

---- MAP tunnel I/F ----

```
ip -6 tunnel change map1 rule_ipv6_prefix 2403:9200::/32
ip -6 tunnel change map1 rule_ipv4_prefix 203.86.225.0/28
ip -6 tunnel change map1 rule_eabits_length 16
ip -6 tunnel change map1 psid_offset 4
ip -6 tunnel change map1 map_border_router 2403:9200:fff0:0::2
ip -6 tunnel change map1 fmr 1
ip -6 tunnel change map1 map_autosetaddr 1
ip -6 tunnel change map1 map_autosetgw 1
```

A.2.9. Furukawa Network Solution Corp.:CE

```
ip nat ap_pool ADDR-POOL1
  ipv6-mape-profile 1
exit
ip ipv6-mape-profile 1
  user-len 16
  br-address 2403:9200:fff0:0::2
  rule-ipv6-prefix reference-interface loopback 1
  rule-ipv6-prefix-len 32
  rule-ipv4-prefix 203.86.225.0 255.255.255.240
exit
interface tunnel 1
  tunnel mode ipip ipv6-mape-profile 1
  tunnel source reference-interface ipv6 loopback 1
  ip mtu 1460
  ip nat inside source list 10 ap_pool ADDR-POOL1
exit
```


A.2.10. Vyatta ASAMAP:CE

```
interfaces {
    map map0 {
        br-address 2403:9200:fff0::2/64
        default-forwarding-mode encapsulation
        default-forwarding-rule true
        role ce
        rule 1 {
            ea-length 16
            ipv4-prefix 203.86.225.0/28
            ipv6-prefix 2403:9200::/32
        }
        tunnel-source eth1
    }
}
```

A.2.11. Internet Initiative Japan Inc. SEIL:CE

```
interface frd0 mtu 1460
interface frd0 tcp-mss 1420
nat napt add private 192.168.1.0-192.168.1.255 interface frd0
nat option port-assignment random
frd mode ce
frd ce-address 2403:9200:fff6:0:cb:56e1:f0f:f600
frd br-address 2403:9200:fff0::2
frd rule add R0 external-ipv4-prefix 203.86.225.0/28 internal-ipv6-prefix 2403:9200::/32 index-length 16 psid-offset 4
```

A.2.12. Yamaha :CE

```
tunnel select 1
tunnel encapsulation ipip
tunnel endpoint address 2403:9200:fff7:0:cb:56e1:f0f:f700 2403:9200:fff0::2
tunnel map-e 203.86.225.0/28 2403:9200::/32 16 4 ::2
ip tunnel mtu 1460
ip tunnel nat descriptor 1000
tunnel enable 1
nat descriptor type 1000 masquerade
nat descriptor timer 1000 30
nat descriptor timer 1000 tcpfin 5
nat descriptor address outer 1000 map-e
```

A.2.13. CERNET OpenWRT :CE

```
# configure eth1 -- IPv4 interface
ifconfig br-lan 192.168.1.1/24

./control start

#janog software interop event
utils/ivictl -r -d -P 2403:9200:fff0:0::2/128
utils/ivictr -r -p 203.86.225.0/28 -P 2403:9200::/32 -R 4096 - M 1 -f -E
utils/ivictr -s -i br-lan -I eth0.1 -H -N -a 192.168.1.0/24 -A 203.86.225.1/28 -P
  2403:9200::/32 -R 4096 - M 1 -o 4 -f -c 1400 -E
service iptables stop
service ip6tables stop
```

Authors' Addresses

Shishio Tsuchiya (editor)
Cisco Systems
Midtown Tower, 9-7-1, Akasaka
Minato-Ku, Tokyo 107-6227
Japan

Phone: +81 3 6434 6543
Email: shtsuchi@cisco.com

Shuichi Ohkubo
Sakura Internet
33F Sumitomo fudosan Nishi shinjuku Bldg., 7-20-1 Nishi shinjuku
Shinjuku-Ku, Tokyo 160-0023
Japan

Phone: +81 3 5332 7070
Email: ohkubo@sakura.ad.jp

Yuya Kawakami
INTERNET MULTIFEED CO.
OTEMACHI 1st.SQUARE EAST TOWER, 3F 1-5-1, Otemachi,
Chiyoda-ku, Tokyo 100-0004
Japan

Phone: +81 3 3282 1040
Email: kawakami@mfeed.ad.jp

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: October 29, 2013

Sheng Jiang (Editor)
Yu Fu
Bing Liu
Huawei Technologies Co., Ltd
Peter Deacon
IEA Software, Inc.
April 27, 2013

RADIUS Attribute for MAP

draft-jiang-software-map-radius-04.txt

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 October 29, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

Mapping of Address and Port (MAP) is a stateless mechanism for running IPv4 over IPv6-only infrastructure. It provides both IPv4 and IPv6 connectivity services simultaneously during the IPv4/IPv6 co-existing period. The Dynamic Host Configuration Protocol for IPv6 (DHCPv6) MAP options has been defined to configure MAP Customer Edge (CE). However, in many networks, the configuration information may be stored in Authentication Authorization and Accounting (AAA) servers while user configuration is mainly from Broadband Network Gateway (BNG) through DHCPv6 protocol. This document defines a Remote Authentication Dial In User Service (RADIUS) attribute that carries MAP configuration information from AAA server to BNG. The MAP RADIUS attribute are designed following the simplify principle. It provides just enough information to form the correspondent DHCPv6 MAP option.

Table of Contents

1. Introduction	3
2. Terminology	3
3. MAP Configuration process with RADIUS	3
4. Attributes	6
4.1. MAP-Configuration Attribute	6
4.2. MAP Rule Options	6
4.3. Sub Options for MAP Rule Option	7
4.3.1. Rule-IPv6-Prefix Sub Option	7
4.3.2. Rule-IPv4-Prefix Sub Option	8
4.3.3. Encapsulation/Translation Flag Sub Option.....	9
4.3.4. PSID Sub Option	10
4.3.5. PSID Length Sub Option	10
4.3.6. PSID Offset Sub Option	11
4.4. Table of attributes	11
5. Diameter Considerations	12
6. Security Considerations	12
7. IANA Considerations	12
8. Acknowledgments	12
9. References	13
9.1. Normative References	13
9.2. Informative References	13

1. Introduction

Recently providers start to deploy IPv6 and consider how to transit to IPv6. Mapping of Address and Port (MAP)

[I-D.ietf-software-map] is a stateless mechanism for running IPv4 over IPv6-only infrastructure. It provides both IPv4 and IPv6 connectivity services simultaneously during the IPv4/IPv6 co-existing period. MAP has adopted Dynamic Host Configuration Protocol for IPv6 (DHCPv6) [RFC3315] as auto-configuring protocol. The MAP Customer Edge (CE) uses the DHCPv6 extension options [I-D.mdt-software-map-dhcp-option] to discover MAP Border Relay (in tunnel model only) and to configure relevant MAP rules.

In many networks, user configuration information may be managed by AAA (Authentication, Authorization, and Accounting) servers. Current AAA servers communicate using the Remote Authentication Dial In User Service (RADIUS) [RFC2865] protocol. In a fixed line broadband network, the Broadband Network Gateways (BNGs) act as the access gateway of users. The BNGs are assumed to embed a DHCPv6 server function that allows them to locally handle any DHCPv6 requests initiated by hosts.

Since the MAP configuration information is stored in AAA servers and user configuration is mainly through DHCPv6 protocol between BNGs and hosts/CEs, new RADIUS attributes are needed to propagate the information from AAA servers to BNGs. The MAP RADIUS attribute are designed following the simplify principle, while providing enough information to form the correspondent DHCPv6 MAP option. [I-D.mdt-software-map-dhcp-option].

2. Terminology

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 [RFC2119].

The terms MAP CE and MAP Border Relay are defined in [I-D.ietf-software-map].

3. MAP Configuration process with RADIUS

The below Figure 1 illustrates how the RADIUS protocol and DHCPv6 cooperate to provide MAP CE with MAP configuration information.

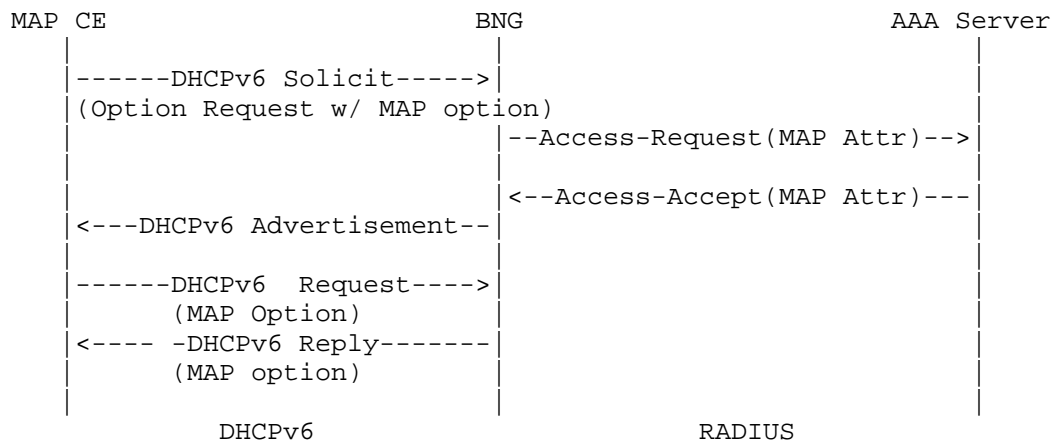


Figure 1: the cooperation between DHCPv6 and RADIUS combining with RADIUS authentication

BNGs act as a RADIUS client and as a DHCPv6 server. First, the MAP CE MAY initiate a DHCPv6 Solicit message that includes an Option Request option (6) [RFC3315] with the MAP option [draft-ietf-softwire-map-dhcp] from the MAP CE. But note that the ORO (Option Request option) with the MAP option could be optional if the network was planned as MAP-enabled as default. When BNG receives the SOLICIT, it SHOULD initiate radius Access-Request message, in which the User-Name attribute (1) SHOULD be filled by the MAP CE MAC address, to the RADIUS server and the User-password attribute (2) SHOULD be filled by the shared MAP password that has been preconfigured on the DHCPv6 server, requesting authentication as defined in [RFC2865] with MAP-Configuration attribute, defined in the next Section. If the authentication request is approved by the AAA server, an Access-Accept message MUST be acknowledged with the IPv6-MAP-Configuration Attribute. After receiving the Access-Accept message with MAP-Configuration Attribute, the BNG SHOULD respond the user an Advertisement message. Then the user can requests for a MAP Option, the BNG SHOULD reply the user with the message containing the MAP option. The recommended format of the MAC address is as defined in Calling-Station-Id (Section 3.20 in [RFC3580]) without the SSID (Service Set Identifier) portion.

Figure 2 describes another scenario, in which the authorization operation is not coupled with authentication. Authorization relevant to MAP is done independently after the authentication process. As similar to above scenario, the ORO with the MAP option in the initial DHCPv6 request could be optional if the network was planned as MAP-enabled as default.

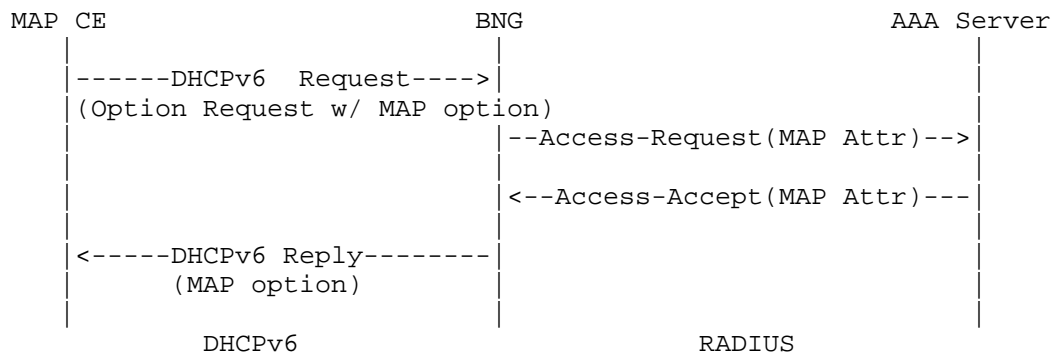


Figure 2: the cooperation between DHCPv6 and RADIUS decoupled with RADIUS authentication

In the abovementioned scenario, the Access-Request packet SHOULD contain a Service-Type attribute (6) with the value Authorize Only (17); thus, according to [RFC5080], the Access-Request packet MUST contain a State attribute that obtained from the previous authentication process.

In both above-mentioned scenarios, Message-authenticator (type 80) [RFC2865] SHOULD be used to protect both Access-Request and Access-Accept messages.

After receiving the MAP-Configuration Attribute in the initial Access-Accept, the BNG SHOULD store the received MAP configuration parameters locally. When the MAP CE sends a DHCPv6 Request message to request an extension of the lifetimes for the assigned address, the BNG does not have to initiate a new Access-Request towards the AAA server to request the MAP configuration parameters. The BNG could retrieve the previously stored MAP configuration parameters and use them in its reply.

If the BNG does not receive the MAP-Configuration Attribute in the Access-Accept it MAY fallback to a pre-configured default MAP configuration, if any. If the BNG does not have any pre-configured default MAP configuration or if the BNG receives an Access-Reject, the tunnel cannot be established.

As specified in [RFC3315], section 18.1.4, "Creation and Transmission of Rebind Messages ", if the DHCPv6 server to which the DHCPv6 Renew message was sent at time T1 has not responded by time T2, the MAP CE (DHCPv6 client) SHOULD enter the Rebind state and attempt to contact any available server. In this situation, the secondary BNG receiving the DHCPv6 message MUST initiate a new Access-Request towards the AAA

server. The secondary BNG MAY include the MAP-Configuration Attribute in its Access-Request.

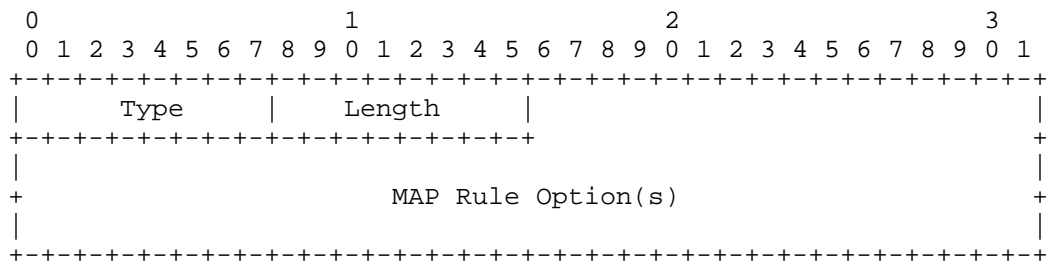
4. Attributes

This section defines MAP-Rule Attribute which is used in the MAP scenario. The attribute design follows [RFC6158] and referring to [I-D.ietf-radext-radius-extensions].

The MAP RADIUS attribute are designed following the simplify principle. The sub options are organized into two categories: the necessary and the optional.

4.1. MAP-Configuration Attribute

The MAP-Configuration Attribute is structured as follows:



Type

TBD

Length

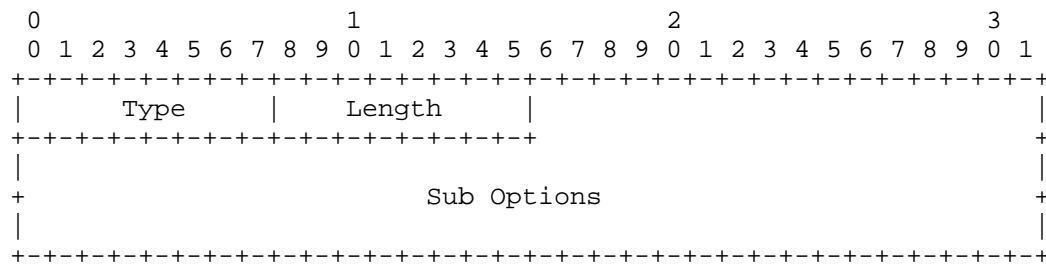
2 + the length of the Rule option(s)

MAP Rule Option (s)

A variable field that may contains one or more Rule option(s), defined in Section 4.2.

4.2. MAP Rule Options

Depending on deployment scenario, one Default Mapping rule and zero or more other type Mapping Rules MUST be included in one MAP-Configuration Attribute.



Type

- 1 Basic Mapping Rule (Not Forwarding Mapping Rule)
- 2 Forwarding Mapping Rule (Not Basic Mapping Rule)
- 3 Default Mapping Rule
- 4 Basic & Forwarding Mapping Rule

Length

- 2 + the length of the sub options

Sub Option

A variable field that contains necessary sub options defined in Section 4.3 and zero or several optional sub options, defined in Section 4.4.

4.3. Sub Options for MAP Rule Option

The sub options do not include EA-Len Embedded-Address length , because it can be calculated by the combine of prefix4len, prefix6-len, PSID and offset bits.

4.3.1. Rule-IPv6-Prefix Sub Option

The Rule-IPv6-Prefix Sub Option is necessary for every MAP Rule option. It should appear for once and only once.

The IPv6 Prefix sub option is follow the framed IPv6 prefix designed in [RFC3162].



SubType

1 (SubType number, for the Rule-IPv6-Prefix6 sub option)

SubLen

20 (the length of the Rule-IPv6-Prefix6 sub option)

Reserved

Reserved for future usage. It should be set to all zero.

prefix6-len

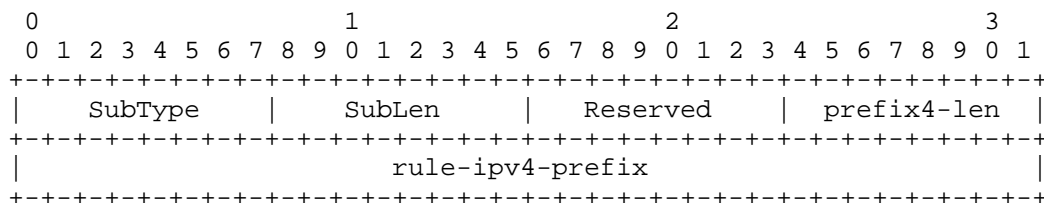
length of the IPv6 prefix, specified in the rule-ipv6-prefix field, expressed in bits

rule-ipv6-prefix

a 128-bits field that specifies an IPv6 prefix that appears in a MAP rule

"For the encapsulation mode the Rule IPv6 prefix can be the full IPv6 address of the BR." [I-D.ietf-software-map]

4.3.2. Rule-IPv4-Prefix Sub Option



SubType

2 (SubType number, for the Rule-IPv4-Prefix6 sub option)

SubLen

8 (the length of the Rule-IPv4-Prefix6 sub option)

Reserved

Reserved for future usage. It should be set to all zero.

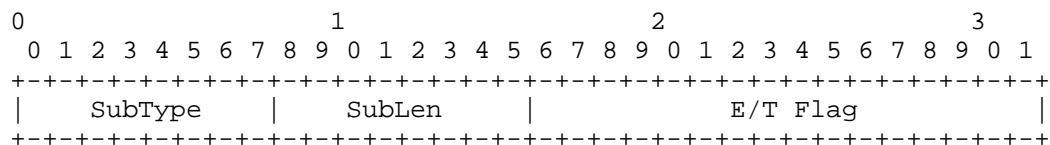
Prefix4-len

length of the IPv6 prefix, specified in the rule-ipv6-prefix field, expressed in bits

rule-ipv4-prefix

a 32-bits field that specifies an IPv4 prefix that appears in a MAP rule

4.3.3. Encapsulation/Translation Flag Sub Option



SubType

3 (SubType number, for the E/T flag sub option)

SubLen

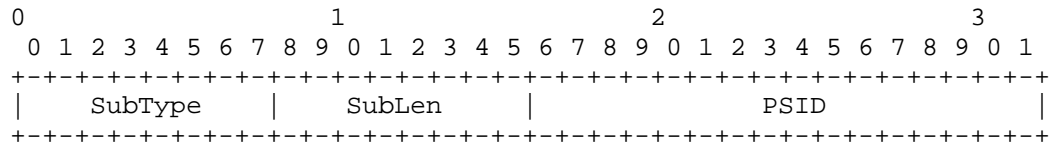
4 (the length of the E/T flag sub option)

E/T Flag

indicate the MAP transport mode: encapsulation or translation.
all 0 for encapsulation, all 1 for translation.

If this sub option is not present, the default is to be assumed as encapsulation mode.

4.3.4. PSID Sub Option



SubType

4 (SubType number, for the PSID Sub Option sub option)

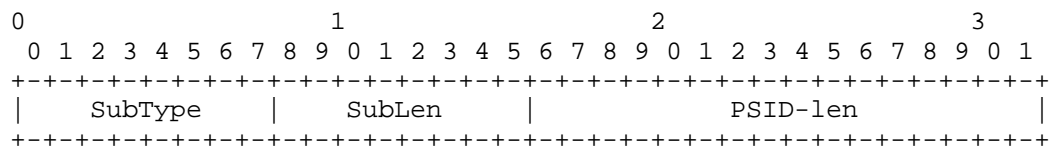
SubLen

4 (the length of the PSID Sub Option sub option)

PSID (Port-set ID)

Explicit 16-bit (unsigned word) PSID value. The PSID value algorithmically identifies a set of ports assigned to a CE. The first k-bits on the left of this 2-octets field is the PSID value. The remaining (16-k) bits on the right are padding zeros.

4.3.5. PSID Length Sub Option



SubType

5 (SubType number, for the PSID Length sub option)

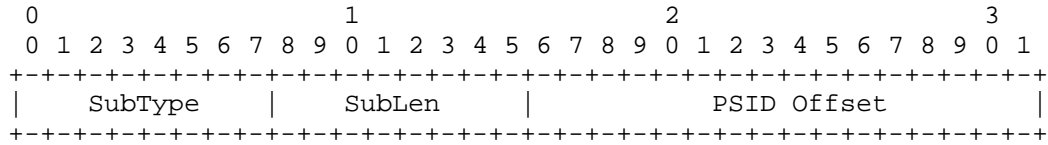
SubLen

4 (the length of the PSID Length sub option)

PSID-len

Bit length value of the number of significant bits in the PSID field. (also known as 'k'). When set to 0, the PSID field is to be ignored. After the first 'a' bits, there are k bits in the port number representing valid of PSID. Subsequently, the address sharing ratio would be 2^k .

4.3.6. PSID Offset Sub Option



SubType

6 (SubType number, for the PSID Offset sub option)

SubLen

4 (the length of the PSID Offset sub option)

PSID Offset

4 bits long field that specifies the numeric value for the MAP algorithm's excluded port range/offset bits (A-bits), as per section 5.1.1 in [I-D.ietf-software-map]. Default must be set to 4.

4.4. Table of attributes

The following table provides a guide to which attributes may be found in which kinds of packets, and in what quantity.

Request	Accept	Reject	Challenge	Accounting	#	Attribute
				Request		
0-1	0-1	0	0	0-1	TBD1	MAP-Configuration
0-1	0-1	0	0	0-1	1	User-Name
0-1	0	0	0	0-1	2	User-Password
0-1	0-1	0	0	0-1	6	Service-Type
0-1	0-1	0-1	0-1	0-1	80	Message-Authenticator

The following table defines the meaning of the above table entries.

0	This attribute MUST NOT be present in packet.
0+	Zero or more instances of this attribute MAY be present in packet.
0-1	Zero or one instance of this attribute MAY be present in packet.
1	Exactly one instance of this attribute MUST be present in packet.

5. Diameter Considerations

This attribute is usable within either RADIUS or Diameter [RFC6733]. Since the Attributes defined in this document will be allocated from the standard RADIUS type space, no special handling is required by Diameter entities.

6. Security Considerations

In MAP scenarios, both CE and BNG are within a provider network, which can be considered as a closed network and a lower security threat environment. A similar consideration can be applied to the RADIUS message exchange between BNG and the AAA server.

Known security vulnerabilities of the RADIUS protocol are discussed in RFC 2607 [RFC2607], RFC 2865 [RFC2865], and RFC 2869 [RFC2869]. Use of IPsec [RFC4301] for providing security when RADIUS is carried in IPv6 is discussed in RFC 3162 [RFC3162].

A malicious user may use MAC address proofing and/or dictionary attack on the shared MAP password that has been preconfigured on the DHCPv6 server to get unauthorized MAP configuration information.

Security considerations for MAP specific between MAP CE and BNG are discussed in [I-D.ietf-softwire-map]. Furthermore, generic DHCPv6 security mechanisms can be applied DHCPv6 intercommunication between MAP CE and BNG.

Security considerations for the Diameter protocol are discussed in [RFC6733].

7. IANA Considerations

This document requires the assignment of two new RADIUS Attributes Types in the "Radius Types" registry (currently located at <http://www.iana.org/assignments/radius-types> for the following attributes:

- o MAP-Configuration TBD1

IANA should allocate the numbers from the standard RADIUS Attributes space using the "IETF Review" policy [RFC5226].

8. Acknowledgments

The authors would like to thank for valuable comments from Peter Lothberg, Wojciech Dec, and Suresh Krishnan .etc.

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.
- [RFC2865] Rigney, C., Willens, S., Rubens, A., and W. Simpson, "Remote Authentication Dial In User Service (RADIUS)", RFC 2865, June 2000.
- [RFC3162] Aboba, B., Zorn, G., and D. Mitton, "RADIUS and IPv6", RFC 3162, August 2001.
- [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.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.
- [RFC5080] Nelson, D. and DeKok A., "Common Remote Authentication Dial In User Service (RADIUS) Implementation Issues and Suggested Fixes", RFC 5080, December 2007.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 5226, May 2008.
- [RFC6158] DeKok, A. and G. Weber, "RADIUS Design Guidelines", RFC 6158, March 2011.
- [RFC6733] V. Fajardo, Ed., J. Arkko, J. Loughney, G. Zorn, Ed., "Diameter Base Protocol", RFC 6733, October 2012.
- [I-D.ietf-software-map]
O. Troan, et al., "Mapping of Address and Port (MAP)",
draft-ietf-software-map, working in progress.
- [I-D.mdt-software-map-dhcp-option]
T. Mrugalski, et al., "DHCPv6 Options for Mapping of
Address and Port", draft-mdt-software-map-dhcp-option,
working in progress.

9.2. Informative References

- [RFC2607] Aboba, B. and J. Vollbrecht, "Proxy Chaining and Policy Implementation in Roaming", RFC 2607, June 1999.

[RFC2869] Rigney, C., Willats, W., and P. Calhoun, "RADIUS Extensions", RFC 2869, June 2000.

[I-D.ietf-radext-radius-extensions]
DeKok, A. and A. Lior, "Remote Authentication Dial In User Service (RADIUS) Protocol Extensions", draft-ietf-radext-radius-extensions, work in process.

Author's Addresses

Sheng Jiang
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd.
Hai-Dian District, Beijing 100095
P.R. China
Email: jiangsheng@huawei.com

Yu Fu
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd.
Hai-Dian District, Beijing 100095
P.R. China
Email: eleven.fuyu@huawei.com

Bing Liu
Huawei Technologies Co., Ltd
Huawei Building, 156 Beiqing Rd.
Hai-Dian District, Beijing 100095
P.R. China
Email: leo.liubing@huawei.com

Peter Deacon
IEA Software, Inc.
P.O. Box 1170
Veradale, WA 99037
USA
Email: peterd@iea-software.com

software
Internet-Draft
Intended status: Informational
Expires: April 16, 2016

R. Maglione, Ed.
W. Dec
Cisco Systems
I. Leung
Rogers Communications
E. Mallette
Bright House Networks
October 14, 2015

Use cases for MAP-T
draft-maglione-software-map-t-scenarios-06

Abstract

The Software working group standardized both encapsulation and translation based stateless IPv4/IPv6 solutions in order to be able to provide IPv4 connectivity to customers in an IPv6-Only environment.

The purpose of this document is to describe some operational use cases that would benefit from a translation based approach and highlights the operational benefits that a translation based solution would allow.

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 16, 2016.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Operational Service Policy Use Cases	3
2.1. Network/Transport Layer Classification classifiers	4
2.2. Device Configuration (DOCSIS)	5
2.3. Service Flow management using Deep Packet Inspection	5
2.4. Service Flow Redirection Policies (Web-redirection)	6
2.5. Service Flow Caching	7
3. Technological Considerations	7
3.1. Encapsulation and Translation Overhead	8
3.2. Efficient Utilization of the Access Network	8
3.2.1. Jumbo Frame Support in the Access	8
3.2.2. Operator Added Packet Overhead and Service Level Agreements	9
4. Conclusions	10
5. Acknowledgements	10
6. IANA Considerations	10
7. Security Considerations	10
8. Informative References	10
Authors' Addresses	10

1. Introduction

The Software working group standardized both encapsulation [RFC7597] and translation [RFC7599] based stateless IPv4/IPv6 solutions developed for the purposes of offering IPv4 connectivity to the customers in an IPv6-Only environment.

There are deployment scenarios that may benefit equally from an encapsulated or translated form of an IPv4/IPv6 stateless addressing solution. There are, however, use cases where using a translation approach could lead to significant operational benefits and potential savings for the operators.

This document describes some use cases that would take advantage of a translation based solution, by highlighting the operational benefits that a translation based approach would allow.

2. Operational Service Policy Use Cases

In Broadband Networks it is common practice for Operators to apply per-subscriber policies on subscriber traffic at the network edge such as a BNG (Broadband Network Gateway), CMTS (Cable Modem Termination System), PGW (PDN Gateway) or like device. Various services may require the application of different policies at these services edges.

Typically a policy would include a classification function and an action function.

- o Service flow classification may occur based on any combination of the following:
 - * Layer-3 identifiers such as source, destination address, protocol or next header, DSCP or Traffic Class;
 - * Layer-4 identifiers such as source or destination port;
 - * service type/destination (i.e. Internet, network service, or other service)
- o Actions may be provisioned against the classified traffic; the following are some examples of actions:
 - * application of different QoS treatment (could be rate-limit, drop, redirect,.. etc) based on Layer 3 or higher layer (Layer 4-7) classification from devices like deep packet inspection appliances;
 - * Service flow redirection on selected types of traffic (i.e. Web portal);
 - * Service flow caching on selected types of traffic.

The rationale for applying such policy at the network edge is based on how tightly coupled this layer of the network is with many key systems within the operators network such as RADIUS, DHCP, access technology awareness and ability to implement subscriber awareness.

In many common deployments today, the customer's policies are maintained in RADIUS server or enforced through other provisioned data in co-operation with service activation such as DHCP and bootstrap configuration. In a cable operator network, while much of the heavily lifting of subscriber management is embedded on the CMTS or OLT, the reality is that classification is shared across CMTS and cable-modem (CM) or across OLT and optical network unit (ONU.) The

CM and ONU classification capabilities are not as robust and flexible as the upstream CMTS, OLT and/or assisting edge router. The implications of that are that the CPE may need to be replaced with a device that has the capability to classify on a larger packet header.

An additional point to consider is that the edge network nodes are also often fitted with, or co-located with higher functioning appliances that employ Deep Packet Inspection and distributed caches used to enhance service performance.

2.1. Network/Transport Layer Classification classifiers

Most of the policies described in Section 2 require the use of network and transport layer classification and filtering mechanisms such as classifiers at the network edge. The application of classifiers and other network layer classification functions on selected subscriber flows are often applied by a AAA server, gleaned from configuration information, provisioned from per-CM DOCSIS configuration files generated from the operator OSS, or sent by a policy control function (PCRF, PCMM, etc).

This section will explain why the application of some types of classifiers (like Layer 3 destination based classifiers and - Layer 3 plus Layer 4 - classifiers,) can be deployed in a more simplistic fashion when using a translated form of a stateless IPv4/IPv6 transition technology such as MAP-T [RFC7599].

A key characteristic of MAP-T is the mapping of the IPv4 address of any destination into the IPv6 destination address, by means of IPv4 to IPv6 mapping rules. This mapping means that the subscriber flows are native IPv6 flows within the operators network. The ability to use a standard IPv6 classifier to identify interesting traffic for classification is well aligned with traditional traffic identification capabilities using IPv4 based classifiers. Such classifiers can be easily applied at the access edge as a standard function commonly available on most platforms deployed.

In contrast, a solution utilizing an IP tunnel based transport (MAP-E [RFC7597] or DS-Lite [RFC6333]), effectively hides the payload's IP layer information, making it difficult to identify by means of an IPv6 classifier. The operator in the latter case (tunneled option) would need additional functionality to classify the same subscriber flows which may not be available on the deployed platforms.

The classifier use case is further extended when considering that many traffic classifications are made using transport layer (Layer 4) information. This is common in operator networks that often apply differential traffic treatment to different services that typically

operate using well defined TCP/UDP ports. In the MAP-T deployment case, these ports are available for classification matching using the same standard access edge node capabilities using IPv6 classifiers. In the case where tunneled forms of a solution are used, these higher layer ports are hidden from the network (base IP layer) and special functionality to correctly classify these service flows is required.

The ability to apply classifiers at the access edge node allows the operator to not only use standard IPv6 classifier functionality, but also use same mechanisms (RADIUS interface parameters/system, or DOCSIS configuration classifier parameters) for applying such classifiers. I.e. custom RADIUS interface extensions or custom DOCSIS classifier extensions to deal with the classifier semantics of an IP tunnel based transport are not required.

2.2. Device Configuration (DOCSIS)

Some access technologies, like DOCSIS, require a modem configuration file for network operation. These configuration files often contain access control and classification information that uses IPv4 and/or IPv6 network and transport layer information.

MAP-T allows use of standard IPv6 classifiers within these configuration files permitting the continued use of the well-known service architecture. Translation technologies which use tunneling may require the operator to update how services are managed as information needed to enforce policy is not longer viewable by the Cable Modem or upstream CMTS. The operator in this case may need to build new service capabilities higher up in the network after the network translator to apply the full range of policies for the subscriber base.

2.3. Service Flow management using Deep Packet Inspection

Several Service Providers today use Deep Packet Inspection devices located at the network edge (such as a BNG) in order to inspect the subscriber's traffic for different purposes: profiling the user's behavior, and classifying the traffic based on higher layer information and/or traffic signatures.

Deep packet inspection devices available today in the market and, more importantly, those already deployed in operator's network may not be able to analyze encapsulated traffic, like IPinIP, and to correlate the inner packet's contents to the outer packet's "subscriber" context - this limitation is consistent across multiple vendors. In order to overcome this limitation when using IP tunnel based transports, without resorting to costly network upgrades, dedicated DPI devices need to be applied at a point in the network

where the IP tunnel transport has been stripped and the payload is directly available for native processing. This not only changes the network architecture, but it increases the number of DPI's devices required: one for IPv6 traffic at the access edge, the other at a location where the IPv4 traffic is exposed (typically a separate location). In addition the operator would need to enforce policies at two architecturally separate places in the network. Furthermore, even with these changes enacted, there remains a critical problem of correlating traffic to a given subscriber: in encapsulation based solutions, the IPv4 address information in the payload is not sufficient to uniquely identify a subscriber given that an IPv4 address will not be unique. As such, additional mechanisms and changes to the accounting infrastructure need to be introduced which when combined with all the previous aspects makes this solution operationally complex.

With MAP-T operators can continue using the current architectural model with DPI devices installed at the access edge; the only requirement would be to have the same device able to recognize specific applications on the native IPv6 transport, which DPI devices based on application signatures are capable of doing. Thus with MAP-T it doesn't matter that an operator might provision the same IPv4 address across multiple subscribers. In addition with MAP-T the access edge would remain the single enforcement point for all user's policies for all traffic. This would allow the operators to continue using a consistent architecture and set of accounting tools for their network.

2.4. Service Flow Redirection Policies (Web-redirection)

Redirecting the user's traffic to web portal is a common practice in Service Provider networks. For example, it is common for operators to inform users about new services, service advisories and/or access to account changes using web-redirection techniques activated on http traffic. In current deployments web-redirection occurs at the Edge node level, where the subscriber's traffic first hits the IP network. The activation/de-activation of redirection policy on selected subscribers may be driven by the AAA/RADIUS through specific RADIUS attributes. In current deployments web-redirection occurs at the Edge node level, where the subscriber's traffic first hits the IP network. The activation/de-activation of redirection policy on selected subscribers may be driven by the AAA/RADIUS through specific RADIUS attributes.

If MAP-T is used the redirection of both IPv6 and IPv4 traffic can be kept at the Edge of the network with the same configuration currently used and by simply translating the Server's address in IPv6 with known mapping rules. In case of tunnel based solution the

redirection of IPv6 and IPv4 cannot occur in a single place, because the redirection of IPv4 traffic must be implemented at or after the v4/v6 gateway responsible for de-encapsulating the traffic. This approach not only would require deploying two separate infrastructures located in different places in order to achieve the redirection for both IPv6 and IPv4 traffic, but also it would not allow continuing using the AAA/RADIUS Server infrastructure in order to enforce the redirect policy at the subscriber's session.

2.5. Service Flow Caching

With the continuing growing of video traffic, especially considering the increase of http video traffic (YouTube like,) it is useful for the Service Providers to be able to cache the video stream at the Edge of the network in order to save bandwidth on upstream links. Using cache devices together with tunnel solutions would introduce similar challenges/issues as the ones described for DPI scenarios, in particular it would require applying caching functionality after the decapsulation point. Obviously this would not eliminate the benefits of the cache. Instead a MAP-T approach would allow caching the subscriber traffic at the edge of the network and gaining the bandwidth savings introduced by the caching. Crucially, any native IPv6 web-caches would be capable of processing IPv6 MAP-T traffic as fully native traffic.

In addition in some deployments today, Web Cache Control Protocol (WCCP) feature is used in order to redirect subscriber's traffic to the cache devices. When a subscriber requests a page from a web server (located in the Internet, in this case), the network node where the WCCP is active, sends the request to a Cache Engine. If the cache engine has a copy of the requested page in storage, the engine sends the user that page. Otherwise, the engine gets the requested page and the objects on that page from the web server, stores a copy of the page and its objects (caches them), and forwards the page and objects to the user. WCCP is another example of web redirect thus, the same considerations described in section Section 2.4 and the benefits introduced by MAP-T also apply here.

3. Technological Considerations

There are additional technological considerations which need to be analyzed by the operator when choosing which transition technology option they would like to deploy. This section describes some of those considerations.

3.1. Encapsulation and Translation Overhead

MAP-E adds an encapsulation tax of 40 bytes, while MAP-T adds a translation tax of 20 bytes (translating from a 20-byte IPv4 header to a 40-byte IPv6 header.) In the downstream direction (from network toward the CPE), with an average packet size of 1000-1100 bytes, the added encapsulation is under 4% in the case of MAP-E. In the case of MAP-T that encapsulation tax drops to about 2%.

In the upstream direction, with an average packet size of ~400 bytes, the effects of the encapsulation tax is more pronounced with an added 10% overhead for MAP-E and 5% additional overhead for MAP-T. As the upstream direction tends to be both (a) more heavily oversubscribed than is the downstream and (b) of lower performance, the greater the header tax the more it upsets the precariously balanced upstream/downstream network loading models.

3.2. Efficient Utilization of the Access Network

Point-to-Multipoint access networks are common across network operators - DOCSIS (1.0, 1.1, 2.0, 3.0), EPON, 10G-EPON, GPON, XGPON, XGPON2, etc. This network type has been incredibly successful, as attested to by all the variants of point-to-multipoint networks deployed, primarily because of their cost effectiveness.

There are a couple challenges that are introduced by adding a significant amount of encapsulation overhead. These challenges affect MAP-T and MAP-E similarly; the effects from MAP-E are simply more pronounced.

The first challenge is that, commonly, point-to-multipoint networks have limited support for jumbo frames. The second challenge is one that results in reduction in effective capacity on the wire, which yields higher cost.

3.2.1. Jumbo Frame Support in the Access

Some access technologies natively support fragmentation, and as a result, can support "jumbo frames" up to a point. A max size IPv4 packet that fits into the payload of a standard-compliant Ethernet frame is 1500 bytes. In the context of this discussion a "jumbo frame" is any Ethernet frame that has more than 1500 bytes in the Ethernet payload. IEEE Std. 802.3 now specifies a larger frame size of up to 2000 bytes, referred to as an envelope frame, where the envelope frame, quoting from IEEE Std.802.3-2012 "is intended to allow inclusion of additional prefixes and suffixes required by higher layer encapsulation protocols. The encapsulation protocols may use up to 482 octets."

In the network access space, particularly one filled with legacy access products which may be 10 years old (or perhaps older), it is not uncommon to find products that just only support a max 1500 byte Ethernet payload. Some may support up to 1532 byte payload (1550 byte Ethernet frame), some 1582 byte payload (1600 byte Ethernet frame), though there's certainly not a uniform supported frame size past the 1500 byte payload.

Since MTU discovery isn't typically used for IPv4 in operator networks and since MTU discovery for IPv6 is not implemented on the IPv4 host stack requesting the communication, there's no effective way to tell the host stack to reduce the size of its IPv4 frame to accommodate the MAP-T or MAP-E overhead with the MTU frame size limitation of the specific access products. There are tools like Maximum Segment Size rewrite that can be implemented to help address the issue for a TCP payload but UDP payload will continue to be impaired.

Thus MAP-T is preferred as there are more deployed access products that could support a 1534-byte or 1538-byte Ethernet frame than can support a 1554-byte or 1558-byte Ethernet frame, which mandates fewer access product replacements.

3.2.2. Operator Added Packet Overhead and Service Level Agreements

One of the traditional challenges with adding additional packet overhead to a customer frame is that it becomes more challenging to provide customer the last-mile bandwidth in their SLA. This is a very simple overprovisioning problem when the maximum size frame is used, as the overhead in that case is a fixed ~1.5% or ~3% for MAP-T and MAP-E respectively.

However in the case of variable packet sizes, the added overhead from either MAP-T or MAP-E can become very significant - from a worse case of ~31% (MAP-T) and ~63% (MAP-E) to the ~1.5% or ~3%. This means that to provide the customer what they purchased operators will either provision more than the customer SLA to account for the added overhead or abide by the "not guaranteed" bandwidth response.

With the average upstream packet sizes being smaller, the 5% (MAP-T) or 10% (MAP-E) added overhead for the average upstream packet size could find itself in an overprovisioned QoS profile.

Many customers, particularly business customers, are very savvy and have a strong belief that when a network operator offers them an SLA, it's not an SLA at a specific packet size. This can be a significant operational difficulty for network operators, one with a real operational cost.

4. Conclusions

The use cases described in this document have highlighted a clear need for a MAP-T solution based on Service Providers' operational requirements.

This document showed that a MAP-T approach is not a duplication of any other existing IPv4/IPv6 migration mechanisms based on IP tunneling, but actually has capabilities to solve Service Provider's problems.

5. Acknowledgements

The authors would like to thank Victor Kuarsingh for his valuable comments and inputs to this document.

6. IANA Considerations

This document does not require any action from IANA.

7. Security Considerations

This document has no additional security considerations beyond those already identified in section 11 of [RFC7599]

8. Informative References

- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, DOI 10.17487/RFC6333, August 2011, <<http://www.rfc-editor.org/info/rfc6333>>.
- [RFC7597] Troan, O., Ed., Dec, W., Li, X., Bao, C., Matsushima, S., Murakami, T., and T. Taylor, Ed., "Mapping of Address and Port with Encapsulation (MAP-E)", RFC 7597, DOI 10.17487/RFC7597, July 2015, <<http://www.rfc-editor.org/info/rfc7597>>.
- [RFC7599] Li, X., Bao, C., Dec, W., Ed., Troan, O., Matsushima, S., and T. Murakami, "Mapping of Address and Port using Translation (MAP-T)", RFC 7599, DOI 10.17487/RFC7599, July 2015, <<http://www.rfc-editor.org/info/rfc7599>>.

Authors' Addresses

Roberta Maglione (editor)
Cisco Systems
Via Torri Bianche 8
Vimercate 20871
Italy

Email: robmg1@cisco.com

Wojciech Dec
Cisco Systems
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands

Email: wdec@cisco.com

Ida Leung
Rogers Communications
8200 Dixie Road
Brampton, ON L6T 0C1
CANADA

Email: Ida.Leung@rci.rogers.com

Edwin Mallette
Bright House Networks
4145 S. Faulkenburg Road
Riverview, Florida 33578
USA

Email: edwin.mallette@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 23, 2013

B. Sarikaya
Huawei USA
February 19, 2013

Multicast Support for MAP-E
draft-sarikaya-software-4rdmulticast-06.txt

Abstract

This memo specifies MAP-E (together with MAP-T and 4rd)'s multicast component so that IPv4 hosts can receive multicast data from IPv4 servers over an IPv6 network. In the encapsulation solution for encapsulation variant of Mapping of Address and Port (MAP), MAP-E, IGMP Proxy at the MAP-E Customer Edge router uses IPv4-in-IPv6 tunnel established by MAP-E to exchange IGMP messages to establish multicast state at MAP-E Border Relay so that MAP-E Border Relay can tunnel IPv4 multicast data to IPv4 hosts connected to MAP-E Customer Edge device. In the Translation Multicast solution for the translation variant of MAP, MAP-T and 4rd, IGMP messages are translated into MLD messages at the CE router which is IGMP/MLD Proxy and sent to the network in IPv6. MAP-T/4rd Border Relay does the reverse translation and joins IPv4 multicast group for MAP-T/4rd hosts. Border Relay as multicast router receives IPv4 multicast data and translates the packet into IPv6 multicast data and sends downstream on the multicast tree. Member CEs receive multicast data, translate it back to IPv4 and transmit to the hosts.

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 August 23, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Terminology	4
3. Requirements	4
4. Architecture	5
4.1. Encapsulation Multicast Architecture	5
4.2. MAP-T and 4rd Translation Architecture	6
5. Encapsulation Multicast Operation	7
5.1. Encapsulation Interface Considerations	9
5.2. Avalanche Problem Considerations	10
6. MAP-T and 4rd Translation Multicast Operation	10
6.1. Address Translation	11
6.2. Protocol Translation	12
6.3. Supporting IPv6 Multicast in MAP-T and 4rd Translation Multicast	13
6.4. Learning Multicast Prefixes for IPv4-embedded IPv6 Multicast Addresses	14
7. Security Considerations	15
8. IANA Considerations	15
9. Acknowledgements	15
10. References	15
10.1. Normative References	15
10.2. Informative references	17
Appendix A. Group Membership Message Translation Details	18
Author's Address	20

1. Introduction

With IPv4 address depletion on the horizon, many techniques are being standardized for IPv6 migration including Mapping of Address and Port (MAP) - Encapsulation, - Translation and 4rd [I-D.ietf-software-map], [I-D.ietf-software-map-t], [I-D.ietf-software-4rd]. MAP/4rd enables IPv4 hosts to communicate with external hosts using IPv6 only ISP network. MAP/4rd Customer Edge (CE) device's LAN side is dual stack and WAN side is IPv6 only. CE tunnels/translates IPv4 packets received from the LAN side to 4rd Border Relays (BR). BRs have anycast IPv6 addresses and receive encapsulated/translated packets from CEs over a virtual interface. MAP/4rd operation is stateless. Packets are received/ sent independent of each other and no state needs to be maintained except for NAT44 operation on IPv4 packets received from the user.

It should be noted that there is no depletion problem for IPv4 address space allocated for any source multicast and source specific multicast [RFC3171]. This document is not motivated by the depletion of IPv4 multicast addresses.

MAP-E, MAP-T and 4rd are unicast only. They do not support multicast. In this document we specify how multicast from home IPv4 users can be supported in MAP-E (as well as MAP-T and 4rd).

In case of MAP-E we integrate the multicast solution into the MAP-E tunnel resulting in a multicast tunneling protocol. Multicast tunneling protocol has the advantage of not requiring multicast enabled IPv6 network between CE routers and MAP-E BRs.

When MAP-E CE router receives an IGMP join message to an Any-Source Multicast (ASM) [RFC1112] or Source-Specific Multicast (SSM) group [RFC4607], it sends an aggregated IGMP membership report message in the IPv4-in-IPv6 tunnel to the border relay. MAP-E BR joins the source in the multicast infrastructure and sends multicast data downstream to all member CEs in the IPv4-in-IPv6 tunnel. When a CE has no membership state, i.e. after all member hosts leave the group(s), its state with the BR expires and the CE can send the next join message in anycast. IPv4 multicast data received at the BR is tunneled to the member CE in IPv6 and CE decapsulates the packet and sends IPv4 multicast data packet to the member hosts.

In case IPv6 network is multicast enabled, MAP-T/4rd can provide multicast service to the hosts using MAP-T/4rd Multicast Translation based solution. A Multicast Translator can be used that receives IPv4 multicast group management messages in IGMP and generates corresponding IPv6 group management messages in MLD and sends them to IPv6 network towards MAP-T/4rd Border Relay. We use

[I-D.ietf-software-map-t] or [I-D.ietf-software-4rd] for sending IPv4 multicast data in IPv6 to the CE routers. At MAP-T/4rd CE router another translator is needed to translate IPv6 multicast data into IPv4 multicast data.

It should be noted that if IPv6 network is multicast enabled the translation multicast solution presented in Section 6 can also be used for MAP-E.

In this document we address MAP-E (and MAP-T/4rd) multicast problem and propose two architectures: Multicast Tunneling and Multicast Translation based solutions. Section 2 is on terminology, Section 3 is on requirements, Section 4 is on architecture, Section 5 is on multicast tunneling protocol Section 6 is on multicast translation protocol, and Section 7 states security considerations.

2. Terminology

This document uses the terminology defined in [I-D.ietf-software-map], [I-D.ietf-software-map-t], [I-D.ietf-software-4rd], [RFC3810] and [RFC3376].

3. Requirements

This section states requirements on MAP-E, MAP-T and 4rd multicast support protocol.

IPv4 hosts connected to MAP-E, MAP-T and 4rd CE router MUST be able to join multicast groups in IPv4 and receive multicast data.

Any source multicast (ASM) SHOULD be supported and source specific multicast (SSM) MUST be supported.

In case of encapsulation solution, MAP-E CE MUST support IGMP Proxy as defined in [RFC4605]. MAP-E BR MUST support IGMP querier downstream and MAP-E BR may support PIM protocol or IGMP router upstream.

In case of translation solution, MAP-T and 4rd CE MUST support IGMP to MLD translation. MAP-T and 4rd CE MUST be MLD Proxy as defined in [RFC4605]. MAP-T and 4rd BR MUST support MLD Querier. MAP-T and 4rd BR MUST support join/leave operations in IPv4 multicast upstream.

4. Architecture

In MAP-E, MAP-T and 4rd, there are hosts (possibly IPv4/ IPv6 dual stack) served by MAP-E, MAP-T and 4rd Customer Edge device. CE is dual stack facing the hosts and IPv6 only facing the network or WAN side. MAP-E, MAP-T and 4rd CE may be local IPv4 Network Address and Port Translation (NAPT) box [RFC3022] by assigning private IPv4 addresses to the hosts. MAP-E, MAP-T and 4rd CEs in the same domain may use shared public IPv4 addresses on their WAN side and if they do they should avoid ports outside of the allocated port set for NAPT operation. At the boundary of the network there is MAP-E, MAP-T and 4rd Border Relay. BR receives IPv4 packets tunneled in IPv6 from CE and decapsulates them and sends them out to IPv4 network.

Unicast MAP-E, MAP-T and 4rd are stateless except for the local NAPT at the CE. Each IPv4 packet sent by CE treated separately and different packets from the same CE may go to different BRs or CEs. CE encapsulates IPv4 packet in IPv6 with destination address set to BR address (usually anycast IPv6 address). BR receives the encapsulated packet and decapsulates and send it to IPv4 network. CEs are configured with Rule IPv4 Prefixes, Rule IPv6 Prefixes and with an BR IPv6 anycast address. BR receives IPv4 packets addressed to this ISP and from the destination address it extracts the destination host's IPv4 address and uses this address as destination address and encapsulates the IPv4 packet in IPv6 and sends it to IPv6-only network.

4.1. Encapsulation Multicast Architecture

Encapsulation variant of MAP called MAP-E network lends itself easily to the Multicast Tunneling architecture. Dual stack hosts are connected to the Customer Edge router and it is multicast enabled. It is assumed that IPv6 only network is the unicast only network and that IPv6 multicast is not enabled or IPv6 multicast is partially enabled. At the boundary of the network MAP-E Border Relay is connected to the native multicast backbone infrastructure.

We place IGMP Proxy at the CE router. CE router serves all the connected hosts. For multicast traffic, CE Router uses MAP-E tunneling interface with MAP-E BR to send/receive IGMP messages using IPv4-in-IPv6 tunnel [RFC2473].

MAP-E BR is IGMP Router towards the CEs and it could be IGMP Router or PIM router upstream. A given relay and all CEs connected to it can be considered to be on a separate logical link. On this link, gateways and relay communicate using IPv4-in-IPv6 tunneling to transmit and receive multicast control messages for membership management and multicast data from the relay to the gateways.

All the elements of MAP-E multicast support system with tunneling are shown in Figure 1.

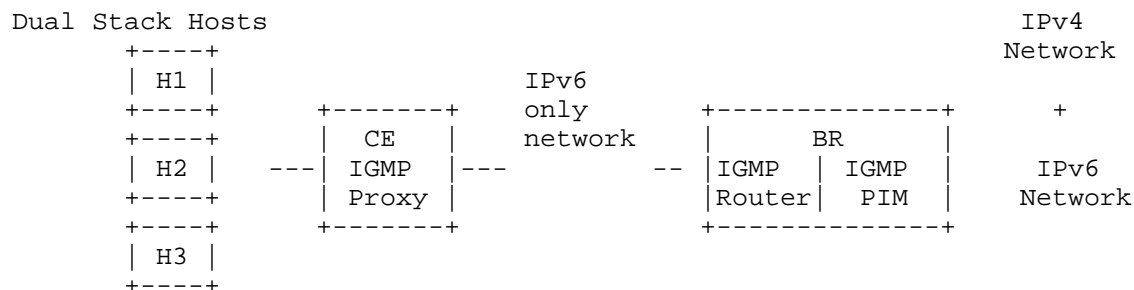


Figure 1: Architecture of MAP-E Multicast Tunneling

4.2. MAP-T and 4rd Translation Architecture

In case IPv6 only network is multicast enabled, translation multicast architecture can be used. CE implements IGMP Proxy function [RFC4605] towards the LAN side and MLD Proxy on its WAN interface. IPv4 hosts send their join requests (IGMP Membership Report messages) to CE. CE as a MLD proxy sends aggregated MLD Report messages upstream towards BR. CE replies MLD membership query messages with MLD membership report messages based on IGMP membership state in the IGMP/MLD proxy.

BR is MLD querier on its WAN side. On its interface to IPv4 network BR may either have IGMP client or PIM. PIM being able to support both IPv4 and IPv6 multicast should be preferred. BR receives MLD join requests, extracts IPv4 multicast group address and then joins the group upstream, possibly by issuing a PIM join message.

IPv4 multicast data received by the BR as a leaf node in IPv4 multicast distribution tree is translated into IPv6 multicast data by the translator using [I-D.ietf-softwire-map-t], [I-D.ietf-softwire-4rd] and then sent downstream to the IPv6 part of the multicast tree to all downstream routers that are members. IPv6 data packet eventually gets to the CE. At the CE, a reverse [I-D.ietf-softwire-map-t], [I-D.ietf-softwire-4rd] operation takes place by the translator and then IPv4 multicast data packet is sent to the member hosts on the LAN interface. [I-D.ietf-softwire-map-t], [I-D.ietf-softwire-4rd] are modified to handle multicast addresses.

In order to support SSM, IGMPv3 MUST be supported by the host, CE and

BR. For ASM, BR MUST be the Rendezvous Point (RP).

MAP-T and 4rd Translation Multicast solution uses the multicast 46 translator in not one but two places in the architecture: at the CE router and at the Border Relay. IPv4 multicast data received at 4rd BR goes through a [I-D.ietf-software-4rd] header-mapping into IPv6 multicast data at the BR and another [I-D.ietf-software-4rd] header-mapping back to IPv4 multicast data at the CE router. Encapsulation variant of [I-D.ietf-software-4rd] is not used. In case of MAP-T, IPv4 data packet is translated using v4 to v6 header translation using multicast addresses instead of the mapping algorithm used in [I-D.ietf-software-map-t].

All the elements of MAP-T and 4rd translation-based multicast support system are shown in Figure 2.

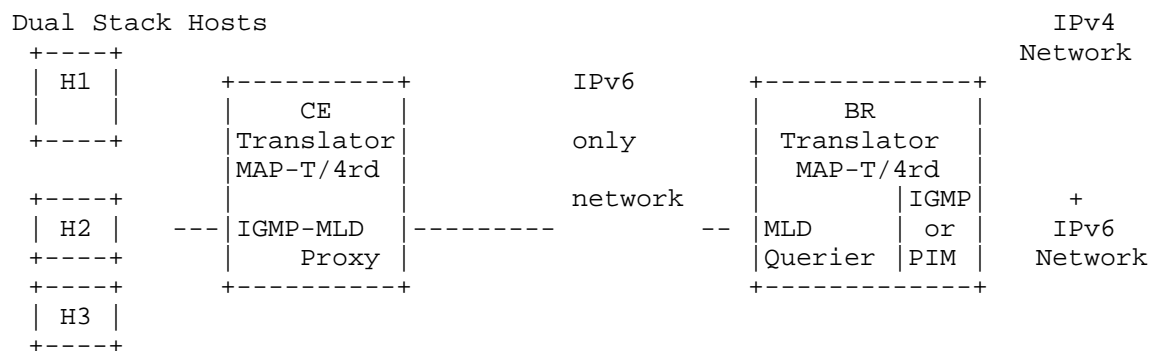


Figure 2: Architecture of MAP-T and 4rd Translation Multicast

5. Encapsulation Multicast Operation

When a host (H1, H2 or H3 in Figure 1) wants to join an IPv4 multicast group G or (S,G), it sends an IGMP report (IGMPv3 report for a source-specific group) to CE router.

CE encapsulates IGMP report messages in IPv6 and sends it over the tunnel to BR in anycast. CE router uses BR's anycast address this CE router is configured with. After CE receives unicast address of BR, it sends all subsequent IGMP messages for G or (S,G) in unicast.

BR (topologically closest to this CE router) receives the message, decapsulates it and then lets IGMP router to process it. On the upstream, an IGMP Join message is sent to subscribe group G or (S,G) or a PIMv4 Join message is sent if PIM is supported. BR establishes

membership state for group G or (S,G). BR sends all related IGMP messages to this CE in unicast using IPv4-in-IPv6 tunneling.

CE now has BR's unicast address which it uses to send all IGMP packets for group G for any source multicast or (S,G) for source specific multicast. If CE receives multiple join messages for the same group G, CE sends an aggregated join message to BR.

If CE receives another join message for a different group G', (S',G') CE encapsulates it and sends it in anycast to the BR. This enables the use of multiple BRs that may be deployed as anchor points and makes downstream multicast data delivery more efficient.

A CE is required to assist in IGMP signaling and data forwarding between the hosts that it serves and the corresponding BRs that are handling the multicast group G or (S,G). CE must have IGMP Proxy for each upstream tunnel interface that has been established with the BR. The CE decides on the mapping of downstream links to a proxy instance connected to an upstream link to a BR based on the unicast source IPv6 address in the packets received from BR. Because of this BRs MUST use the unicast source IPv6 address in packets sent to CEs. Encapsulation at the CE is according to [RFC2473] with an IPv4 payload carrying IGMP messages.

On the reception of IGMP reports from the hosts, the CE must identify the corresponding proxy instance from the incoming interface and perform regular IGMP proxy operations of inserting, updating or removing multicast forwarding state on the incoming interface and will merge state updates into the IGMP proxy membership database. It will then send an aggregated Report via the upstream tunnel to the BR when the membership database changes.

On the reception of IGMP queries, the CE proxy instance will answer the Queries on behalf of all active downstream receivers maintained in its membership database. Queries sent by the BR do not force the CE to trigger corresponding messages immediately towards hosts.

BR acts as the default multicast querier for the corresponding CE. It implements the function of the designated multicast router or a further IGMP proxy. After BR receives IGMP Join message it adds the tunnel to the CE in its outgoing interface list for the group (G) or the source, group (S,G) that the host wants to join. BR establishes group-/source-specific multicast forwarding states at its corresponding downstream tunnel interfaces. Afterwards, BR maintains/removes these group-/source-specific multicast forwarding states. BR treats its tunnel interfaces as multicast-enabled downstream links, serving zero to many listening nodes. BR will send a join message upstream towards the source of the multicast group to

build a multicast tree in the native multicast infrastructure and becomes a leaf node in the multicast tree.

BR will send any group management messages (IGMP Report or Query messages) downstream to specific CEs on the tunnel interface by encapsulating these IGMP messages in IPv6 using [RFC2473].

As for multicast data, the data packets from the source received at the BR will be replicated to all interfaces in its outgoing interface list as well as the tunnel outgoing interface for all member CEs. BR sends multicast data in IPv4-in-IPv6 tunnel to the CE with the data packet encapsulated. Encapsulation is according to [RFC2473] with an IPv4 payload.

CE receives Multicast Data message over the tunnel interface associated with the tunnel to BR. After decapsulation, multicast traffic arriving at the CE on an upstream interface will be forwarded according to the group-specific or source-specific forwarding states as acquired for each downstream interface within the IGMP proxy instance.

5.1. Encapsulation Interface Considerations

Legacy IPv4 in IPv6 tunneling is performed as in [RFC2473]. Packets upstream from CE carry only IGMP signaling messages and they are not expected to be subject to fragmentation. However packets downstream, i.e. multicast data to CE may be subject to fragmentation.

Source and destination addresses of IGMP messages in IPv4-in-IPv6 software from CE is as follows:

Source address of IPv6 header is CE IPv6 address, e.g. 2001:db8:0:1::1, destination address is BR anycast address, possibly shared of the MAP domain.

Source address of IGMP messages is CE's IPv4 interface address, e.g. 192.0.0.2, destination address is the all-systems multicast address of 224.0.0.1 for IGMP Query, all IGMPv3-capable multicast routers of 224.0.0.22 for IGMPv3 Report, the multicast group specified in the Group Address field of the Report for IGMPv1 or IGMPv2 Report.

Source and destination addresses of IGMP messages in IPv4-in-IPv6 software from BR is as follows:

Source address of IPv6 header is BR's unicast IPv6 address, e.g. 2001:db8:0:2::1, destination address is CE IPv6 address, e.g. 2001:db8:0:1::1.

Source address of IGMP messages is CE's IPv4 interface address, e.g. 192.0.2.1, destination address is the all-systems multicast address of 224.0.0.1 for IGMP Query, all IGMPv3-capable multicast routers of 224.0.0.22 for IGMPv3 Report, the multicast group specified in the Group Address field of the Report for IGMPv1 or IGMPv2 Report.

Source and destination addresses of multicast data messages in IPv4-in-IPv6 software is as follows:

Source address of IPv6 header is BR IPv6 unicast address, e.g. 2001:db8:0:2::1, destination address is CE IPv6 address, e.g. 2001:db8:0:1::1.

Source address of IPv4 multicast data is unicast IPv4 address of the multicast source, e.g. the content provider, destination address is IPv4 multicast group address.

BR decapsulates datagrams carrying IGMP messages from CE's and then IGMP/PIM router processing takes over. Network Address Translation (NAT) is not applied on IGMP messages.

5.2. Avalanche Problem Considerations

In Section 5 BR replicates the data packets from the source received to all outgoing interfaces for all member CEs. This replication (often called avalanche problem) can be very costly if there are very large number of downstream member CEs such as in IPTV application. Note that the avalanche problem is faced by all multicast solutions that use tunneling to bypass non-multicast enabled access network.

In multicast MAP-E, one approach that can be used is to deploy MAP-E BRs close to the user. BRs colocated at the access network gateway such as at the Border Network Gateway (BNG) could reduce the packet duplication bottleneck considerably.

In multicast MAP-E, another approach is to exploit multiple BRs that can be deployed in the network. MAP-E CE can use BR anycast address when sending an encapsulated upstream IGMP join request and then use the unicast source address of this BR in subsequent IGMP messages.

6. MAP-T and 4rd Translation Multicast Operation

In this section we specify how the host can subscribe and receive IPv4 multicast data from IPv4 content providers based on the architecture defined in Figure 2 in two parts: address translation and protocol translation. Translation details are given in Appendix A.

6.1. Address Translation

IPv4-only host, H1 will join IPv4 multicast group by sending IGMP Membership Report message upstream towards the IGMP Proxy in Figure 2. MLD Proxy first creates a synthesized IPv6 address of IPv4 multicast group address using IPv4-embedded IPv6 multicast address format [I-D.ietf-mboned-64-multicast-address-format]. ASM_MPREFIX64 for any source multicast groups and SSM_MPREFIX64 for source specific multicast groups are used. Both are /96 prefixes.

SSM_MPREFIX64 is set to ff3x:0:8000::/96, with 'x' set to any valid scope. ASM_MPREFIX64 values are formed as shown in Figure 3. M bit MUST BE set to 1. "flgs" and "scop" fields are defined in [RFC3956]. The usage of the "rsv" bits is the same as defined in [RFC3306]. "sub-group-id" field MUST follow the recommendations specified in [RFC3306] if unicast-based prefix is used or the recommendations specified in [RFC3956] if embedded-RP is used. The default value is all zeros.

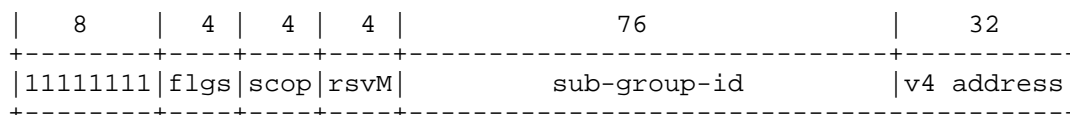


Figure 3: ASM_MPREFIX64 Formation

Each translator in the upstream BR is assigned a unique ASM_MPREFIX64 prefix. CE (MLD Proxy in CE) can learn this value by means out of scope with this document. With this, CE can easily create an IPv6 multicast address from the IPv4 group address a.b.c.d that the host wants to join.

Source-Specific Multicast (SSM) can also be supported similar to the Any Source Multicast (ASM) described above. In case of SSM, IPv4 multicast addresses use 232.0.0.0/8 prefix. IPv6 SSM_MPREFIX64 is set to FF3x:0:8000::/96.

Since SSM translation requires a unique address for each IPv4 multicast source, an IPv6 unicast prefix must be configured to the translator in the upstream BR to represent IPv4 sources. This prefix is prepended to IPv4 source addresses in translated packets.

The join message from the host for the group ASM_MPREFIX64:a.b.c.d or SSM_MPREFIX64:a.b.c.d or an aggregate join message will be received by MLD querier at the BR. BR as multicast anchor checks the group address and recognizes ASM_MPREFIX64 or SSM_MPREFIX64 prefix. It next checks the last 32 bits is an IPv4 multicast address in range 224/8 - 239/8. If all checks succeed, IGMPv4 Client joins a.b.c.d

using IGMP on its IPv4 interface.

Joining IPv4 groups can also be done using PIM since PIM supports both IPv4 and IPv6. The advantage of using PIM is that there is no need to enable IGMP support in neighboring IPv4 routers. The advantage of using IGMP is that IGMP is a simpler protocol and it is supported by a wider range of routers. The use of PIM or IGMP is left as an implementation choice.

6.2. Protocol Translation

The hosts will send their subscription requests for IPv4 multicast groups upstream to the default router, i.e. Customer Edge device. After subscribing the group, the host can receive multicast data from the CE. The host implements IGMP protocol's host part.

Customer Edge device is IGMP Proxy facing the LAN interface. After receiving the first IGMP Report message requesting subscription to an IPv4 multicast group, a.b.c.d, MLD Proxy in the CE's WAN interface synthesizes an IPv6 multicast group address corresponding to a.b.c.d and sends an MLD Report message upstream to join the group.

When CE is a NAT or NAT box assigning private IPv4 addresses to the hosts, IP Multicast requirements for a Network Address Translator (NAT) and a Network Address Port Translator (NAPT) stated in [RFC5135] apply to IGMP messages and IPv4 multicast data packets.

When MAP-T or 4rd BR receives IPv4 multicast data for an IPv4 group a.b.c.d it [I-D.ietf-softwire-4rd] translates/encapsulates IPv4 packet into IPv6 multicast packet and sends it to IPv6 synthesized address corresponding to a.b.c.d using ASM_MPREFIX64 or SSM_MPREFIX64. The header mapping described in [I-D.ietf-softwire-4rd] Section 4.2 (using Table 1) is used except for mapping the source and destination addresses. In this document we use the multicast address translation described in Section 6.1 and propose it as a complementary enhancement to the translation algorithm in [I-D.ietf-softwire-4rd].

The IP/ICMP translation translates IPv4 packets into IPv6 using minimum MTU size of 1280 bytes (Section 4.3 in [I-D.ietf-softwire-4rd]) but this can be changed for multicast. Path MTU discovery for multicast is possible in IPv6 so 4rd BR can perform path MTU discovery for each ASM group and use these values instead of 1280. For SSM, a different MTU value MUST be kept for each SSM channel. Because of this 8 bytes added by IPv6 fragment header in each data packet can be tolerated.

Since multicast address translation does not preserve checksum

neutrality, [I-D.ietf-software-4rd] translator/encapsulator at 4rd BR must however modify the UDP checksum to replace the IPv4 addresses with the IPv6 source and destination addresses in the pseudo-header which consists of source address, destination address, protocol and UDP length fields before calculating the new checksum.

IPv6 multicast data must be translated back to IPv4 at the 4rd CE (e.g. using Table 2 in Section 4.3 of [I-D.ietf-software-4rd]). Such a task is much simpler than the translation at 4rd BR because IPv6 header is much simpler than IPv4 header and IPv4 link on the LAN side of 4rd CE is a local link. The packet is sent on the local link to IPv4 group address a.b.c.d for IPv6 group address of ASM_MPREFIX64: a.b.c.d or SSM_MPREFIX64:a.b.c.d.

In case an IPv4 multicast source sends multicast data with the don't fragment (DF) flag set to 1, [I-D.ietf-software-4rd] header mapping sets the D bit in IPv6 fragment header before sending the packet downstream as in Fig. 3 in Section 4.3 of [I-D.ietf-software-4rd]. This feature of [I-D.ietf-software-4rd] preserves the semantics of DF flag at the BR and CE.

Because MAP-T/4rd is stateless, Multicast MAP-T/4rd should stay faithful to this as much as possible. Border Relay acts as the default multicast querier for all CEs that have established multicast communication with it. In order to keep a consistent multicast state between a CE and BR, CE MUST use the same IPv6 multicast prefixes (ASM_MPREFIX64/SSM_REFIX64) until the state becomes empty. After that point, the CE may obtain different values for these prefixes, effectively changing to a different 4rd BR.

6.3. Supporting IPv6 Multicast in MAP-T and 4rd Translation Multicast

IPv6 multicast can be supported natively since IPv6-only network is assumed to be multicast enabled. MAP-T or 4rd Customer Edge device has MLD Proxy function. Proxy operation for MLD [RFC3810] is described in [RFC4605].

CE receives MLD join requests from the hosts and then sends aggregated MLD Report messages upstream towards BR. No address or protocol translation is needed at the CE or at the BR. IPv6 Hosts in MAP-T or 4rd domain use any source multicast block FF0X [RFC4291] or source specific multicast block FF3X::8000:0-FF3X::FFFF:FFFF for dynamic allocation by a host [RFC4607], [RFC3307].

MAP-T or 4rd Border Relay is MLD querier. It serves all CEs downstream. After receiving an MLD join message, BR sends PIM join message upstream to join IPv6 multicast group. Multicast membership database is maintained based on the aggregated Reports received from

downstream interfaces in the multicast tree.

MAP-T or 4rd Border Relay is a Rendezvous Point (RP) for ASM groups. For SSM, BR MUST support MLDv2.

IPv6 multicast data received from the Single Source Multicast or Any Source Multicast sources are replicated according to the multicast membership database and the data packets are sent downstream on the multicast tree and eventually received by the CEs that have one of more members of this multicast group.

MLD Proxy in the CE receives multicast data then forwards the packet downstream. Each member host receives IPv6 multicast data packet from its Layer 2 interface.

6.4. Learning Multicast Prefixes for IPv4-embedded IPv6 Multicast Addresses

CE can be pre-configured with Multicast Prefix64 of ASM_MPREFIX64 and SSM_MPREFIX64 that are supported in their network. However automating this process is also desired.

A new router advertisement option, a Multicast ASM Translation Prefix option, can be defined for this purpose. The option contains IPv6 ASM multicast translation prefix, ASM_MPREFIX64. A new router advertisement option, a Multicast SSM Translation Prefix option, can be defined for this purpose. The option contains IPv6 SSM multicast prefix translation prefix SSM_MPREFIX64.

After the host gets the multicast prefixes, when an application in the host wishes to join an IPv4 multicast group the host MUST use ASM_MPREFIX64 or SSM_MPREFIX64 and then obtain the synthesized IPv6 group address before sending MLD join message.

Source-specific multicast (SSM) group membership message payloads in IGMPv3 and MLDv2 contain address literals and their translation requires another multicast translation prefix option. IPv4 source addresses in IGMPv3 Membership Report message are unicast addresses of IPv4 sources and they have to be translated into unicast IPv6 source addresses in MLDv2 Membership Report message. A new router advertisement option, a Multicast Translation Unicast Prefix option can be defined for this purpose. The option contains IPv6 unicast Network-Specific Prefix U_PREFIX64. The host can be configured by its default router using router advertisements containing the prefixes [I-D.sarikaya-softwire-6man-raoptions]. 64:ff9b::/96 is the global value called well-known prefix that is assigned to U_PREFIX64 [RFC6052]. Organization specific values called Network-Specific Prefixes can also be used. Since multicast is potentially inter-

domain, the use of well-known prefix for U_PREFIX64 is recommended.

Note that U_PREFIX64 is also used in multicast data packet address translation. Source-specific multicast source address in multicast data packets coming from SSM sources MUST be translated using U_PREFIX64.

7. Security Considerations

4rd Encapsulation Multicast control and data message security can be provided by the security architecture, mechanisms, and services described in [RFC4301], [RFC4302] and [RFC4303]. 4rd Translation Multicast control and data message security are as described in [RFC4607] for source specific multicast.

8. IANA Considerations

TBD.

9. Acknowledgements

TBD.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.
- [RFC1112] Deering, S., "Host extensions for IP multicasting", STD 5, RFC 1112, August 1989.
- [RFC2113] Katz, D., "IP Router Alert Option", RFC 2113, February 1997.
- [RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, October 1999.
- [RFC3171] Albanna, Z., Almeroth, K., Meyer, D., and M. Schipper, "IANA Guidelines for IPv4 Multicast Address Assignments",

RFC 3171, August 2001.

- [RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [RFC4607] Holbrook, H. and B. Cain, "Source-Specific Multicast for IP", RFC 4607, August 2006.
- [RFC3307] Haberman, B., "Allocation Guidelines for IPv6 Multicast Addresses", RFC 3307, August 2002.
- [RFC2491] Armitage, G., Schulter, P., Jork, M., and G. Harter, "IPv6 over Non-Broadcast Multiple Access (NBMA) networks", RFC 2491, January 1999.
- [RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, December 1998.
- [RFC2765] Nordmark, E., "Stateless IP/ICMP Translation Algorithm (SIIT)", RFC 2765, February 2000.
- [RFC3022] Srisuresh, P. and K. Egevang, "Traditional IP Network Address Translator (Traditional NAT)", RFC 3022, January 2001.
- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.
- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, December 2005.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.
- [RFC5135] Wing, D. and T. Eckert, "IP Multicast Requirements for a Network Address Translator (NAT) and a Network Address Port Translator (NAPT)", BCP 135, RFC 5135, February 2008.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation

Algorithm", RFC 6145, April 2011.

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

[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-softwire-map-t]
Li, X., Bao, C., Dec, W., Troan, O., Matsushima, S., and T. Murakami, "Mapping of Address and Port using Translation (MAP-T)", draft-ietf-softwire-map-t-01 (work in progress), February 2013.

[I-D.ietf-softwire-4rd]
Jiang, S., Despres, R., Penno, R., Lee, Y., Chen, G., and M. Chen, "IPv4 Residual Deployment via IPv6 - a Stateless Solution (4rd)", draft-ietf-softwire-4rd-04 (work in progress), October 2012.

[I-D.ietf-mboned-64-multicast-address-format]
Boucadair, M., Qin, J., Lee, Y., Venaas, S., Li, X., and M. Xu, "IPv6 Multicast Address With Embedded IPv4 Multicast Address", draft-ietf-mboned-64-multicast-address-format-04 (work in progress), August 2012.

10.2. Informative references

[RFC3306] Haberman, B. and D. Thaler, "Unicast-Prefix-based IPv6 Multicast Addresses", RFC 3306, August 2002.

[RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[I-D.sarikaya-softwire-6man-raoptions]
Sarikaya, B., "IPv6 RA Options for Translation Multicast Prefixes", draft-sarikaya-softwire-6man-raoptions-00 (work in progress), August 2012.

[I-D.perreault-mboned-igmp-ml-d-translation]

Perreault, S. and T. Tsou, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Translation ("IGMP/MLD Translation")", draft-perreault-mboned-igmp-ml-d-translation-01 (work in progress), April 2012.

Appendix A. Group Membership Message Translation Details

IGMP Report messages (IGMP type number 0x12 and 0x16, in IGMPv1 and IGMPv2 and 0x22 in IGMPv3) are translated into MLD Report messages (MLDv1 ICMPv6 type number 0x83 and MLDv2 type number 0x8f). IGMP Query message (IGMP type number 0x11) is translated into MLD Query message (ICMPv6 type number 0x82)

[I-D.perreault-mboned-igmp-ml-d-translation].

Destination address in ASM, i.e. IGMPv1, IGMPv2 and MLDv1, is the multicast group address so the destination address in IGMP message is translated into the destination address in MLD message using [I-D.ietf-mboned-64-multicast-address-format].

Destination address in SSM, i.e. IGMPv3 and MLDv2 is translated as follows: it could be all nodes on link, which has the value of 224.0.0.1 (IGMPv3) and ff02::1 (MLDv2), all routers on link, which has the value of 224.0.0.2 (IGMPv3) and ff02::2 (MLDv2), all IGMP/MLD-capable routers on link, which has the value of 224.0.0.22 (IGMPv3) and ff02::16 (MLDv2).

Source address of MLD message that CE sends is set to link-local IPv6 address of CE's WAN side interface. Source address of MLD message that BR sends is set to link-local IPv6 address of BR's downstream interface.

Multicast Address or Group Address field in IGMP message payloads is translated using [I-D.ietf-mboned-64-multicast-address-format] as described above into the corresponding field in MLD message.

Source Address in IGMPv3 message payloads is translated using U_PREFIX64, the IPv6 unicast prefix to be used by SSM source. [RFC6052] defines in Section 2.3 the address translation algorithm of embedding an IPv4 source address and obtaining an IPv6 source address using a network specific prefix like U_PREFIX64. At the BR on its upstream interface or at the CE on its LAN interface, IPv4 addresses are extracted from the IPv4-embedded IPv6 addresses.

Maximum Response Time (MRT) field in IGMPv2 and IGMPv3 queries are translated into Maximum Response Delay (MRD) in MLDv1 and MLDv2

queries, respectively. In the corresponding MLD message, MRD is set to 100 times the value of MRT. At the BR on its upstream interface or at the CE on its LAN interface, MRT value is obtained by dividing MRD into 100 and rounding it to the nearest integer.

IGMP messages are sent with a Router Alert IPv4 option [RFC2113]. The translated MLD message are sent with a Router Alert option in a Hop-By-Hop IPv6 extension header [RFC2711]. In both cases, 2-octet value is set to 0.

Author's Address

Behcet Sarikaya
Huawei USA
5340 Legacy Dr. Building 175
Plano, TX 75024

Phone:
Email: sarikaya@ieee.org

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 23, 2012

B. Sarikaya
Huawei USA
June 21, 2012

Multicast Support for Dual Stack Lite
draft-sarikaya-softwire-dslitemulticast-01.txt

Abstract

This memo specifies modifications required to Dual-Stack Lite (DS-Lite) so that both IPv4 hosts can receive multicast data from IPv4 servers.

The DS-Lite solution is based on DS-Lite Basic Bridging BroadBand element (B4) proxying Internet Group Management Protocol (IGMP) and then tunneling IGMP messages over IPv4-in-IPv6 softwire to DS-Lite Address Family Transition Router element (AFTR). IPv4 multicast data received at AFTR is tunneled over IPv4-in-IPv6 softwire to B4 and then delivered to the hosts. This solution integrates well with DS-Lite unicast solution by using IPv4-in-IPv6 softwire and works with unicast IPv6 network connecting B4 with AFTR.

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 December 23, 2012.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Terminology	3
3. Requirements	3
4. Architecture	4
5. DS-Lite Multicast Operation	4
5.1. Tunnel Interface Considerations	6
6. Multicast Support for Host-Based Architecture	7
7. Avalanche Problem	7
8. IANA Considerations	7
9. Acknowledgements	8
10. References	8
10.1. Normative References	8
10.2. Informative references	8
Author's Address	10

1. Introduction

With IPv4 address depletion on the horizon, many techniques are being standardized for IPv6 migration including DS-Lite [RFC6333] and 6rd [RFC5969]. DS-Lite enables IPv4 hosts to communicate with external hosts using IPv6 only network and moves the traditional NAT to the network. B4 element's LAN side is dual stack and WAN side is IPv6 only. B4 tunnels IPv4 packets received from the LAN side to AFTR element after encapsulating IPv4 packet in an IPv6 packet. AFTR decapsulates the packet, does a NAT operation and then sends the packet out to IPv4 public internet.

DS-Lite as defined in [RFC6333] is unicast only, it does not support multicast. In this document we specify multicast extensions to DS-Lite in order to provide IP multicast communication to home IPv4 users in DS-Lite.

2. Terminology

This document uses the terminology defined in [RFC6333] and [RFC3376].

3. Requirements

This section states requirements on DS-Lite multicast support protocol.

DS-Lite multicast solution MUST integrate with DS-Lite unicast solution, it MUST not introduce additional mechanisms to the existing B4 to AFTR communication.

DS-Lite multicast solution MUST not require additional capabilities in IPv6 network connecting B4 to AFTR other than what unicast DS-Lite solution requires.

DS-Lite B4 MUST support IGMP Proxy as defined in [RFC4605]. DS-Lite B4 MAY support MLD Proxy.

DS-Lite AFTR MUST support IGMP Querrier. DS-Lite AFTR MAY support MLD Querrier.

Both any source multicast (ASM) and source specific multicast (SSM) MUST be supported.

4. Architecture

In DS-Lite, there are hosts (possibly IPv4/ IPv6 dual stack) served by B4 element. B4 is dual stack facing the hosts and IPv6 only facing the network or WAN side. At the boundary of the network there is AFTR. AFTR receives IPv4 packets tunneled in IPv6 from B4 and decapsulates them and sends them out to IPv4 network.

In order to support multicast communication B4 implements IGMP Proxy function [RFC4605]. IPv4 hosts send their join requests (IGMP Membership Report messages) to B4. B4 as a proxy sends aggregated Report messages upstream towards AFTR.

AFTR is the default multicast querier for B4. AFTR implements multicast router function or it could be another IGMP proxy.

All the elements of DS-Lite multicast support system are shown in Figure 1.

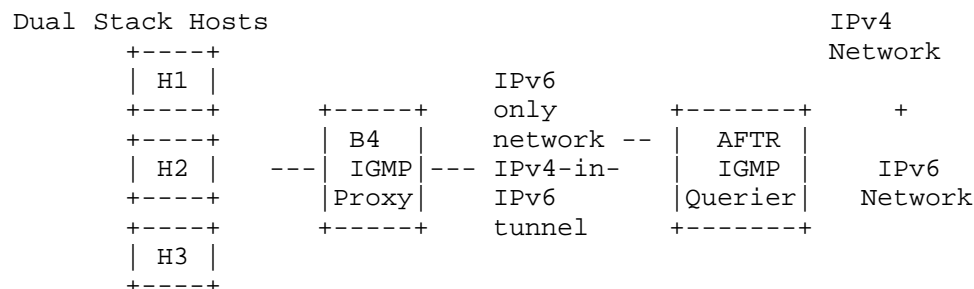


Figure 1: Architecture of DS-Lite Multicast Protocol

5. DS-Lite Multicast Operation

In this section we specify how the host can subscribe and receive IPv4 multicast data from IPv4 content providers based on the architecture defined in Section 4.

The hosts will send their subscription requests for IPv4 multicast groups upstream to the default router, i.e. B4 Element. After subscribing to the group, the host can receive multicast data from the B4. The host implements IGMP protocol's host part.

In order to support SSM, IGMPv3 MUST be supported by the host, B4 and

AFTR.

B4 Element is IGMP Proxy. After receiving the first IGMP Report message requesting subscription to an IPv4 multicast group, B4 establishes a tunnel interface with a AFTR. The tunnel is IPv6 based but it will carry IPv4 traffic, IGMP messages back and forth and IPv4 multicast data messages downstream. This is similar to [RFC6224] Section 4.4 but the operation is much simpler. In DS-Lite environment there is no requirement to handle host mobility. B4 does not have to keep more than one tunnel interfaces, a single interface is sufficient. IGMP Proxy at the B4 does not have to have more than one proxy instances, a single instance is sufficient.

B4 is regular IGMP proxy and it keeps IGMP proxy membership database. B4 inserts multicast forwarding state on the incoming interface, and merges state updates into the IGMP proxy membership database. B4 updates or removes elements from the database as required. B4 will then send an aggregated Report via the upstream tunnel to the AFTR when the membership database changes.

B4 answers IGMP queries from AFTR based on the membership database. B4's downstream link follows the traditional multipoint channel forwarding and does not pose any specific problems.

B4 receives IPv4 multicast data from the AFTR tunneled over the tunnel interface. B4 decapsulates the packet and then forwards it downstream. Each member host receives the data packet based on Layer 2 multicast interface. No packet duplication is necessary.

AFTR acts as the as the default multicast querier for all B4s that have established an IPv6 tunnel with it. In order to keep a consistent multicast state between a B4 and AFTR, once a B4 is connected it will stay connected until the state becomes empty. After that point, the B4 may continue to use the tunnel for IPv4 unicast traffic.

According to aggregated IGMP reports received from a B4, AFTR establishes group/source-specific multicast forwarding states at its corresponding downstream tunnel interfaces. After that, AFTR maintains or removes the state as required by the aggregated reports received from B4.

At the upstream interface, AFTR procures for aggregated multicast membership maintenance. Based on the multicast-transparent operations of the B4s, the AFTR treats its tunnel interfaces as multicast enabled downstream links, serving zero to many listening nodes.

Multicast traffic arriving at the AFTR is transparently forwarded according to its multicast forwarding information base. Multicast data is first replicated and then forwarded in IPv4-in-IPv6 tunnel from AFTR to the corresponding B4.

5.1. Tunnel Interface Considerations

Legacy IPv4 in IPv6 tunneling is performed as in [RFC2473] and [RFC4213]. Considerations specified in [RFC6333] apply. Packets upstream from B4 carry only IGMP signaling messages and they are not expected to fragmentation. However packets downstream, i.e. multicast data to B4 may be subject to fragmentation.

Source and destination addresses of IGMP messages in IPv4-in-IPv6 software from B4 is as follows:

Source address of IPv6 header is B4 IPv6 address, e.g. 2001:db8:0:1::1, destination address is AFTR address, e.g. 2001:db8:0:2::1.

Source address of IGMP messages is B4's IPv4 interface address, e.g. 192.0.0.2, destination address is the all-systems multicast address of 224.0.0.1 for IGMP Query, all IGMPv3-capable multicast routers of 224.0.0.22 for IGMPv3 Report, the multicast group specified in the Group Address field of the Report for IGMPv1 or IGMPv2 Report.

Source and destination addresses of IGMP messages in IPv4-in-IPv6 software from AFTR is as follows:

Source address of IPv6 header is AFTR address, e.g. 2001:db8:0:2::1, destination address is B4 IPv6 address, e.g. 2001:db8:0:1::1.

Source address of IGMP messages is AFTR's IPv4 interface address, e.g. 192.0.2.1, destination address is the all-systems multicast address of 224.0.0.1 for IGMP Query, all IGMPv3-capable multicast routers of 224.0.0.22 for IGMPv3 Report, the multicast group specified in the Group Address field of the Report for IGMPv1 or IGMPv2 Report.

Source and destination addresses of multicast data messages in IPv4-in-IPv6 software is as follows:

Source address of IPv6 header is AFTR address, e.g. 2001:db8:0:2::1, destination address is B4 IPv6 address, e.g. 2001:db8:0:1::1.

Source address of IPv4 multicast data is unicast IPv4 address of the multicast source, e.g. the content provider, destination address is IPv4 multicast group address.

AFTR decapsulates datagrams carrying IGMP messages from B4's and then IGMP router processing takes over. Network Address Translation (NAT) is not applied on IGMP messages.

6. Multicast Support for Host-Based Architecture

In this section we specify multicast support for Host-Based DS-Lite architecture described in Appendix B2 of [RFC6333].

In host-based DS-Lite, the host accesses the service provider network directly with an IPv6 global address. Host sends its IPv4 datagrams in IPv6 using an IPv4-in-IPv6 software tunnel to an AFTR, i.e. it implements DS-Lite B4. Source address of all IPv4 datagrams is the pre-configured well-known IPv4 non-routable address.

For multicast, there are two choices: the host could implement host side of IGMP protocol or for mobile router type of hosts, the host implements IGMP proxy as in Section 5.

Host encapsulates IGMP messages as described in Section 5.1 and sends them to AFTR. AFTR does not perform IPv4-IPv4 NAT translations on IGMP datagrams instead they are processed by IGMP router at the AFTR.

Multicast data received from AFTR for a multicast group that the host has subscribed is decapsulated by the host, if the host is IGMP client, it processes the data. If the host is IGMP proxy, it consults multicast state for the group and forwards the data downstream so that the members can receive the data.

7. Avalanche Problem

When multicast datagrams are received at the AFTR, AFTR consults its membership database and duplicates the packets for each member B4 interface and then these datagrams are forwarded in IPv4-in-IPv6 software downstream. This may cause an avalanche of downstream packets if the number of member B4's is high.

Avalanche problem can be eased by network partitioning. AFTR can be deployed closer to the users. For example in broadband networks, AFTR can be deployed at the Broadband Network Gateway (BNG) nodes.

8. IANA Considerations

None.

9. Acknowledgements

TBD.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.
- [RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick, "Internet Group Management Protocol (IGMP) / Multicast Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.
- [RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, December 1998.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.
- [RFC6224] Schmidt, T., Waehlich, M., and S. Krishnan, "Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains", RFC 6224, April 2011.
- [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-Stack Lite Broadband Deployments Following IPv4 Exhaustion", RFC 6333, August 2011.

10.2. Informative references

- [RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.
- [RFC4286] Haberman, B. and J. Martin, "Multicast Router Discovery", RFC 4286, December 2005.
- [RFC4541] Christensen, M., Kimball, K., and F. Solensky,

"Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006.

- [RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

Author's Address

Behcet Sarikaya
Huawei USA
5340 Legacy Drive Building 175
Plano, TX 75074

Phone:
Email: sarikaya@ieee.org

