Internet Engineering Task Force
Internet-Draft
Intended status: Experimental
Expires: October 27, 2013

R. Despres
RD-IPtech
S. Jiang, Ed.
Huawei Technologies Co., Ltd
R. Penno
Cisco Systems, Inc.
Y. Lee
Comcast
G. Chen
China Mobile
M. Chen
Freebit Co, Ltd.
April 25, 2013

IPv4 Residual Deployment via IPv6 - a Stateless Solution (4rd) draft-ietf-softwire-4rd-05

Abstract

The 4rd automatic tunneling mechanism makes IPv4 Residual Deployment possible via IPv6 networks without maintaining for this per-customer states in 4rd-capable nodes (reverse of the IPv6 Rapid Deployment of 6rd). To cope with the IPv4 address shortage, customer sites can be assigned shared public IPv4 addresses with restricted port sets. 4rd can also support the scenarios that customer sites are assigned full public IPv4 addresses or a set of public IPv4 addresses.

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 27, 2013.

Copyright Notice

Table of Contents

1. Introduction

For deployments of residual IPv4 service via IPv6 networks, the need for a stateless solution, i.e. one where no per-customer state is needed in IPv4-IPv6 gateway nodes of the provider, is expressed in [I-D.ietf-softwire-stateless-4v6-motivation]. This document specifies such a solution, named "4rd" for IPv4 Residual Deployment. With it, IPv4 packets are transparently tunneled across IPv6 networks (reverse of 6rd [RFC5969] in which IPv6 packets are statelessly tunneled across IPv4 networks). While IPv6 headers are too long to be mapped into IPv4 headers, so that 6rd requires encapsulation of full IPv6 packets in IPv4 packets, IPv4 headers can be reversibly translated into IPv6 headers in such a way that, during IPv6 domain traversal, UDP packets having checksums and TCP packets are valid IPv6 packets. IPv6-only middle boxes that perform deep-packet-inspection can operate on them, in particular for port inspection and web caches.

In order to deal with the IPv4-address shortage, customers can be assigned shared public IPv4 addresses, with statically assigned restricted port sets. As such, it is a particular application of the A+P approach of [RFC6346].

Deploying 4rd in the networks that have enough public IPv4 address, customer sites can also be assigned full public IPv4 addresses. 4rd also supports the scenarios that a set of public IPv4 addresses are assigned to customer sites.

The design of 4rd builds on a number of previous proposals made for IPv4-via-IPv6 transition technologies listed in Section 8.

In some use cases, IPv4-only applications of 4rd-capable customer nodes can also work with stateful NAT64s of [RFC6146], provided these are upgraded to support 4rd tunnels in addition their IP/ICMP translation of [RFC6145]. The advantage is then a more complete IPv4 transparency than with double translation.

How the 4rd model fits in the Internet architecture is summarized in Section 3. The protocol specification is detailed in Section 4. Section 5 and Section 6 respectively deal with security and IANA considerations. Previous proposals that influenced this specification are listed in Section 8. A few typical 4rd use cases are presented in Appendices.

2. Terminology

The key words "MUST", "SHOULD", "MAY", and "OPTIONAL" in this

document are to be interpreted as described in [RFC2119].

ISP: Internet-Service Provider. In this document, the service it offers can be DSL, fiber-optics, cable, or mobile. The ISP can also be a private-network operator.

4rd (IPv4 Residual Deployment): An extension of the IPv4 service where public IPv4 addresses can be statically shared among several customer sites, each one being assigned an exclusive port set. This service is supported across IPv6-routing domains.

4rd domain (or Domain): An ISP-operated IPv6 network across which 4rd is supported according to the present specification.

Tunnel packet: An IPv6 packet that transparently conveys an IPv4 packet across a 4rd domain. Its header has enough information to reconstitute the IPv4 header at Domain exit. Its payload is the original IPv4 payload.

CE (Customer Edge): A customer-side tunnel endpoint. It can be in a node that is a host, a router, or both.

BR (Border Relay): An ISP-side tunnel-endpoint. Because its operation is stateless (neither per CE nor per session state) it can be replicated in as many nodes as needed for scalability.

4rd IPv6 address: IPv6 address used as destination of a Tunnel packet sent to a CE or a BR.

NAT64+: An ISP NAT64 of [RFC6146] that is upgraded to support 4rd tunneling when IPv6 addresses it deals with are 4rd IPv6 addresses.

4rd IPv4 address: A public IPv4 address or, in case of a shared public IPv4 address, a public transport address (public IPv4 address plus port number).

PSID (Port-Set Identifier): A flexible-length field that algorithmically identifies a port set.

4rd IPv4 prefix: A flexible-length prefix that may be a a public IPv4 prefix, a public IPv4 address, or a public IPv4 address followed by a PSID.

Mapping rule: A set of parameters that are used by BRs and CEs to derive 4rd IPv6 addresses from 4rd IPv4 addresses. Mapping rules are also used by each CE to derive a 4rd IPv4 prefix from an IPv6 prefix that has been delegated.

EA bits (Embedded Address bits): Bits that are the same in a 4rd IPv4 address and in the 4rd IPv6 address derived from it.

BR mapping rule: The mapping rule applicable to off-domain IPv4 addresses (addresses reachable via BRs). It can also apply to some or all of CE-assigned IPv4 addresses.

CE mapping rule: A mapping rule that is applicable only to CE-assigned IPv4 addresses (shared or not).

NAT64+ mapping rule: Mapping rule applicable to IPv4 addresses reachable via a NAT64+.

CNP (Checksum Neutrality preserver): A field of 4rd IPv6 addresses that ensures that TCP-like checksums do not change when IPv4 addresses are replaced by 4rd IPv6 addresses.

4rd Tag: A 16-bit tag whose value permits, in 4rd CEs, BRs, and NAT64+s, to distinguish 4rd IPv6 addresses from other IPv6 addresses.

3. The 4rd Model

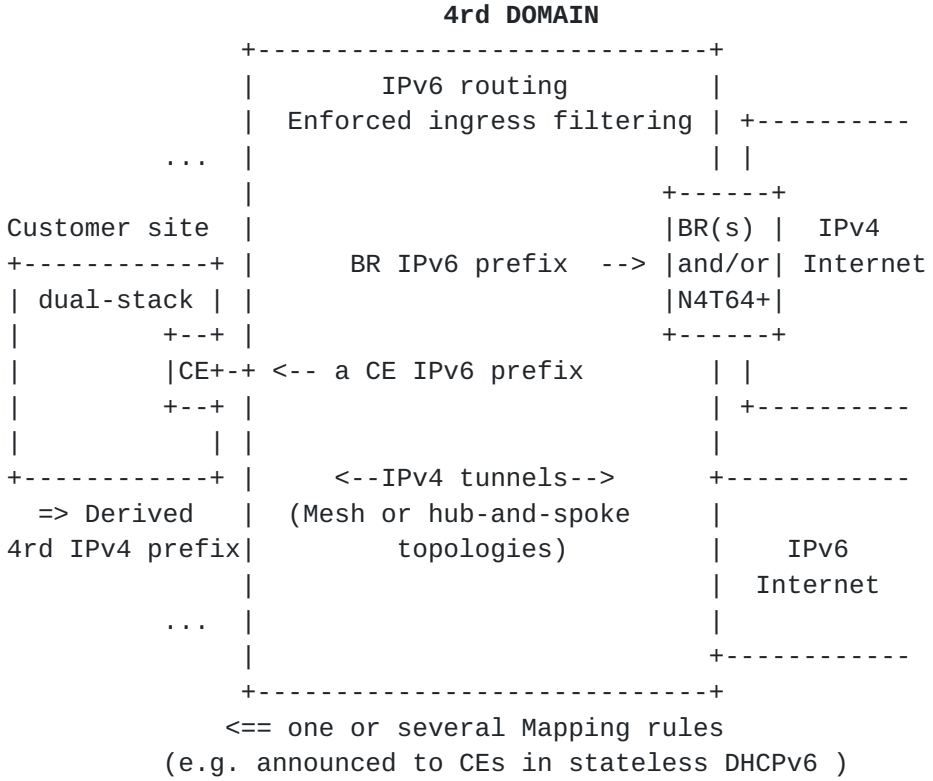


Figure 1

How the 4rd model fits in the Internet architecture is represented in Figure 1.

A 4rd domain is an IPv6 network that includes one or several 4rd BRs at its border with the public IPv4 Internet, and can advertise its IPv4-IPv6 Mapping rule(s) to CEs according to Section 4.9.

BRs of a 4rd Domain are all identical as far as 4rd is concerned. On 4rd CE, the IPv4 packets will be transformed (detailed in Section 4.3) into IPv6 packets that have the same anycast IPv6 prefix, which is the 80-bit BR prefix, in their destination addresses. They are then routed to any of the BRs. The 80-bit BR IPv6 prefix is an arbitrarily chosen /64 prefix from the IPv6 address space of the network operator and appended 0x0300 (16-bit 4rd Tag, see R-9 in Section 4.5).

Using the Mapping rule that applies, each CE derives its 4rd IPv4 prefix from its delegated IPv6 prefix, or one of them if it has several, details in Section 4.4. If the obtained IPv4 prefix has more than 32 bits, the assigned IPv4 address is shared among several CEs. Bits beyond the first 32 specify a set of ports whose use is reserved for the CE.

IPv4 traffic is automatically tunnelled across the Domain. By default, IPv4 traffic between two CEs follows a direct IPv6 route between them (mesh topology). If the ISP configures the Hub&spoke option, each IPv4 packet from a CE to another is routed via a BR.

During Domain traversal, each tunnelled TCP/UDP IPv4 packet looks like a valid TCP/UDP IPv6 packet. Thus, TCP/UDP access-control lists that apply to IPv6, and possibly some other functions using deep packet inspections, also apply to IPv4.

For IPv4 anti-spoofing protection to extend to IPv4, ingress filtering has to be effective in IPv6 (Section 4.4 and Section 5).

If an ISP wishes to support dynamic IPv4 address sharing, in addition or in place of 4rd stateless address sharing, it can do it by means of a stateful NAT64. By upgrading this NAT to add 4rd-tunnels support, which makes it a NAT64+, CEs that are assigned no static IPv4 space can benefit from complete IPv4 transparency between CE and NAT64. (Without this NAT64 upgrade, IPv4 traffic is translated to IPv6 and back to IPv4, which looses the DF=MF=1 combination of IPv4, that which is recommended for host fragmentation in Section 8 of [RFC4821].)

IPv4 packets are kept unchanged by Domain traversal except that:

- The IPv4 Time to live (TTL), unless it is 1 or 255 at Domain entry, decreases during Domain traversal by the number of traversed routers. This is acceptable because it is undetectable end to end, and because TTL values that can be used with some protocols to test adjacency of communicating routers are preserved ([RFC4271], [RFC5082]). Effect on the traceroute utility, which uses TTL expiry to discover routers of end-to-end paths, is noted in Section 4.3.

- IPv4 packets whose lengths are =< 68 octets always have their "Don't fragment flags" DF=1 at Domain exit even if they had DF=0 at Domain entry. This is acceptable because these packets are too short to be fragmented [RFC0791] so that their DF bits have no meaning. Besides, both [RFC1191] and [RFC4821] recommend that sources always set DF=1.

- Unless the Tunnel-traffic-class option applies to a Domain (Section 4.2), IPv4 packets may have their TOS fields modified after Domain traversal (Section 4.7).

4. Protocol Specifications

This section describes detailed 4rd protocol specifications. They are mainly organized by functions. As a brief summary, an 4rd CE SHOULD follow R-1, R-2, R-3, R-4, R-6, R-7, R-8, R-9, R-10, R-11, R-12, R-13, R-14, R-16, R-17, R-18, R-19, R-20, R-21, R-22, R-23, R-24, R-25, R-26 and R-27; while an 4rd BR SHOULD follow R-2, R-3, R-4, R-5, R-6, R-9, R-12, R-13, R-14, R-15, R-19, R-20, R-21, R-22 and R-24.

4.1. NAT44 on CE

R-1: A CE node who is assigned a shared public IPv4 address MUST include a NAT44 [RFC1631]. This NAT44 MUST only use external ports that are in the CE assigned port set.

NOTE: This specification only concerns IPv4 communication between IPv4-capable endpoints. For communication between IPv4-only endpoints and IPv6 only remote endpoints, the BIH specification of [RFC6535] can be used. It can coexist in a node with the CE function, including if the IPv4-only function is a NAT44 [RFC1631].

4.2. Mapping rules and other Domain parameters

R-2: CEs and BRs MUST be configured with the following Domain parameters:

A. One or several Mapping rules, each one comprising:

1. Rule IPv4 prefix

2. EA-bits length

3. Rule IPv6 prefix

4. WKPs authorized (OPTIONAL)

B. Domain PMTU

C. Hub&spoke topology (Yes or No)

D. Tunnel traffic class (OPTIONAL)

"Rule IPv4 prefix" is used to find, by a longest match, which Mapping rule applies to a 4rd IPv4 address (Section 4.5). A Mapping rule whose Rule IPv4 prefix is longer than /0 is a CE mapping rule. BR and NAT64+ mapping rules, which must apply to all off-domain IPv4

addresses, have /0 as their Rule IPv4 prefixes.

"EA-bits length" is the number of bits that are common to 4rd IPv4 addresses and 4rd IPv6 addresses derived from them. In a CE mapping rule, it is also the number of bits that are common to a CE delegated IPv6 prefix and the 4rd IPv4 prefix derived from it. BR and NAT64+ mapping rules have EA-bits lengths equal to 32.

"Rule IPv6 prefix" is the prefix that is substituted to the Rule IPv4 prefix when a 4rd IPv6 address is derived from a 4rd IPv4 address (Section 4.5). In a BR mapping rule or a NAT64+ mapping rule, it MUST be a /80 prefix whose 64~79 bits are the 4rd Tag.

"WKPs authorized" may be set for mapping rules that assign shared IPv4 addresses to CEs. (These rules are those whose length of the Rule IPv4 prefix plus the EA-bits length exceeds 32.) If set, well-known ports may be assigned to some CEs having particular IPv6 prefixes. If not set, fairness is privileged: all IPv6 prefixes concerned with the Mapping rule have ports sets having identical values (no port set includes any of the well known ports).

"Domain PMTU" is the IPv6 path MTU that the ISP can guarantee for all its IPv6 paths between CEs and between BRs and CEs. It MUST be at least 1280 [RFC2460].

"Hub&spoke topology", if set to Yes, requires CEs to tunnel all IPv4 packets via BRs. If set to No, CE-to-CE packets take the same routes as native IPv6 packets between the same CEs (mesh topology).

"Tunnel traffic class", if provided, is the IPv6 traffic class that BRs and CEs MUST set in Tunnel packets. In this case, evolutions of the IPv6 traffic class that may occur during Domain traversal are not reflected in TOS fields of IPv4 packets at Domain exit (Section 4.7).

4.3. Reversible Packet Translations at Domain entries and exits

R-3: Domain-entry nodes that receive IPv4 packets with IPv4 options MUST discard these packets, and return ICMPv4 error messages to signal IPv4-option incompatibility (Type = 12, Code = 0, Pointer = 20) [RFC0792]. This limitation is acceptable because there are a lot firewalls in current IPv4 Internet also filter IPv4 packets with IPv4 options.

R-4: Domain-entry nodes that receive IPv4 packets without IPv4 options MUST convert them to Tunnel packets, with or without IPv6 fragment headers depending on what is needed to ensure IPv4 transparency (Figure 2). Domain-exit nodes MUST convert them back to IPv4 packets.

An IPv6 fragmentation header MUST be included at tunnel entry (Figure 2) if, and only if, one or several of the following conditions hold:

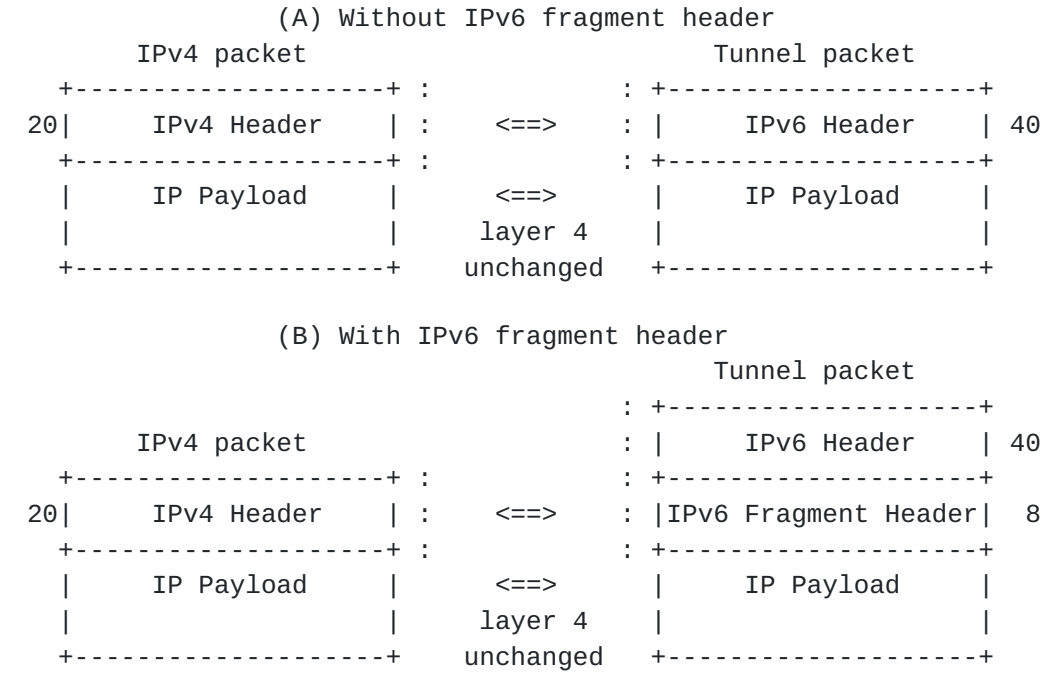
* The Tunnel_traffic_class option applies to the Domain.

* TTL = 1 OR TTL = 255.

* The IPv4 packet is already fragmented, or may be fragmented later on, i.e. if MF=1 OR Offset>0 OR (Total length > 68 AND DF=0).

In order to optimize cases where fragmentation headers are unnecessary, the NAT44 of a CE that has one SHOULD send packets with TTL = 254.

R-5: In Domains whose chosen topology is Hub&spoke, BRs that receive 4rd IPv6 packets whose embedded destination IPv4 addresses match a CE mapping rule MUST do the equivalent of reversibly translating their headers to IPv4 and then reversibly translate them back to IPv6 as though packets would be entering the Domain.

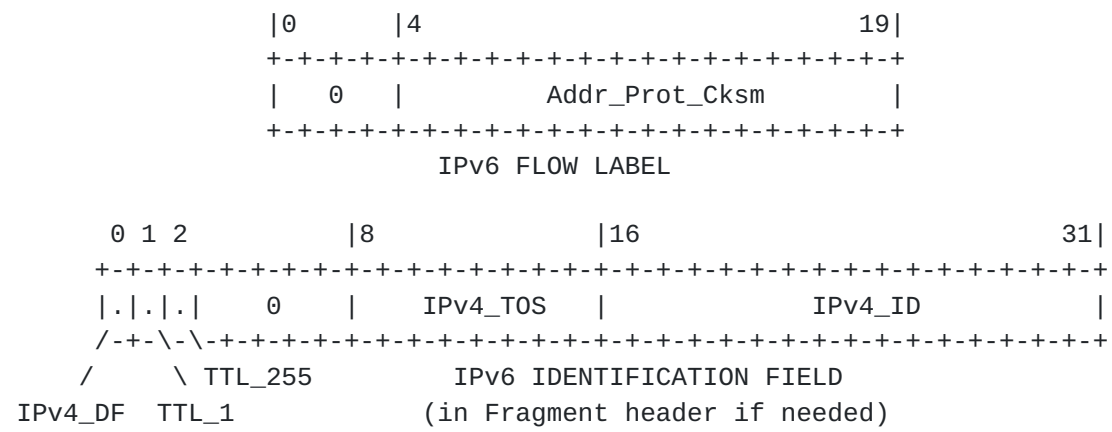


Reversible Packet Translation

Figure 2

R-6: Values to be set in IPv6-header fields at Domain entry are detailed in Table 1 (no-fragment-header case) and Table 2 (fragment-header case). Those to be set in IPv4 header fields at Domain exit are detailed in Table 3 (no-fragment-header case) and Table 4 (fragment-header case).

To convey IPv4-header informations that have no equivalent in IPv6, some ad-hoc fields are placed in IPv6 flow labels and in Identification fields of IPv6 fragment headers, as detailed in Figure 3.



4rd Identification fields of IPv6 Fragment headers

Figure 3

IPv6 FIELD	VALUE (fields from IPv4 header)
Version	6
Traffic class	TOS
Addr_Prot_Cksm	Sum of Addresses and Protocol (Note 1)
Payload length	Total length - 20
Next header	Protocol
Hop limit	Time to live
Source address	See Section 4.5
Destination address	See Section 4.5

IPv4-to-IPv6 Reversible Header Translation (without Fragment header)

Table 1

IPv6 FIELD	VALUE (fields from IPv4 header)
Version	6
Traffic class	TOS OR Tunnel_traffic_class (Section 4.7)
Addr_Prot_Cksm	Sum of Addresses and Protocol (Note 1)
Payload length	Total length - 12
Next header	44 (Fragment header)
Hop limit	IF Time to live = 1 or 255 THEN 254
	ELSE Time to live (Note 2)
Source address	See Section 4.5
Dest. address	See Section 4.5
2nd next header	Protocol
Fragment offset	IPv4 Fragment offset
M	More-fragments flag (MF)
IPv4_DF	Don't-fragment flag (DF)
TTL_1	IF Time to live = 1 THEN 1 ELSE 0 (Note 2)
TTL_255	IF Time to live = 255 THEN 1 ELSE 0 (Note 2)
IPv4_TOS	Type of service (TOS)
IPv4_ID	Identification

IPv4-to-IPv6 Reversible Header Translation (with Fragment header)

Table 2

IPv4 FIELD	VALUE (fields from IPv6 header)
Version	4
Header length	5
TOS	Traffic class
Total Length	Payload length + 20
Identification	0
DF	1
MF	0
Fragment offset	0
Time to live	Hop count
Protocol	Next header
Header checksum	Computed as per [RFC0791] (Note 3)
Source address	Bits 80-111 of source address
Dest. address	Bits 80-111 of source address

IPv6-to-IPv4 Reversible Header Translation (without Fragment header)

Table 3

IPv4 FIELD	VALUE (fields from IPv6 headers)
Version	4
Header length	5
TOS	Traffic class OR IPv4_TOS (Section 4.7)
Total Length	Payload length + 12
Identification	IPv4_ID
DF	IPv4_DF
MF	M
Fragment offset	Fragment offset
Time to live (Note 2)	IF TTL_255 = 1 THEN 255TTL_1 = 1 THEN 1 ELSEIF TTL_1 = 1 THEN 1 ELSE Hop count
Protocol	2nd Next header
Header checksum	Computed as per [RFC0791] (Note 3)
Source address	Bits 80-111 of source address
Destination address	Bits 80-111 of destination address

IPv6 to IPv4 Reversible Header Translation (with Fragment header)

Table 4

NOTE 1: The need to save in the IPv6 header a checksum of both IPv4 addresses and the IPv4 protocol field results from the following facts: (1) Header checksums, present in IPv4 but not in IPv6, protect addresses or protocol integrity; (2) In IPv4, ICMP messages and null-checksum UDP datagram depend on this protection because, unlike other datagram, they have no other address-and-protocol integrity protection. The sum MUST be performed in ordinary 2's complement arithmetic.

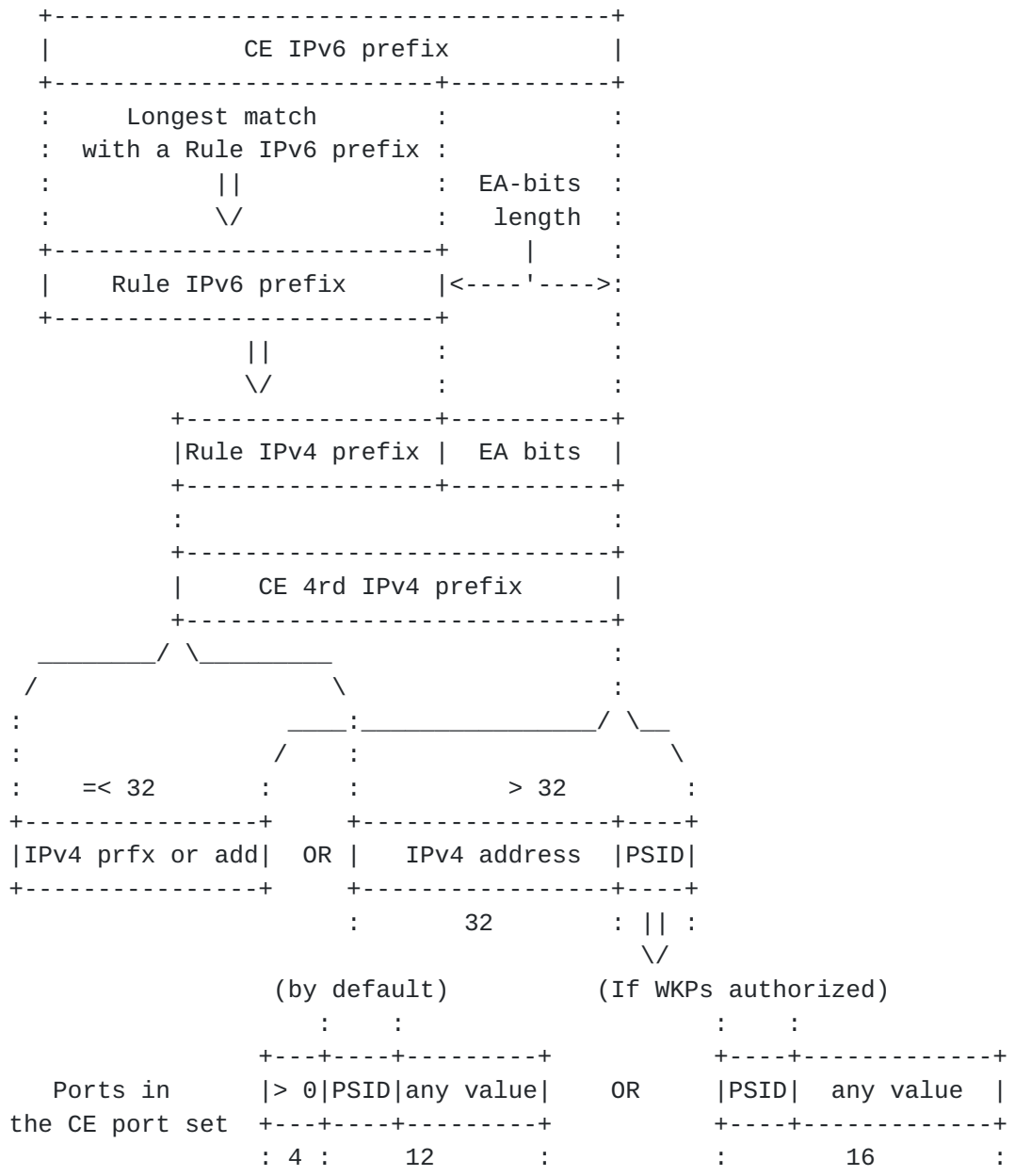
IP-layer Packet length is another field covered by the IPv4 IP-header checksum. It is not included in the saved checksum because: (1) doing so would have conflicted with [RFC6437] (flow labels must be the same in all packets of each flow); (2) ICMPv4 messages have good enough protection with their own checksums; (3) the UDP length field provides to null-checksum UDP datagrams the same level of protection after Domain traversal as without Domain traversal (consistency between IP-layer and UDP-layer lengths can be checked).

NOTE 2: TTL treatment has been chosen to permit adjacency tests between two IPv4 nodes situated at both end of a 4rd tunnel. TTL values to be preserved for this are TTL=255 and TTL=1. For other values, TTL decrease between to IPv4 nodes is the same as though traversed IPv6 routers would be IPv4 routers.

Effect of this TTL treatment on IPv4 traceroute (section 1 of [RFC1393]) is specific: (1) the number of routers of the end-to-end path includes traversed IPv6 routers; (2) IPv6 routers of a Domain are listed after IPv4 routers of Domain entry and exit; (3) the IPv4 address shown for an IPv6 router is the IPv6-only dummy IPv4 address of Section 4.8; (4) the response time indicated for an IPv6 router is that of the next router.

NOTE 3: Provided the sum of obtained IPv4 addresses and protocol matches Addr_Prot_Cksm. If not, the packet MUST be silently discarded.

4.4. Address Mapping from CE IPv6 Prefixes to 4rd IPv4 prefixes



From CE IPv6 prefix to 4rd IPv4 address and Port set

Figure 4

R-7: A CE whose delegated IPv6 prefix matches the Rule IPv6 prefix of one or several Mapping rules MUST select the CE mapping rule for which the match is the longest. It then derives its 4rd IPv4 prefix as shown in Figure 4: (1) the CE replaces the Rule IPv6 prefix by the Rule IPv4 prefix. The result is the CE 4rd IPv4 prefix. (2) If this CE 4rd IPv4 prefix has less than 32

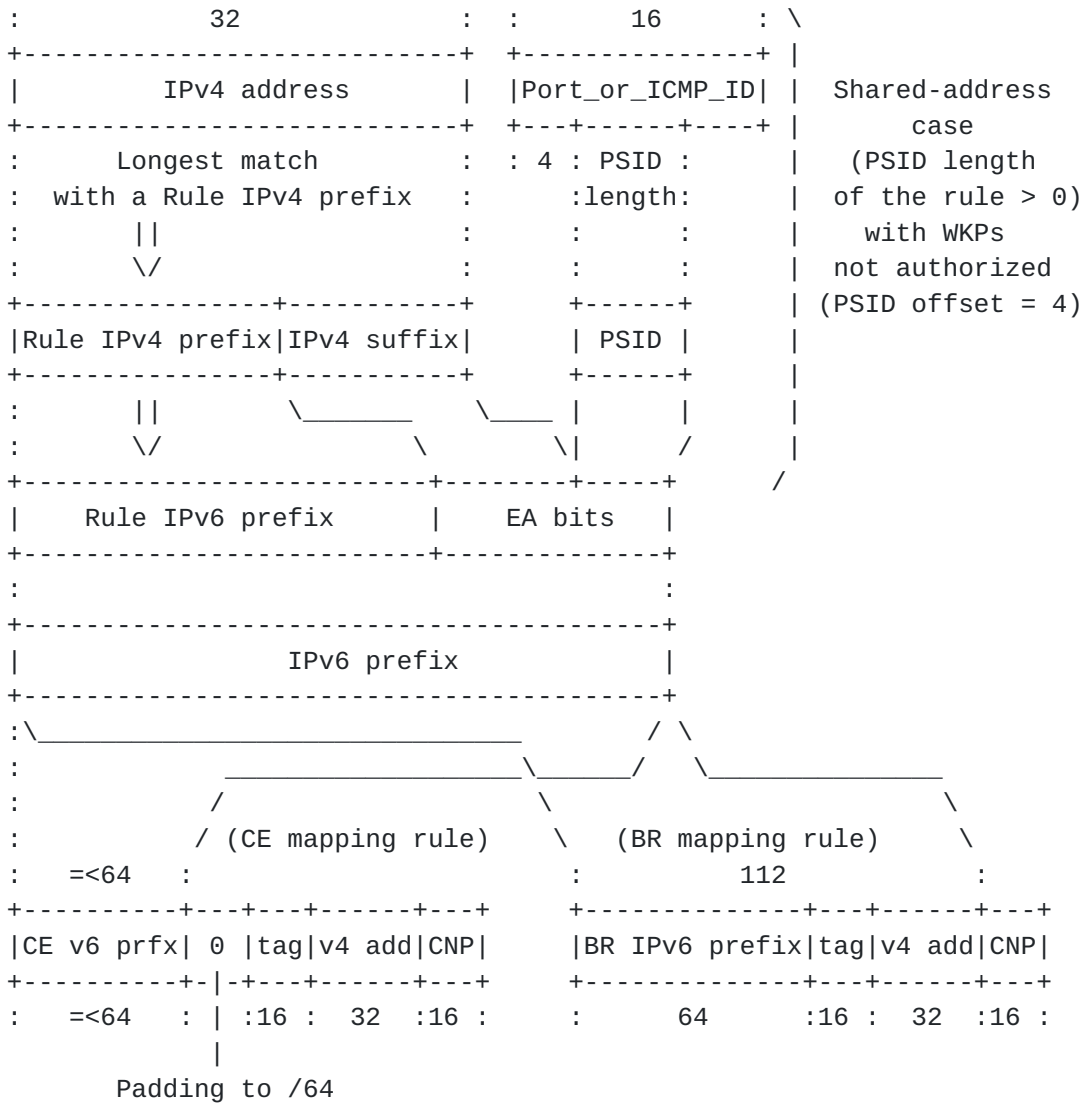
bits, the CE takes it as its assigned IPv4 prefix. If it has exactly 32 bits, the CE takes it as its IPv4 address. If it has more than 32 bits, the CE MUST takes the first 32 bits as its shared public IPv4 address, and bits beyond the first 32 as its Port-set identifier (PSID). Ports of its restricted port set are by default those that have any non-zero value in their first 4 bits (the PSID offset), followed by the PSID, and followed by any values in remaining bits. If the WKP authorized option applies to the Mapping rule, there is no 4-bit offset before the PSID so that all ports can be assigned.

NOTE: The choice of the default PSID position in Port fields has been guided by the following objectives: (1) for fairness, avoid having any of the well-known ports 0-1023 in the port set specified by any PSID value; (2) for compatibility RTP/RTCP [RFC4961], include in each port set pairs of consecutive ports; (3) in order to facilitate operation and training, have the PSID at a fixed position in port fields; (4) in order to facilitate documentation in hexadecimal notation, and to facilitate maintenance, have this position nibble aligned. Ports that are excluded from assignment to CEs are 0-4095 instead of just 0-1023 in a trade-off to favor nibble alignment of PSIDs and overall simplicity.

R-8: A CE whose delegated IPv6 prefix has its longest match with the Rule IPv6 prefix of the BR mapping rule MUST take as IPv4 address the 32 bit that, in the delegated IPv6 prefix, follow this Rule IPv6 prefix. If this is the case while the Hub&spoke option applies to the Domain, or if the Rule IPv6 prefix is not a /80, there is a configuration error in the Domain. An implementation-dependent administrative action MAY be taken.

A CE whose delegated IPv6 prefix matches the Rule IPv6 prefix of neither any CE Mapping rule nor the BR mapping rule, and is in a Domain that has a NAT64+ mapping rule, MUST be noted as having the unspecified IPv4 address.

4.5. Address Mapping from 4rd IPv4 addresses to 4rd IPv6 Addresses



From 4rd IPv4 address to 4rd IPv6 address

Figure 5

R-9: BRs, and CEs that are assigned public IPv4 addresses, shared or not, MUST derive 4rd IPv6 addresses from 4rd IPv4 addresses by the steps below or their functional equivalent (Figure 5 details the shared public IPv4 address case):

(1) If Hub&spoke topology does not apply to the Domain, or if it applies but the IPv6 address to be derived is a source address from a CE or a destination address from a BR, find the CE mapping rule whose Rule IPv4 prefix has the longest match with the IPv4 address.

If no Mapping rule is thus obtained, take the BR mapping rule.

If the obtained Mapping rule assigns IPv4 prefixes to CEs, i.e. if length of the Rule IPv4 prefix plus EA-bits length is 32 - k, with k >= 0, delete the last k bits of the IPv4 address.

Otherwise, i.e. if length of the Rule IPv4 prefix plus EA-bits length is 32 + k, with k > 0, take k as PSID length, and append to the IPv4 address the PSID copied from bits p to p+3 of the Port_or_ICMP_ID field where: (1) p, the PSID offset, is 4 by default, and 0 if the WKPs authorized option applies to the rule; (2) The Port_or_ICMP_ID field is in bits of the IP payload that depend on whether the address is source or destination, on whether the packet is ICMP or not, and, if it is ICMP, whether it is an error message or an echo message. This field is:

a. If the packet Protocol is not ICMP, the port field associated with the address (bits 0-15 for a source address, and bits 16-31 for a destination address).

b. If the packet is an ICMPv4 echo or echo-reply message, the ICMPv4 Identification field (bits 32-47).

c. If the packet is an ICMPv4 error message, the port field associated with the address in the returned packet header (bits 240-255 for a source address, bits 224-239 for a destination address).

NOTE 1: Using Identification fields of ICMP messages as port fields permits to exchange Echo requests and Echo replies between shared-address CEs and IPv4 hosts having exclusive IPv4 addresses. Echo exchanges between two shared-address CEs remain impossible, but this is a limitation inherent to address sharing (one reason among many to use IPv6).

NOTE 2: When the PSID is taken in the port field of the IPv4 payload, it is, to avoid dependency on any particular layer-4 protocol having port fields, without checking that the protocol is indeed one that has a port field . A packet may consequently go, in case of source mistake, from a BR to a shared-address CE with a protocol

that is not supported by this CE. In this case, the CE NAT44 returns an ICMPv4 "protocol unreachable" error message. The IPv4 source is thus appropriately informed of its mistake.

(2) Replace in the result the Rule IPv4 prefix by the Rule IPv6 prefix.

(3) If the result is shorter than a /64, append to the result a null padding up to 64 bits, followed by the 4rd tag (0x0300), and followed by the IPv4 address.

NOTE: The 4rd tag is a 4rd-specific mark. Its function is to ensure that 4rd IPv6 addresses are recognizable by CEs without any interference with the choice of subnet prefixes in CE sites. (These choices may have been done before 4rd is enabled.)

For this, the 4rd tag has its "u" and "g" bits of [RFC4291] both set to 1, so that they maximumly differ from these existing IPv6 address schemas. So far, u=g=1 has not been used in any IPv6 addressing architecture.

With the 4rd tage, IPv6 packets can be routed to the 4rd function within a CE node based on a /80 prefix that no native-IPv6 address can contain.

(4) Add to the result a Checksum-neutrality preserver (CNP). Its value, in one's complement arithmetic, is the opposite of the sum of 16-bit fields of the IPv6 address other than the IPv4 address and the CNP themselves (i.e. 5 consecutive fields in address-bits 0-79).

NOTE: CNP guarantees that Tunnel packets are valid IPv6 packets for all layer-4 protocols that use the same checksum algorithm as TCP. This guarantee does not depend on where checksum fields of these protocols are placed in IP payloads. (Today, such protocols are UDP [RFC0768], TCP [RFC0793], UDP-Lite [RFC3828], and DCCP [RFC5595]. Should new ones be specified, BRs will support them without needing an update.)

R-10: 4rd-capable CE SHOULD always prohibit all addresses that use its advertised prefix and have IID starting with 0x0300 (4rd Tag), by using Duplicate Address Detection [RFC4862].

R-11: A CE that is assigned the unspecified IPv4 address (see Section 4.4) MUST use, for packets tunneled between itself and the Domain NAT64+, addresses as detailed in Figure 6:(a) for its IPv6 source, (b) as IPv6 destinations that depend on IPv4 destinations. The NAT64+, being NAT64 conforming [RFC6146], uses (b) as source addresses that depend on IPv4 sources. It finds in its binding information base addresses conforming to (a). Finding the 4rd tag in them, it uses 4rd tunneling rather than IPv4 to IPv6 translation.

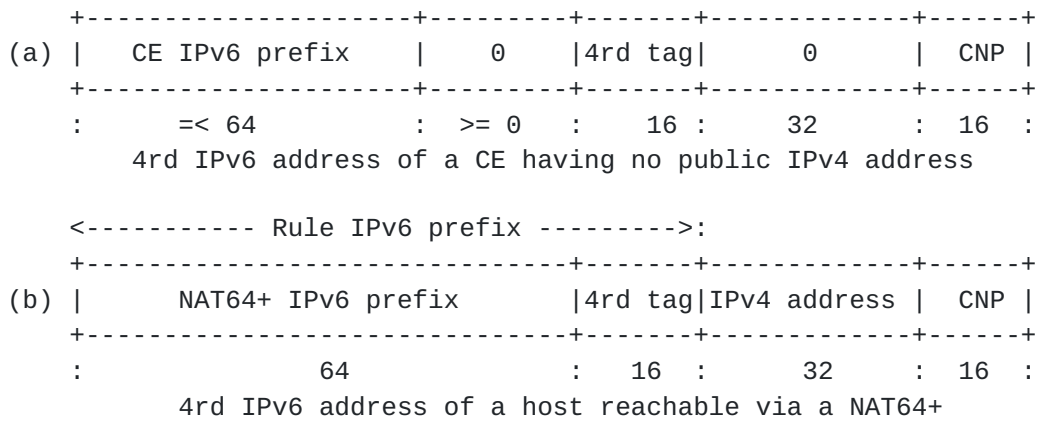


Figure 6

R-12: For anti-spoofing protection, CEs and BRs MUST check that the source address of each received Tunnel packet is that which, according to Section 4.5, is derived from the source 4rd IPv4 address. For this, the IPv4 address used to obtain the source 4rd IPv4 address is that embedded in the IPv6 source address (in its bits 80-111). (This verification is needed because IPv6 ingress filtering [RFC3704] applies only to IPv6 prefixes, without guarantee that Tunnel packets are built as specified in Section 4.5.)

R-13: For additional protection against packet corruption at a link layer that might be undetected at this layer during Domain traversal, CEs and BRs SHOULD verify that source and destination IPv6 addresses have not been modified. This can be done by checking that they remain checksum neutral (see the Note on CNP above).

4.6. Fragmentation Processing

4.6.1. Fragmentation at Domain Entry

R-14: If an IPv4 packet enters a CE or BR with a size such that the derived Tunnel packet would be longer than the Domain PMTU, the packet has to be either discarded or fragmented. The Domain-entry node MUST discard it if the packet has DF=1, with an ICMP error message returned to the source. It MUST fragment it otherwise, with the payload of each fragment not exceeding PMTU - 48. The first fragment has its offset equal to the received offset. Following fragments have offsets increased by lengths of previous-fragments payloads. Functionally, fragmentation is supposed to be done in IPv4 before applying to each fragment the reversible header translation of Section 4.3.

4.6.2. Ports of Fragments addressed to Shared-Address CEs

Because ports are available only in first fragments of IPv4 fragmented packets, a BR needs a mechanism to send to the right shared-address CEs all fragments of fragmented packets.

For this, a BR MAY systematically reassemble fragmented IPv4 packets before tunneling them. But this consumes large memory space, opens denial-of-service-attack opportunities, and can significantly increase forwarding delays. This is the reason for the following requirement:

R-15: BRs SHOULD support an algorithm whereby received IPv4 packets can be forwarded on the fly. The following is an example of such algorithm:

(1) At BR initialization, if at least one CE mapping rule concerns shared public IPv4 addresses (length of Rule IPv4 prefix + EA-bits length > 32), the BR initializes an empty "IPv4-packet table" whose entries have the following items:

- IPv4 source
- IPv4 destination
- IPv4 identification
- Destination port

(2) When the BR receives an IPv4 packet whose matching Mapping rule is one of shared public IPv4 addresses (length of Rule IPv4 prefix + EA-bits length > 32), the BR searches the table for an entry whose IPv4 source, IPv4 destination, and IPv4 Identification, are those of

the received packet. The BR then performs actions detailed in Table 5 depending on which conditions hold.

- CONDITIONS -								
First Fragment (offset=0)	Y	Y	Y	Y	N	N	N	N
Last fragment (MF=0)	Y	Y	N	N	Y	Y	N	N
An entry has been found	Y	N	Y	N	Y	N	Y	N
- RESULTING ACTIONS -								
Create a new entry	-	-	-	X	-	-	-	-
Use port of the entry	-	-	-	-	X	-	X	-
Update port of the entry	-	-	X	-	-	-	-	-
Delete the entry	X	-	-	-	X	-	-	-
Forward the packet	X	X	X	X	X	-	X	-

Table 5

(3) The BR performs garbage collection for table entries that remain unchanged for longer than some limit. This limit, normally longer that the maximum time normally needed to reassemble a packet is not critical. It should however not be longer than 15 seconds [RFC0791].

R-16: For the above algorithm to be effective, CEs that are assigned shared public IPv4 addresses MUST NOT interleave fragments of several fragmented packets.

R-17: CEs that are assigned IPv4 prefixes, and are in nodes that route public IPv4 addresses rather than only using NAT44s, MUST have the same behavior as described just above for BRs.

4.6.3. Packet Identifications from Shared-Address CEs

When packets go from CEs that share the same IPv4 address to a common destination, a precaution is needed to guarantee that packet Identifications set by sources are different. Otherwise, packet reassembly at destination could otherwise be confused because it is based only on source IPv4 address and Identification. Probability of such confusions may in theory be very low but, in order to avoid creating new attack opportunities, a safe solution is needed.

R-18: A CE that is assigned a shared public IPv4 address MUST only use packet Identifications that have the CE PSID in their bits 0 to PSID length - 1.

R-19: A BR or a CE that receives a packet from a shared-address CE MUST check that bits 0 to PSID length - 1 of their packet Identifications are equal to the PSID found in source 4rd IPv4 address.

4.7. TOS and Traffic-Class Processing

IPv4 TOS and IPv6 Traffic class have the same semantic, that of the differentiated-services field, or DS field, specified in [RFC2474] and [RFC6040]. Their first 6 bits contain a differentiated services codepoint (DSCP), and their two last bits can convey explicit congestion notifications (ECNs), which both may evolve during Domain traversal. [RFC2983] discusses how the DSCP can be handled by tunnel end points. The Tunnel traffic class option permits to ignore DS-field evolutions occurring during Domain traversal, if the desired behavior is that of generic tunnels conforming to [RFC2473].

R-20: Unless the Tunnel traffic class option is configured for the Domain, BRs and CEs MUST copy the IPv4 TOS into the IPv6 Traffic class at Domain entry, and copy back the IPv6 Traffic class into the IPv4 TOS at Domain exit.

R-21: If the Tunnel traffic class option is configured for a Domain, BRs and CEs MUST at Domain entry take the configured Tunnel traffic class as IPv6 Traffic class, and copy the received IPv4 TOS into the IPv4_TOS of the fragment header (Figure 3). At Domain exit, they MUST copy back the IPv4_TOS of the fragment header into the IPv4 TOS.

4.8. Tunnel-Generated ICMPv6 Error Messages

If a Tunnel packet is discarded on its way across a 4rd domain because of an unreachable destination, an ICMPv6 error message is returned to the IPv6 source. For the IPv4 source of the discarded packet to be informed of packet loss, the ICMPv6 message has to be converted into an ICMPv4 message.

R-22: If a CE or BR receives an ICMPv6 error message [RFC4443], it MUST synthesize an ICMPv4 error packet [RFC0792]. This packet MUST contain the first 8 octets of the discarded-packet IP payload. The reserved IPv4 dummy address (TBD, (see Section 6) MUST be used as its source address .

Like in [RFC6145], ICMPv6 Type = 1 and Code = 0 (Destination

unreachable, No route to destination") MUST be translated into ICMPv4 Type = 3 and Code = 0 (Destination unreachable, Net unreachable), and ICMPv6 Type = 3 and Code = 0 (Time exceeded, Hop limit exceeded in transit) MUST be translated into ICMPv4 Type = 11 and Code = 0 (Destination unreachable, Net unreachable).

4.9. Provisioning 4rd Parameters to CEs

Domain parameters listed in Section 4.2 are subject to the following constraints:

R-23: Each Domain MUST have a BR mapping rule and/or a NAT64+ mapping rule. (The BR mapping rule is only used by CEs that are assigned public IPv4 addresses, shared or not. The NAT64+ mapping rule is only used by CEs that are assigned the unspecified IPv4 address (Section 4.4), and therefore need an ISP NAT64 to reach IPv4 destinations.

R-24: Each CE and each BR MUST support up to 32 Mapping rules.

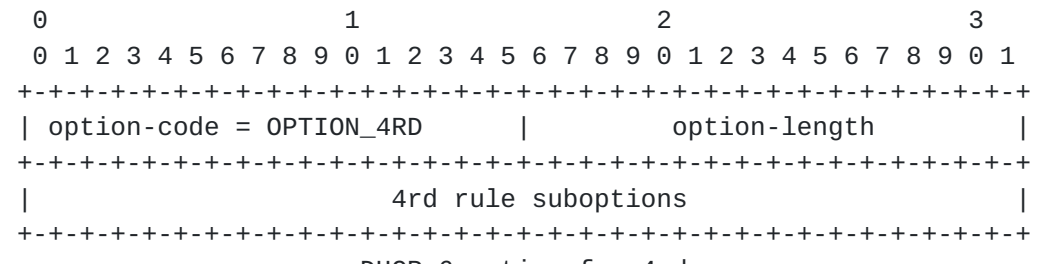
This number of is to ensure that independently acquired CEs an BR nodes can always interwork. (Its value, which is not critical, can easily be changed if another value would be found more desirable by the WG.)

ISPs that need Mapping rules for more IPv4 prefixes than this number SHOULD split their networks into multiple Domains. Communication between these domains can be done in IPv4, or by some implementation-dependent but equivalent other means.

R-25: For mesh topologies, where CE-CE paths don't go via BRs, all mapping rules of the Domain MUST be sent to all CEs. For hub-and-spoke topologies, where all CE-CE paths go via BRs, each CE MAY be sent only the BR mapping rule of the Domain plus, if different, the CE mapping rule that applies to its CE IPv6 prefix.

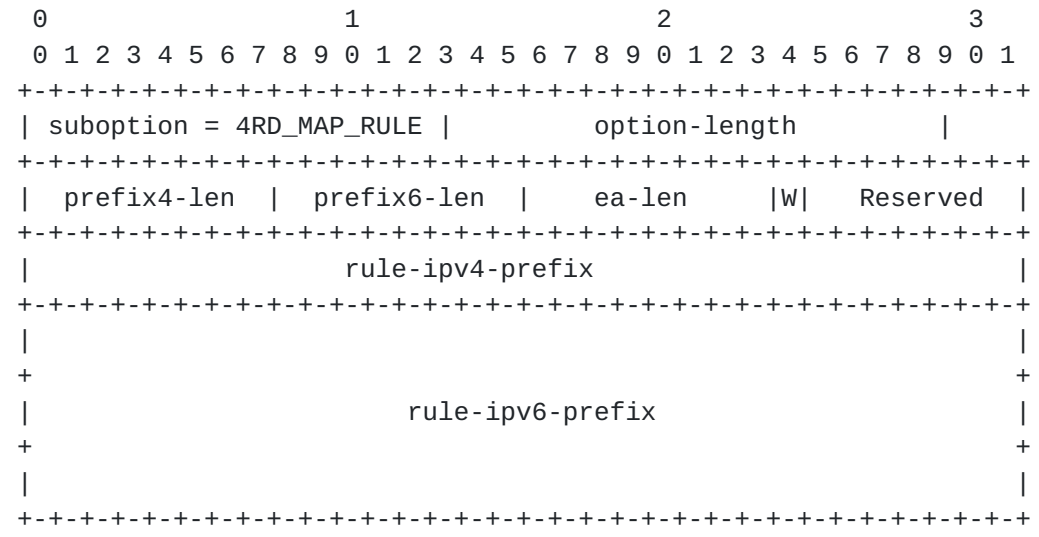
R-26: In a Domain where the chosen topology is Hub&spoke, all CEs MUST have IPv6 prefixes that match a CE mapping rule. (Otherwise, packets sent to CEs whose IPv6 prefixes would match only the BR mapping rule would, with longest-match selected routes, be routed directly to these CEs. This would be contrary to the Hub&spoke requirement).

R-27: CEs MUST be able to acquire parameters of 4rd domains (Section 4.2) in DHCPv6 (ref. [RFC2131]). Formats of DHCPv6 options to be used are detailed in Figure 7 and Figure 8, with field values specified after each Figure.



DHCPv6 option for 4rd

- option-code: TBD1, OPTION_4RD (see Section 6)
- option-length: the length of suboptions in octets
- 4rd rule suboptions: the 4RD DHCPv6 option SHOULD contain at least one 4RD_MAP_RULE suboption and maximum one 4RD_NON_MAP_RULE suboption. the length of suboptions in octets

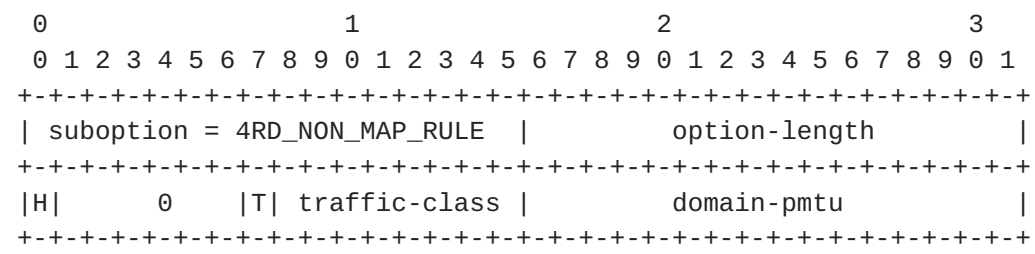


suboption for Mapping-rule parameters

Figure 7

- option-code: 0, 4RD_MAP_RULE suboption (see Section 6)
- option-length: 20
- prefix4-len: number of bits of the Rule IPv4 prefix

- prefix6-len: number of bits of the Rule IPv6 prefix
- ea-len: EA-bits length
- W: WKP authorized, = 1 if set
- rule-ipv4-prefix: the Rule IPv4 prefix, left aligned
- rule-ipv6-prefix: Rule IPv6 prefix, left aligned



suboption for non-mapping-rule parameters of 4rd-domains

Figure 8

- suboption-code: 1, 4RD_NON_MAP_RULE suboption (see Section 6)
- option-length: 4
- H: Hub&spoke topology (= 1 if Yes)
- T: Traffic-class flag (= 1 if a Tunnel traffic class is provided)
- traffic-class: Tunnel-traffic class
- domain-pmtu: Domain PMTU (at least 1280)

Other means than DHCPv6 that may prove useful to provide 4rd parameters to CEs are off-scope for this document. The same or similar parameter formats would however be recommended to facilitate training and operation.

5. Security Considerations

Spoofing attacks

With IPv6 ingress filtering effective in the Domain [RFC3704], and with consistency checks between 4rd IPv4 and IPv6 addresses of Section 4.5, no spoofing opportunity in IPv4 is introduced by 4rd.

Routing-loop attacks

Routing-loop attacks that may exist in some automatic-tunneling scenarios are documented in [RFC6324]. No opportunity for routing-loop attacks has been identified with 4rd.

Fragmentation-related attacks

As discussed in Section 4.6, each BR of a Domain that assigns shared public IPv4 should maintain a dynamic table for fragmented packets that go to these shared-address CEs.

This opens a BNR vulnerability to a denial of service attack from hosts that would send very large numbers of first fragments and would never send last fragments having the same packet identifications. This vulnerability is inherent to IPv4 address sharing, be it static or dynamic. Compared to what it is with algorithms that reassemble IPv4 packets in BRs, it is however significantly mitigated by the algorithm of Section 4.6.2 which uses much less memory space.

6. IANA Considerations

IANA is requested to allocate the following:

- One DHCPv6 option codes TBD1 for OPTION_4RD of Section 4.9 respectively (to be added to section 24.3 of [RFC3315]. Suboption values of 4RD_MAP_RULE (0) and 4RD_NON_MAP_RULE (1) should also be recorded into the DHCPv6 option code space.

- A reserved IPv4 address to be used as the "IPv4 dummy address" of Section 4.8. Its proposed value is 192.70.192.254 (Section 4.8). This address is taken in the /24 range that has been proposed for a similar purpose in [draft-xli-behave-icmp-address-04]. It is subject to IANA confirmation.

7. Relationship with Previous Works

The present specification has been influenced by many previous IETF drafts, in particular those accessible at http://tools.ietf.org/html/draft-xxxx where xxxx are the following

(in order of their first versions):

- bagnulo-behave-nat64 (2008-06-10)
- xli-behave-ivi (2008-07-06)
- despres-sam-scenarios (2008-09-28)
- boucadair-port-range (2008-10-23)
- ymbk-aplusp (2008-10-27)
- xli-behave-divi (2009-10-19)
- thaler-port-restricted-ip-issues (2010-02-28)
- cui-softwire-host-4over6 (2010-05-05)
- xli-behave-divi-pd (2011-07-02)
- dec-stateless-4v6 (2011-03-05)
- matsushima-v6ops-transition-experience (2011-03-07)
- despres-intarea-4rd (2011-03-07)
- deng-aplusp-experiment-results (2011-03-08)
- murakami-softwire-4rd (2011-07-04)
- operators-softwire-stateless-4v6-motivation (2011-05-05)
- murakami-softwire-4v6-translation (2011-07-04)
- despres-softwire-4rd-addmapping (2011-08-19)
- boucadair-softwire-stateless-requirements (2011-09-08)
- chen-softwire-4v6-add-format (2011-10-2)
- mawatari-softwire-464xlat (2011-10-16)
- mdt-softwire-map-dhcp-option (2011-10-24)
- mdt-softwire-mapping-address-and-port (2011-11-25)
- mdt-softwire-map-translation (2012-01-10)

o mdt-softwire-map-encapsulation (2012-01-27)

8. Acknowledgements

This specification has benefited over several years from independent proposals, questions, comments, constructive suggestions, and useful criticisms, coming from numerous IETF contributors.

Authors would like to express recognition to all these contributors, and more especially to the following, in alphabetical order of first names: Brian Carpenter, Behcet Sarikaya, Bing Liu, Cameron Byrne, Congxiao Bao, Dan Wing, Erik Kline, Francis Dupont, Gabor Bajko, Gang Chen, Hui Deng, Jan Zorz, Jacni Quin (who was an active co-author of some earlier versions of this specification), James Huang, Jari Arkko, Laurent Toutain, Leaf Yeh, Lorenzo Colitti, Mark Townsley, Marcello Bagnulo, Mohamed Boucadair, Nejc Skoberne, Olaf Maennel, Ole Troan, Olivier Vautrin, Peng Wu, Qiong Sun, Rajiv Asati, Ralph Droms, Randy Bush, Satoru Matsushima, Simon Perreault, Stuart Cheshire, Teemu Savolainen, Tetsuya Murakami, Tomasz Mrugalski, Tina Tsou, Tomasz Mrugalski, Washam Fan, Wojciech Dec, Xiaohong Deng, Xing Li, Yu Fu.

9. References

9.1. Normative References

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

[RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.

[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.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS

Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

[RFC2983] Black, D., "Differentiated Services and Tunnels", RFC 2983, October 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.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

[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.

[RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, November 2010.

9.2. Informative References

[I-D.ietf-pcp-base]
Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-19 (work in progress), December 2011.

[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.

[I-D.shirasaki-nat444]
Yamagata, I., Shirasaki, Y., Nakagawa, A., Yamaguchi, J., and H. Ashida, "NAT444", draft-shirasaki-nat444-04 (work in progress), July 2011.

[RFC0768] Postel, J., "User Datagram Protocol", STD 6, RFC 768, August 1980.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, November 1990.

[RFC1393] Malkin, G., "Traceroute Using an IP Option", RFC 1393, January 1993.

[RFC1631] Egevang, K. and P. Francis, "The IP Network Address Translator (NAT)", RFC 1631, May 1994.

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

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

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.

[RFC3828] Larzon, L-A., Degermark, M., Pink, S., Jonsson, L-E., and G. Fairhurst, "The Lightweight User Datagram Protocol (UDP-Lite)", RFC 3828, July 2004.

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

[RFC4821] Mathis, M. and J. Heffner, "Packetization Layer Path MTU Discovery", RFC 4821, March 2007.

[RFC4961] Wing, D., "Symmetric RTP / RTP Control Protocol (RTCP)", BCP 131, RFC 4961, July 2007.

[RFC5595] Fairhurst, G., "The Datagram Congestion Control Protocol (DCCP) Service Codes", RFC 5595, September 2009.

[RFC5969] Townsley, W. and O. Troan, "IPv6 Rapid Deployment on IPv4 Infrastructures (6rd) -- Protocol Specification", RFC 5969, August 2010.

[RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.

[RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", RFC 6146, April 2011.

[RFC6324] Nakibly, G. and F. Templin, "Routing Loop Attack Using IPv6 Automatic Tunnels: Problem Statement and Proposed

Mitigations", RFC 6324, August 2011.

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

[RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, November 2011.

[RFC6535] Huang, B., Deng, H., and T. Savolainen, "Dual-Stack Hosts Using "Bump-in-the-Host" (BIH)", RFC 6535, February 2012.

Appendix A. Textual representation of Mapping rules

In the next sections, each Mapping rule will be represented as follows, using 0bXXX to represent binary number XXX, and square brackets [ ] for what is optional:

{Rule IPv4 prefix, EA-bits length, Rule IPv6 prefix [, WKPs authorized]}

EXAMPLES:

```
{0.0.0.0/0, 32, 2001:db8:0:1:300::/80}
                          a BR mapping rule
{198.16.0.0/14, 22, 2001:db8:4000::/34}
                          a CE mapping rule
{0.0.0.0/0, 32, 2001:db8:0:1::/80}
                          a NAT64+ mapping rule)
{198.16.0.0/14, 22, 2001:db8:4000::/34, Yes}
                          a CE mapping rule and Hub&spoke Topology
```

Appendix B. Configuring multiple Mapping Rules

As far as mapping rules are concerned, the simplest deployment model is that in which the Domain has only one rule (the BR mapping rule). To assign an IPv4 address to a CE in this model, an IPv6 /112 is assigned to it comprising the BR /64 prefix, the 4rd tag, and the IPv4 address. This model has however the following limitations: (1) shared IPv4 addresses are not supported; (2) IPv6 prefixes used for 4rd are too long to be used also for native IPv6 addresses; (3) if the IPv4 address space of the ISP is split with many disjoint IPv4 prefixes, the IPv6 routing plan must be as complex as an IPv4 routing plan based on these prefixes.

With more mapping rules, CE prefixes used for 4rd can be those used for native IPv6. How to choose CE mapping rules for a particular deployment needs not being standardized.

The following is only a particular pragmatic approach that can be used for various deployment scenarios. It is used in some of the use cases that follow.

(1) Select a "Common_IPv6_prefix" that will appear at the beginning of all 4rd CE IPv6 prefixes.

(2) Choose all IPv4 prefixes to be used, and assign one of them to each CE mapping rule i.

(3) For each CE mapping rule i, do the following:

A. choose the length of its Rule IPv6 prefix (possibly the same for all CE mapping rules).

B. Determine its PSID_length(i). A CE mapping rule that assigns shared addresses with a sharing ratio 2^Ki, has PSID_length = Ki. A CE mapping rule rule that assigns IPv4 prefixes of length L < 32, is considered to have a negative PSID_length = L - 32.

C. Derive EA-bits length (i) = 32 - L(Rule IPv4 prefix(i)) + PSID_length(i).

D. Derive the length of Rule_code(i), the prefix to be appended to the Common prefix to get the Rule IPv6 prefix of rule i:

```
L(Rule_code(i)) = L(CE IPv6 prefix(i))
                - L(Common_IPv6_prefix]
                - (32 - L(Rule IPv4 prefix(i)))
                - PSID_length(i)
```

E. Derive Rule_code(i) with the following constraints: (1) its length is L(Rule_code(i); it does not overlap with any of the previously obtained Rule codes (for instance, 010, and 01011 do overlap, while 00, 011, and 010 do not); it has the lowest possible value as a fractional binary number (for instance, 0100 < 10 < 11011 < 111). Thus, rules whose Rule_code lengths are 4, 3 , 5, and 2, give Rule_codes 0000, 001, 00010, and 01)

F. Take Rule IPv6 prefix(i)= the Common_IPv6_prefix followed by Rule_code(i).

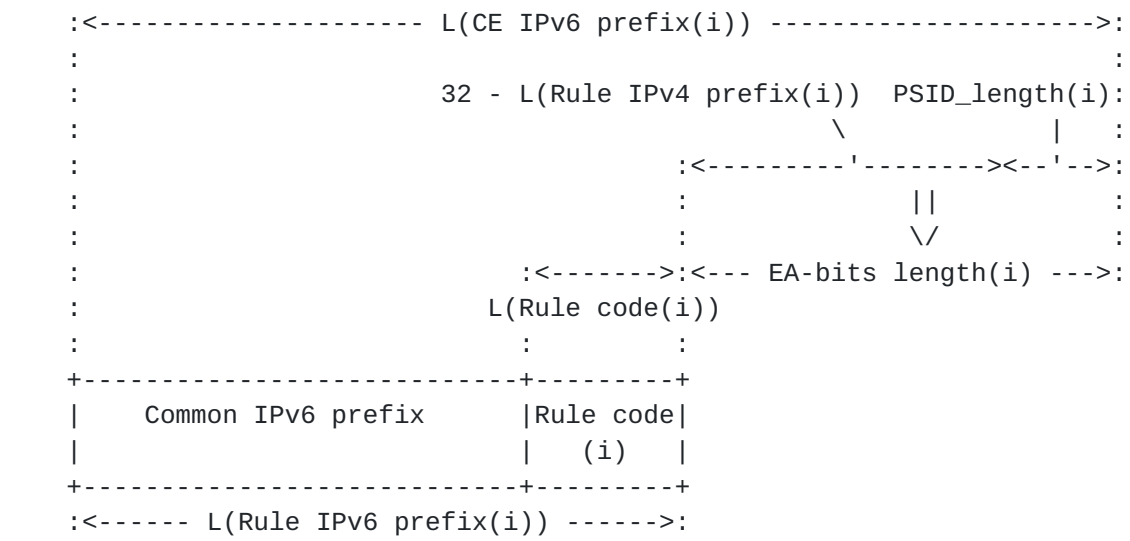


Figure 9

Appendix C. ADDING SHARED IPv4 ADDRESSES TO AN IPv6 NETWORK

C.1. With CEs within CPEs

We consider an ISP that offers IPv6-only service to up to 2^20 customers. Each customer is delegated a /56, starting with common prefix 2001:db8:0::/36. It wants to add public IPv4 service to customers that are 4rd-capable. It prefers to do it with stateless operation in its nodes, but has largely less IPv4 addresses than IPv6 addresses so that a sharing ratio is necessary.

The only IPv4 prefixes it can use are 192.8.0.0/15, 192.4.0.0/16, 192.2.0.0/16, and 192.1.0.0/16 (neither overlapping nor aggregetable). This gives 2^(32-15) + 3*2^(32-16) IPv4 addresses, i.e. 2^18 + 2^16). For the 2^20 customers to have the same sharing ratio, the number of IPv4 addresses to be shared has to be a power of 2. The ISP can therefore renounce to use one /16, say the last one. (Whether it could be motivated to return it to its Internet Registry is off-scope for this document.) The sharing ratio to apply is then 2^20 / 2^18 = 2^2 = 4, giving a PSID length of 2.

Applying principles of Appendix B with L[Common IPv6 prefix] = 36, L[PSID] = 2 for all rules, and L[CE IPv6 prefix(i)] = 56 for all rules, Rule codes and Rule IPv6 prefixes are:

CE Rule IPv4 prefix	EA bits length	Rule-Code length	Code (binary)	CE Rule IPv6 prefix
192.8.0.0/15	19	1	0	2001:db8:0::/37
192.4.0.0/16	18	2	10	2001:db8:800::/38
192.2.0.0/16	18	2	11	2001:db8:c00::/38

Mapping rules are then the following:

```
{192.8.0.0/15, 19, 2001:0db8:0000::/37}
{192.4.0.0/16, 18, 2001:0db8:0800::/38}
{192.2.0.0/16, 18, 2001:0db8:0c00::/38}
{0.0.0.0/0,    32, 2001:0db8:0000:0001:300::/80}
```

The CE whose IPv6 prefix is, for example, 2001:db8:0bbb:bb00::/56, derives its IPv4 address and its port set as follows (Section 4.4):

```
CE IPv6 prefix     : 2001:0db8:0bbb:bb00::/56
Rule IPv6 prefix(i): 2001:0db8:0800::/38 (longest match)
EA-bits length(i)  : 18
EA bits            : 0b11 1011 1011 1011 1011
Rule IPv4 prefix(i): 0b1100 0000 0000 0100 (192.4.0.0/16)
IPv4 address       : 0b1100 0000 0000 0100 1110 1110 1110 1110
                   : 192.4.238.238
PSID               : 0b11
Ports              : 0bYYYY 11XX XXXX XXXX
                     with YYYY > 0, and X...X any value
```

An IPv4 packet sent to address 192.4.238.238 and port 7777 is tunneled to the IPv6 address obtained as follows (Section 4.5):

```
IPv4 address       : 192.4.238.238 (0xC004 EEEE)
                   : 0b1100 0000 0000 0100 1110 1110 1110 1110
Rule IPv4 prefix(i): 192.4.0.0/16  (longest match)
                   : 0b1100 0000 0000 0100
IPv4 suffix (i)    : 0b1110 1110 1110 1110
EA-bits length (i) : 18
PSID length (i)    : 2  (= 16 + 18 - 32)
Port field         : 0b 0001 1110 0110 0001 (7777)
PSID               : 0b11
Rule IPv6 prefix(i): 2001:0db8:0800::/38
CE IPv6 prefix     : 2001:0db8:0bbb:bb00::/56
IPv6 address       : 2001:0db8:0bbb:bb00:300:c004:eeee:YYYY
                     with YYYY = the computed CNP
```

C.2. With some CEs behind Third-party Router CPEs

We now consider an ISP that has the same need as in the previous section except that, instead of using only its own IPv6 infrastructure, it uses that of a third-party provider, and that some of its customers use CPEs of this provider to use specific services it offers. In these CPEs, a non-zero index is used to route IPv6 packets to the physical port to which CEs are attached, say 0x2. Each such CPE delegates to the CE nodes the customer-site IPv6 prefix followed by this index.

The ISP is supposed to have the same IPv4 prefixes as in the previous use case, 192.8.0.0/15, 192.4.0.0/16, and 192.2.0.0/16, and to use the same Common IPv6 prefix, 2001:db8:0::/36.

We also assume that only a minority of customers use third-party CPEs, so that it is sufficient to use only one of the two /16s for them.

Mapping rules, are then (see Appendix C.1):

```
{192.8.0.0/15, 19, 2001:0db8:0000::/37}
{192.4.0.0/16, 18, 2001:0db8:0800::/38}
{192.2.0.0/16, 18, 2001:0db8:0c00::/38}
{0.0.0.0/0,    32, 2001:0db8:0000:0001:300::/80}
```

CEs that are behind third-party CPEs derive their own IPv4 addresses and port sets as in Appendix C.1.

In a BR, and also in a CE if the topology is mesh, the IPv6 address that is derived from IPv4 address 192.4.238.238 and port 7777 is obtained as in the previous section, except for the two last steps which are modified:

```
  IPv4 address       : 192.4.238.238 (0xC004 EEEE)
                     : 0b1100 0000 0000 0100 1110 1110 1110 1110
  Rule IPv4 prefix(i): 192.4.0.0/16  (longest match)
                     : 0b1100 0000 0000 0100
  IPv4 suffix (i)    : 0b1110 1110 1110 1110
  EA-bits length (i) : 18
  PSID length (i)    : 2  (= 16 + 18 - 32)
  Port field         : 0b 0001 1110 0110 0001 (7777)
  PSID               : 0b11
  Rule IPv6 prefix(i): 2001:0db8:0800::/38
  CE IPv6 prefix       : 2001:0db8:0bbb:bb00::/60
  IPv6 address         : 2001:0db8:0bbb:bb00:300:192.4.238.238:YYYY
                         with YYYY = the computed CNP
```

Appendix D. REPLACING DUAL-STACK ROUTING BY IPv6-ONLY ROUTING

In this use case, we consider an ISP that offers IPv4 service with public addresses individually assigned to its customers. It also offers IPv6 service, having deployed for this dual-stack routing. Because it provides its own CPEs to customers, it can upgrade all its CPEs to support 4rd. It wishes to take advantage of this capability to replace dual-stack routing by IPv6-only routing without changing any IPv4 address or IPv6 prefix.

For this, the ISP can use the single-rule model described at the beginning of Appendix B. If the prefix routed to BRs is chosen to start with 2001:db8:0:1::/64, this rule is:

{0.0.0.0/0, 32, 2001:db8:0:1:300::/80}

All what is needed in the network before disabling IPv4 routing is the following:

- In all routers, where there is an IPv4 route toward x.x.x.x/n, add a parallel route toward 2001:db8:0:1:300:x.x.x.x::/(80+n)

- Where IPv4 address x.x.x.x was assigned to a CPE, now delegate IPv6 prefix 2001:db8:0:1:300:x.x.x.x::/112.

NOTE: In parallel with this deployment, or after it, shared IPv4 addresses can be assigned to IPv6 customers. It is sufficient that IPv4 prefixes used for this be different from those used for exclusive-address assignments. Under this constraint, Mapping rules can be set up according to the same principles as those of Appendix C.

Appendix E. ADDING IPv6 AND 4rd SERVICE TO A NET-10 NETWORK

In this use case, we consider an ISP that has only deployed IPv4, possibly because some of its network devices are not yet IPv6 capable. Because it did not have enough IPv4 addresses, it has assigned private IPv4 addresses of [RFC1918] to customers, say 10.x.x.x. It thus supports up to 2^24 customers (a "Net-10" network, using the NAT444 model of [I-D.shirasaki-nat444]).

Now, it wishes to offer IPv6 service without further delay, using for this 6rd [RFC5969]. It also wishes to offer incoming IPv4 connectivity to its customers with a simpler solution than that of PCP [I-D.ietf-pcp-base].

This appendix describes an example that adds IPv6 (using 6rd) and 4rd services to the "Net-10" private IPv4 network.

The IPv6 prefix to be used for 6rd is supposed to be 2001:db8::/32, and the public IPv4 prefix to be used for shared addresses is supposed to be 198.16.0.0/16 (0xc610). The resulting sharing ratio is 2^24 / 2^(32-16) = 256, giving a PSID length of 8.

The ISP installs one or several BRs, at its border to the public IPv4 Internet. They support 6rd, and 4rd above it. The BR prefix /64 is supposed to be that which is derived from IPv4 address 10.0.0.1 (i.e. 2001:db8:0:100:/64).

In accordance with [RFC5969], 6rd BRs are configured with the following parameters IPv4MaskLen = 8, 6rdPrefix = 2001:db8::/32; 6rdBRIPv4Address = 192.168.0.1 (0xC0A80001).

4rd Mapping rules are then the following:

```
{198.16.0.0/16, 24, 2001:db8:0:0:300::/80}
{0.0.0.0/0,      32, 2001:db8:0:100:300:/80,}
```

Any customer device that supports 4rd in addition to 6rd can then use its assigned shared IPv4 address with 240 assigned ports.

If its NAT44 supports port forwarding to provide incoming IPv4 connectivity (statically, or dynamically with UPnP an/or NAT-PMP), it can use it with ports of the assigned port set (a possibility that does not exist in Net-10 networks without 4rd/6rd).

Authors' Addresses

Remi Despres
RD-IPtech
3 rue du President Wilson
Levallois,
France

Email: despres.remi@laposte.net

Sheng Jiang (editor)
Huawei Technologies Co., Ltd
Q14, Huawei Campus, No.156 BeiQing Road
Hai-Dian District, Beijing, 100095
P.R. China

Email: jiangsheng@huawei.com

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

Email: repenno@cisco.com

Yiu Lee
Comcast
One Comcast Center
Philadelphia, PA 1903
USA

Email: Yiu_Lee@Cable.Comcast.com

Gang Chen
China Mobile
53A, Xibianmennei Ave.
Xuanwu District, Beijing 100053
China

Email: phdgang@gmail.com

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

Email: fibrib@gmail.com