Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: July 31, 2012

R. Despres
RD-IPtech

January 28, 2012

IPv4 Residual Deployment via IPv6 - Unified Solution (4rd-U)
draft-despres-softwire-4rd-u-03

Abstract

This document specifies an automatic tunneling mechanism tailored for residual deployment of public IPv4 via IPv6 networks (the reverse of 6rd whose purpose is rapid deployment of IPv6 via IPv4). In order to deal with the IPv4-address shortage, customers can be assigned shared IPv4 addresses with statically assigned restricted port sets. Operation is stateless, with neither per-connection nor per-customer state needed in network nodes.

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 July 31, 2012.

Table of Contents

1. Introduction

This specification addresses the need for a stateless solution permitting deployments of residual IPv4 service via IPv6 networks, as expressed in [I-D.ietf-softwire-stateless-4v6-motivation]. The solution, named 4rd for IPv4 residual deployment, needs neither per-connection nor per-customer states in network nodes. With it, IPv4 packets are tunneled across IPv6 networks in a design reverse of that of 6rd [RFC5969] whereby IPv6 packets are tunneled across IPv4 networks.

In order to deal with the IPv4-address shortage, customers can be assigned shared IPv4 addresses with statically assigned restricted port sets.

The design of 4rd builds on a number of previous proposals made for IPv4-via-IPv6 transition technologies listed in Section 9. It includes in a common framework two formats of tunnel packets. The Header-mapping variant is similar to a double-translation solution based on the IPv6/IPv4 translation of [RFC6145]. Like them, it permits middle boxes to operate on tunneled IPv4 packets as though they would be native IPv6 packets, in particular for port-based access control lists, and for website redirects. In addition to these, it preserves complete network transparency to IPv4 packets, and is self contained. The Encapsulation variant does not have this middle-box compatibility but is algorithmically simpler.

Terminology is defined in Section 2. How the 4rd model fits in the Internet architecture is summarized in Section 3. The protocol specification is detailed in Section 4. Section 5 illustrates a few typical 4rd use cases. Section 6 and Section 7 respectively deal with security and IANA considerations. Previous proposals that influenced this specification are listed in Section 9.

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].

2. Terminology

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 with restricted port sets assigned to customers.

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

BR (Border Relay): A 4rd-capable node at the border between a 4rd domain and the IPv4 Internet. Because its operation is stateless, it can be replicated in as many instances as needed for scalability.

CE (Customer Edge node): A 4rd-capable customer node attached to a 4rd domain. It can be a host, a router, or both.

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

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

4rd address: An IPv4 address or, in case of a shared IPv4 address, an IPv4 address plus a port number.

Tunnel packet: An IPv6 packet that conveys an IPv4 packet across a 4rd domain. Its payload is either the original IPv4 payload (Header-mapping variant), or the complete IPv4 packet (Encapsulation variant).

Mapping rule: A set of parameters that BRs and CEs use to derive IPv6 addresses from 4rd addresses. They are also used by CEs to derive their own 4rd prefixes from their IPv6 delegated prefixes.

EA bits (Embedded Address bits): Bits that, in prefixes and addresses, are the same in 4rd and in IPv6.

Default mapping rule: The mapping rule that applies to off-domain IPv4 addresses.

CNP (Checksum Neutrality preserver): A field that, in the Header-mapping variant, ensures that contributions to UDP/TCP-like checksums of the transport layer remain valid despite replacements of IPv4 addresses by IPv6 addresses.

3. The 4rd Model

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

A CE that is assigned a shared IPv4 address, or that that is assigned a single addresses and acts as router, MUST have a NAT44 function (IPv4 NAT of [RFC1631]). This NAT44 MUST only use external ports that belong to its assigned port set. It MUST treat external Identifications of Echo requests it sends as though they would be port numbers. and as Identifications of Echo requests it sends across the Domain. It MUST also avoid sending interleaved fragments of several datagrams (Section 4.5.3).

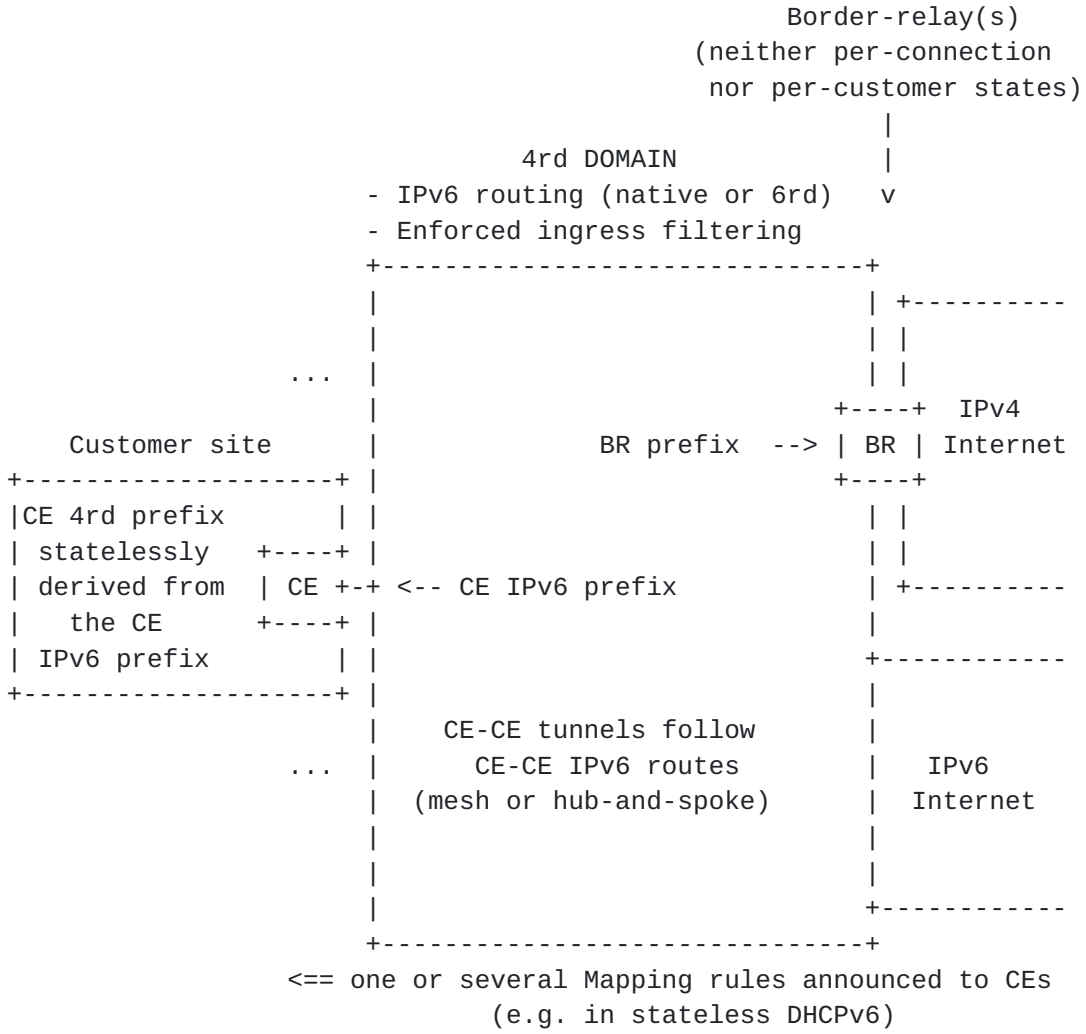


Figure 1

4. Protocol Specification

4.1. 4rd Domain parameters

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

- Header variant (Header mapping or Encapsulation)
- Topology variant (Mesh or Hub&Spoke)
- Domain PMTU
- Tunnel traffic class (optional)
- Domain IPv6 suffix (optional)
- Mapping rules, each one comprising:
  - Rule IPv4 prefix
  - EA-bits length
  - Rule IPv6 prefix

In a Domain that supports shared-address CEs and has duplicated BRs, each BR MUST also have the following parameter:

- BR packet-ID prefix

These parameters are REQUIRED to comply with the following:

R-1 "Header variant" determines whether Tunnel packets are built according to the Header-mapping variant or the Encapsulation variant (Section 4.2 to Section 4.4).

R-2 "Topology variant" determines whether CEs address their Tunnel packets directly to IPv6 addresses of other CEs (Mesh variant), or whether they always address them to BR IPv6 addresses (Hub& spoke variant).

R-3 "Domain PMTU" is the IPv6 path MTU that the ISP can guarantee for all its cross-domain paths. In accordance with [RFC2460], it MUST be at least 1280.

R-4 "Tunnel traffic class", if provided, is the IPv6 traffic class that BRs and CEs MUST set in Tunnel packets. In this case, tunnel traversal is treated in IPv4 as a single-link traversal. Without it, Explicit Congestion Notification of [RFC3168] MAY be propagated from intermediate IPv6 nodes to IPv4 destinations, and IPv4 Time to live values progress with the number of traversed IPv6 links.

R-5 "Domain IPv6 suffix", which is optional, is only used in particular Domains where CEs are placed in customer sites behind third-party CPEs, and where these CPEs use some address bits to route packets among their physical ports. Its effect is detailed in Section 4.3.1 and Section 4.4.1. A use case where it applies is presented in Section 5.2.2.

R-6 "BR datagram-ID prefix" is used to ensure that datagrams that go from CEs that share the same IPv4 address to a common destination via the IPv4 Internet have datagram Identifications that cannot be confused. If the Domain supports shared addresses, all BRs MUST have different values of this parameter, as detailed in Section 4.5.3. (If there is only one BR, or if the Domain supports only non-shared IPv4 addresses, this parameter is not needed, it can be by default a /0 to be ineffective.)

R-7 "Rule IPv4 prefix" is used to find, with a longest match, which Mapping rule applies to a 4rd address. All Mapping rules MUST have different Rule IPv4 prefixes.

R-8 "EA-bits length" of a Mapping rule specifies the number of bits of 4rd addresses that, if this rule applies, are embedded in IPv6 addresses. A rule that has the length of its Rule IPv4 prefix plus its EA-bits length larger than 32, is one that applies to shared-address CEs. The number of bits beyond 32, is the PSID length of the rule. It, determines its sharing ratio.

R-9 "Rule IPv6 prefix" of a Mapping rule is the prefix that, in IPv6 addresses derived from 4rd addresses with this rule, replaces the Rule IPv4 prefix. All Mapping rules MUST have different Rule IPv4 prefixes.

R-10 Each Domain MUST have one and only one "Default mapping rule". This rule MUST have Rule IPv4 prefix = 0.0.0/0, EA-bits length = 32. is Rule IPv6 prefix MUST be a /80 whose format is specified in Section 4.4.2.

R-11 Rules other than the Default mapping rule MUST concern CE IPv6 prefixes that can be the same for native IPv6 and for 4rd, i.e. are at most /64s. (For length of the Rule IPv6-prefix plus EA-bits length, plus length of the Domain-IPv6-suffix if there is one MUST NOT exceed 64).

R-12 Each CE and each BR MUST be capable to support up to 32 Mapping rules. (Note: this number, which is not critical, is easy to change, should a working-group consensus require another value.) ISPs that need Mapping rules for more than 32 IPv4 prefixes SHOULD split their networks into multiple Domains, with Domain to Domain communication driven by IPv4 addresses.

R-13 ISPs that provide their own CPEs to all their customers MAY limit CE functions of these CPEs to those that are needed for parameter values they choose to deploy. (For example, they MAY support only an Encapsulation and Mesh variant, or only a Header-mapping and Hub&spoke variant, as applicable to their deployment choice).

4.2. Headers of Encapsulation and Header-Mapping Variants

R-14 Domain-entry nodes MUST discard IPv4 packets they receive with one or several IPv4 options. (They MUST then return the ICMPv4 error message of [RFC0792] that signals such an IPv4-option incompatibility: Type = 12, Code = 0, Pointer = 20). Note: This limitation is made to privilege simplicity, and taking in consideration that IPv4 options, which are very rarely used, are not necessary for normal IPv4 operation.

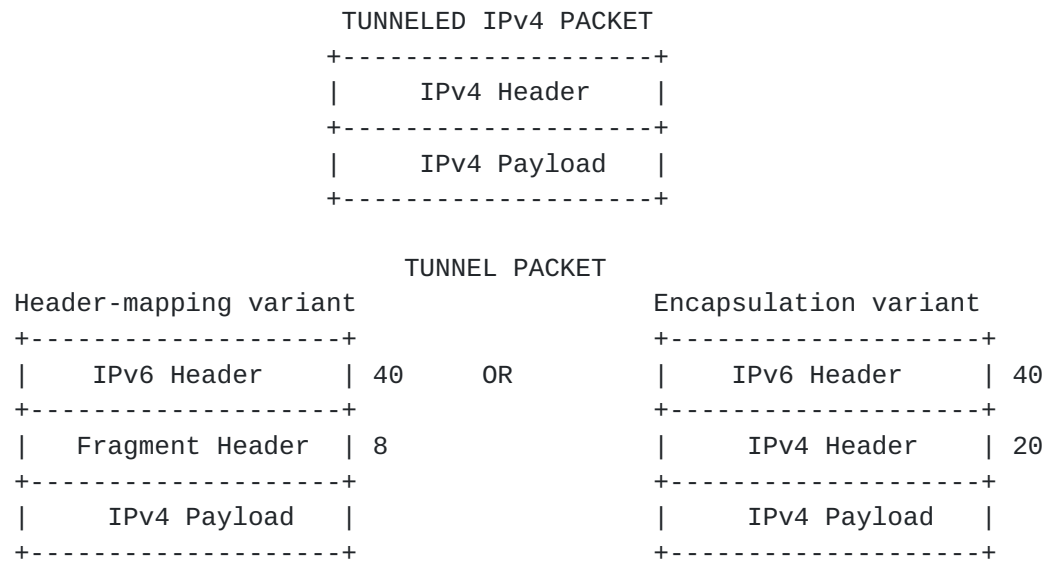


Figure 2

R-15 Domain-entry nodes MUST convert IPv4 packets they receive without any IPv4 option into Tunnel packets. Domain-exit nodes MUST convert back IPv6 packets into IPv4 packets. Headers of Tunnel packets have the two variants shown in (Figure 2): in the Header-mapping variant, the tunnel header comprises an IPv6 header followed by an IPv6 fragment header (total 48 octets); in the Encapsulation variant, it comprises an IPv6 header followed by an IPv4 header (total 60 octets).

R-16 Domain-entry nodes MUST discard IPv4 packets received with one or several IPv4 options (and return ICMPv4 error message of [RFC0792] that signals such an IPv4-option incompatibility: Type = 12, Code = 0, Pointer = 20). Note: this limitation is introduced to privilege simplicity, and in view that IPv4 options, very rarely used, are unnecessary for normal IPv4 operation.

R-17 In the Header-mapping variant, the IPv6 fragment header is used to convey fields of IPv4 headers that have no equivalents in IPv6 headers but have equivalents in IPv6 fragment headers, namely the More-fragment bit M and Offset.

* The IPv6 fragment header MUST also be used to convey the DF bit, an IPv4 field that has no equivalent in IPv6 where the Don't-fragment condition holds by default [RFC2460].

* If the Domain has a configured Tunnel Traffic class, the IPv6 fragment header MUST also be used to convey the original IPv4 TOS so that it can be restored at Domain exit.

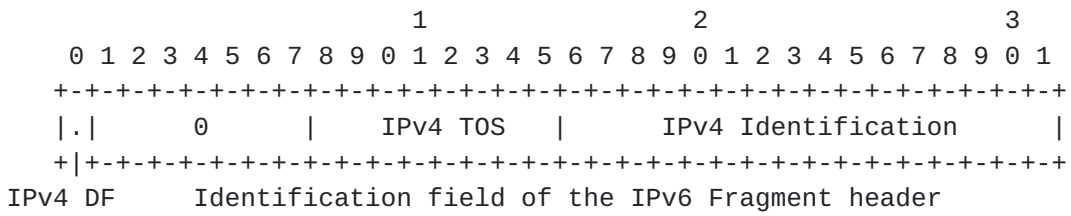


Figure 3

* For this to be possible, it is taken advantage of the fact that the Identification field of an IPv6 fragment headers has 32 bits while 16 bits are enough to contain a copy of that of an IPv4 header. Copies of the IPv4 DF bit and TOS MUST be placed as shown in Figure 3.

IPv6 FIELDS	VALUES SET AT DOMAIN ENTRYS
Version	6
Traffic class	TOS or parameter - see Section 4.6
Flow label	0
Payload length	Total length - 12
Next header	44 (Fragment header)
Hop limit	Time to live
Source address	See Section 4.4.1 & Section 4.4.2
Dest. address	See Section 4.4.1 & Section 4.4.2
2nd next header	Protocol
Frag. offset	Frag. offset
M	More fragments (MF)
IPv4 DF	Don't fragment (DF)
IPv4 TOS	Type of service (TOS)
IPv4 Identification	Identification

Table 1

IPv4 FIELDS	VALUES SET AT DOMAIN EXIT
Version	4
Header length	5
TOS	See Section 4.6
Total Length	Payload length + 12
DF	IPv4 DF
MF	M
Fragment offset	Fragment offset
Time to live	Hop limit
Protocol	2nd Next header
Header checksum	Computed as per [RFC0791]
Source address	Bits 80-11 of source address
Destination address	Bits 80-11 of destination address

Table 2

* Other than that, field values that Domain-entry nodes MUST set in Tunnel-packet headers are straightforward. For reference, they are detailed in Table 1. Those that Domain-exit nodes MUST set in restored IPv4 headers are detailed in Table 2.

R-18 In the Encapsulation variant, Domain-entry nodes MUST add an IPv6 header in front of IPv4 packets they have to forward. Its Next header MUST be set to 4 according to [RFC2003]. Domain-exit nodes MUST decapsulate IPv4 packets. If the Domain has no Tunnel traffic class parameter,they MUST replace TOS values of decapsulated packets by Traffic-class values received in IPv6 headers.

4.3. From CE IPv6 Prefixes to IPv4 Addresses and Port sets

4.3.1. From CE IPv6 Prefix to CE 4rd Prefix

R-19 Each CE MUST derive its own 4rd prefix from its delegated IPv6 prefix as detailed in Figure 4. The first step MUST consist in finding the Mapping rule whose IPv6 prefix has the the longest match with the CE delegated prefix. If none is found, the IPv6 prefix is not one of 4rd. Another IPv6 prefix can be tried if the CE has been delegated are several. If still no rule is found, it is a sign that this CE is left out of 4rd service in this Domain. If a rule is found, the CE MUST replace the Rule IPv6 prefix by the Rule IPv4 prefix. It the Domain has a Domain IPv6 suffix, the CE MUST truncate the result by deleting its last k bits, where k is the Domain-IPv6-suffix length. The result is the CE 4rd prefix.

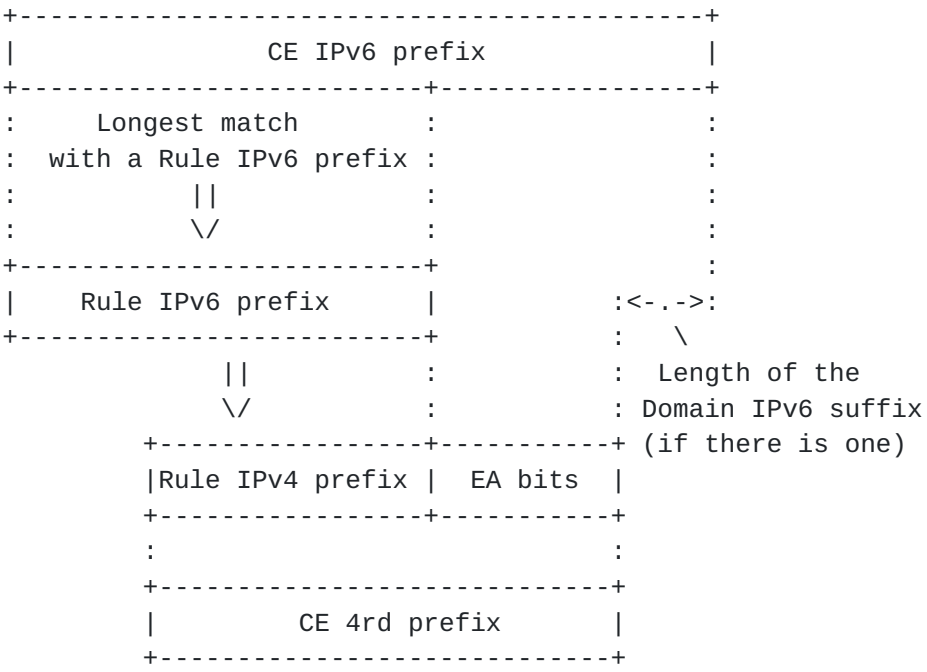


Figure 4

R-20 If this CE 4rd prefix is longer than a /32, the CE MUST take its bits beyond the first 32 as the CE PSID. Ports of the PSID-specified port set MUST then be derived as specified in Section 4.3.2). If the 4rd prefix is a /32, the CE MUST take the 4rd prefix. If it is shorter than a /32, the CE MUST take it as its IPv4 prefix.

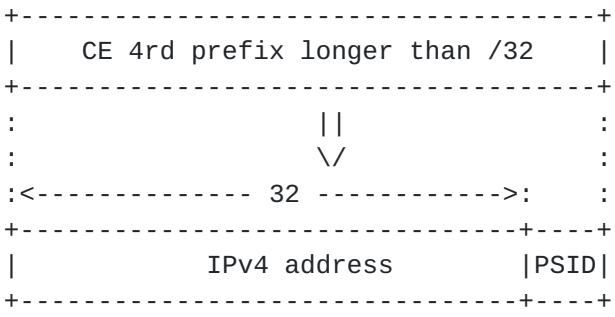


Figure 5

4.3.2. From PSID to Port Set

R-21 Each CE that has a PSID as a result of Section 4.3.1 MUST derive ports of its restricted as shown in Figure 6: non-zero value in their first 4 bits, followed by the PSID, and any value in the remaining bits.

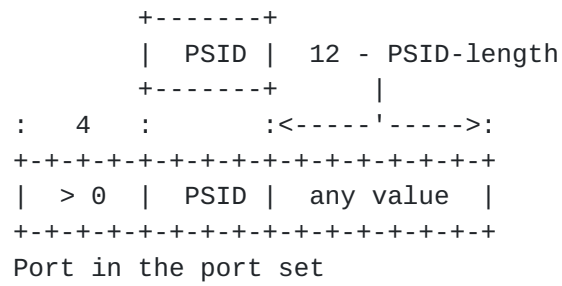
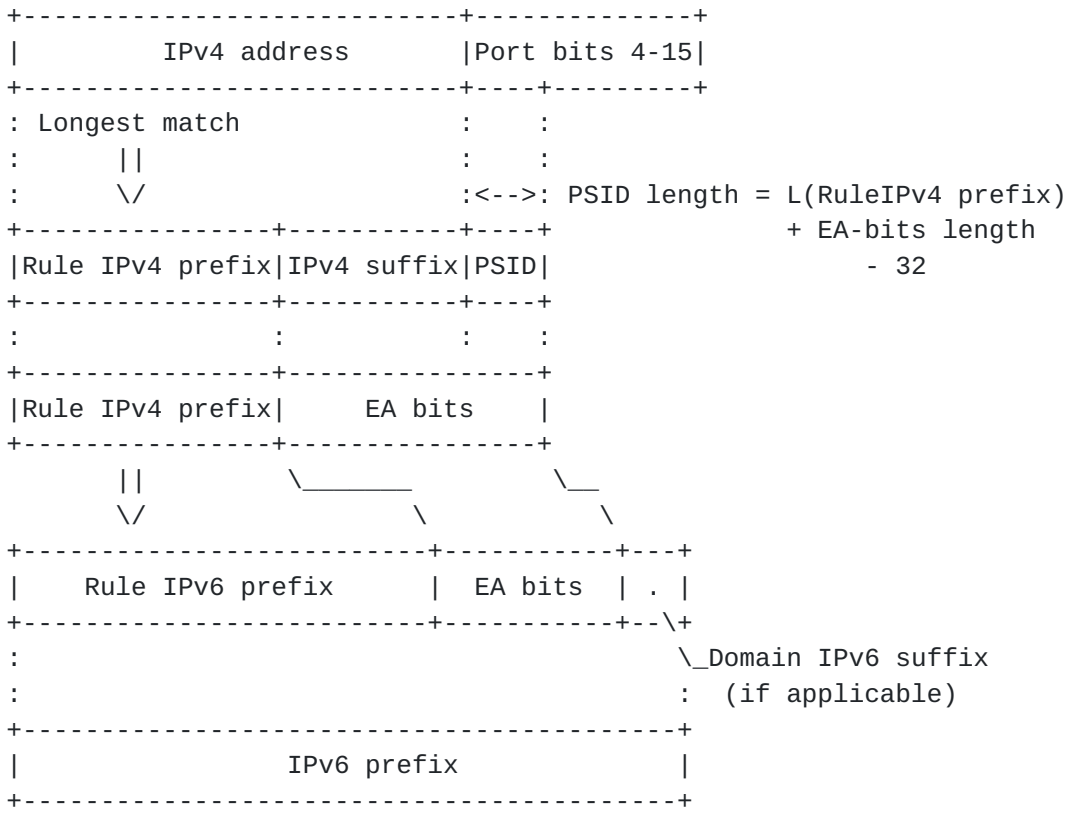


Figure 6

NOTE: The choice of the 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 (these ports have more value than others); (2) for compatibility RTP/RTCP [RFC4961], port sets need to include pairs of consecutive ports (for this, PSID length MUST be limited to 11, a negligible constraint in practice since a sharing ratio of 2048 is already beyond realistic needs); (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. With the first 4 port bits required to be > 0, excluded ports are 0-4095, i.e. more than the required 0-1023. This is a trade-off in view of other objectives, in particular nibble alignment.

4.4. From IPv4 addresses and Ports to IPv6 Addresses

4.4.1. From 4rd Address to IPv6 Prefix



From 4rd address to IPv6 prefix (shared-address case)

Figure 7

R-22 BRs, and CEs in the Mesh variant, MUST derive IPv6 addresses from a 4rd addresses with the following steps (or their functional equivalent):

(1) Find the Mapping rule whose Rule IPv4 prefix has the longest match with the IPv4 address.

(2) If the Rule IPv4 prefix plus EA-bits length of this rule is 32 + k with a positive k, append to the IPv4 address the PSID found in bits 4 to 4+k of the port field.

(3) The port field to be used if a PSID is needed is found in the IPv4-packet payload at a location that depends on whether the address is a source or a destination address, on whether the packet is ICMP or not, and, if it is ICMP whether it is an error message or an echo message. Precise locations in IPv4 payloads are the following:

+ If the packet Protocol is not ICMP, bits 0-15 for an IPv4 source address, and bits 16-31 for a destination address.

+ If the packet is an ICMPv4 error message, bits 240-255 for a source address; bits 224-239 for a destination address.

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

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

(5) If the Domain has a Domain IPv6 suffix (see Section 4.1), and if the rule is not the Default mapping rule, append it to the result to obtain an IPv6 prefix. Steps up to this one are illustrated in Figure 7,

(6) Derive from this derived IPv6 prefix an IPv6 address according to Section 4.4.2.

R-23 In the Hub&spoke variant, CEs MUST always use the Default-mapping rule to derive IPv6 prefixes of Tunnel-packet destinations.

NOTE: Using Identification fields of ICMP messages as port fields permits to exchange Echo requests and Echo replies between shared-address CEs and and IPv4 hosts having exclusive IPv4 addresses. Echo exchanges between two shared-address CEs remain impossible. This limitation to is inherent to address sharing, independently of its being static or dynamic. Using IPv6 is the obvious way to avoid this limitation.

4.4.2. From IPv6 Prefix to IPv6 Address

R-24 An IPv6 prefix derived from a Mapping rule than is not the Default mapping rule is, as specified in Section 4.1, at most a /64. In this case, the IPv6 address MUST be completed with a null padding field up to 64 bits, the V octet (see below), an empty octet, the IPv4 address, and a CNP or null 16-bit field depending on whether the applicable variant is Header mapping or Encapsulation.

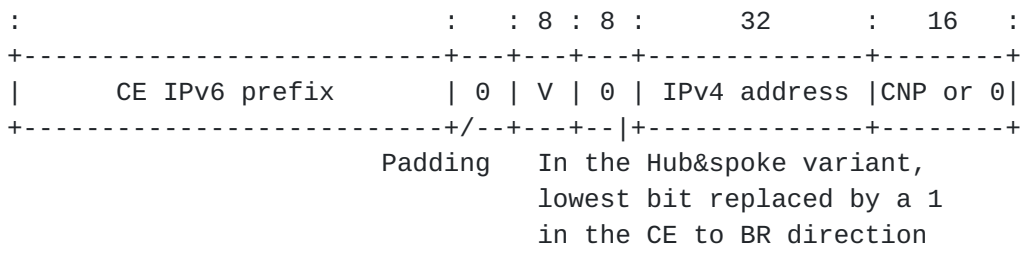


Figure 8

* The CNP is, in one's complement arithmetic, the opposite of the sum of the 5 first 16-bit words of the IPv6 address. This guarantees that, as far as UDP-like checksums of the transport layer are concerned, Tunnel packets are valid in IPv6. At this time, it applies to UDP, TCP and DCCP. It will also automatically apply to any protocols that may use the same checksums in the future (e.g. a SCTP with a simplified checksum, should it be found useful).

* The V octet is a 4rd-specific mark. Its function is to ensure that 4rd does not interfere with the choice of subnet prefixes in CE sites. It can also facilitate maintenance by facilitating distinction between 4rd Tunnel packets and native-IPv6 packets. Within CEs, IPv6 packets can safely be routed to the 4rd function based on a /80 prefix because no internal route for native IPv6 can have a destination prefix that start with this one. For this, the V octet MUST have its "u" and "g" bits of [RFC4291] both set to 1. This is REQUIRED to avoid "u" = 0 (reserved for local-scope

Interface IDs, in which case all other bits have any values), and to avoid "u" = 1 and "g"= 0 ("u" = 1 is reserved for Interface IDs using the EUI-64 format, in which case unicast addresses have "g" = 0). The proposed value of other bits of the octet is 0, giving V = 0x03. As indicated in Section 7, this value needs to be submitted to IANA.

* In the Hub&spoke variant, a precaution is needed so that not only off-domain IPv4 addresses are mapped with the Default mapping rule, but also some CE-assigned IPv4 addresses (use case of Section 5.3). For addresses destined to such CEs to be routed differently in the BR-to-CE direction and in the CE-to-BR direction, CEs MUST set bit 79 to 1 (the last one before copied IPv4 addresses). Routes to BRs of the Hub& spoke variant MUST therefore have 80-bit prefixes whose last bit is a 1.

R-25 An IPv6 prefix that is derived from a 4rd address with the Default mapping MUST, according to Section 4.1, be a /112 (80 bits of Rule IPv6 prefix plus 32 EA bits). In order to ensure that all Tunnel-packet addressed to BRs are as distinguishable from native IPv6 addresses as those addressed to CEs, the Rule IPv6 prefix of the Default mapping rule MUST have the V octet in bits 64-71. The following octet SHOULD be 0, with its last bit modified as above in the CE to BR direction of the Hub& spoke variant. Other than that, values of bit 71 and of the last last 16 bits are set as above for IPv6 prefixes limited to /64s (

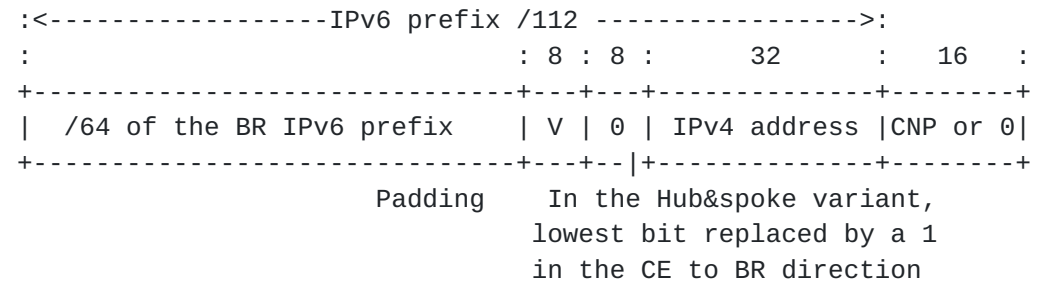


Figure 9

4.5. Fragmentation Considerations

4.5.1. General

R-26 If an IPv4 packet enters a CE or BR with a size such that the size of a directly derived Tunnel packet would exceed the Domain PMTU the packet has to be either fragmented or discarded. The Domain-entry node MUST discard it if it has DF = 1 (with an ICMP error message returned to the source). It MUST fragment it otherwise. (The length of each payload fragment MUST be at most the Domain PMTU - k, where k is 48 in the Header-mapping variant, and k is 60 in the Encapsulation variant.)

4.5.2. Ports of Fragments sent to Shared-Address CEs

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

For this, a BR could systematically reassemble fragmented IPv4 datagrams before tunneling them, but this consumes large memory space, opens denial-of-service-attack opportunities, and can significantly increase forwarding delays.

R-27 BRs SHOULD support an algorithm whereby it can forward IPv4 packets from the Internet on the fly, for example the following algorithm:

(1) At BR initialization, if at least one CE mapping rule concerns shared IPV4 addresses (length of Rule IPv4 prefix + EA-bits length > 32), the BR initializes an empty "IPv4-datagram 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 addresses (length of Rule IPv4 prefix + EA-bits length > 32), the the BR searches the table for an entry whose IPv4 source, IPv4 destination, and IPv4 Identification, are those of the received packet. It then performs actions detailed in Table 3 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	Y	N
An entry has been found	Y	N	Y	N	Y	N	Y	N
- RESULTING ACTIONS -								
Create a new entry	-	-	-	X	-	-	-	-
Use the port of the entry	-	-	-	-	X	-	-	-
Update port of the entry	-	-	X	-	-	-	X	-
Delete the entry	X	-	-	-	X	-	-	-
Forward the packet	X	X	X	X	X	-	X	-

Table 3

(3) BRs SHOULD operate garbage collection set up for table entries that remain unchanged for longer than a fragmented-datagram reception may take. This value, which is not critical, MAY be set to 15 seconds (ref [RFC0791].

R-28 CEs that are able to route public IPv4 addresses in their sites when assigned IPv4 prefixes (as opposed to always having NAT44s at site entrances), MUST have the same behavior as that described above for BRs.

4.5.3. Datagram Identifications from Shared-Address CEs

When datagrams go from different shared-address CEs to a common off-domain destination, there is no guarantee that packet Identification values set by sources are different. Because datagram reassembly in the destination is based only on source address and packet Identification, fragments of different sources can be confused. Probability of this happening may in theory be very low but, in order to avoid creating new attack opportunities, a safe solution is needed.

R-29 BRs SHOULD support an algorithm that ensures that datagram Identifications from shared address CEs never create reassembly ambiguity in common destinations, for example the following one:

(1) At BR initialization, if at least one CE mapping rule concerns shared IPV4 addresses (i.e. if the sum of its Rule-IPv4-prefix length and EA-bits length exceeds 32), the BR initializes an empty table and a "Datagram-ID generator" whose characteristics are specified below. Entries of the table have the following items:

- IPv6 prefix

- CE packet ID

- BR datagram ID

(2) When a BR has an IPv4 packet to forward, it determines whether it comes from a shared-address CE (length of Rule IPv4 prefix + EA-bits length > 32). If yes, the BR searches the table for an entry whose IPv6 prefix is equal to the source IPv6 prefix, and takes actions detailed in Table 4.

- CONDITIONS -										
First Fragment (offset=0)	Y	Y	Y	Y	N	N	N	N	N	N
Last fragment (MF=0)	Y	Y	N	N	Y	Y	Y	N	N	N
An entry is found	Y	N	Y	N	Y	Y	N	Y	Y	N
CE pkt ID = Entry pkt ID	?		?		Y	N		Y	N	
- RESULTING ACTIONS -										
Generate a new BR pkt ID	X	X	X	X	-	-	-	-	-	-
Create a new entry	-	-	-	X	-	-	-	-	-	-
Use BR pkt ID of entry	-	-	-	-	X	-	-	X	-	-
Update pkt IDs of entry	-	-	X	-	-	-	-	-	-	-
Delete the entry	X	-	-	-	X	X	-	-	X	-
Forward the packet	X	X	X	X	X	-	-	-	-	-

Table 4

(3) The "Datagram-ID generator" provides Identification values that, for each possible CE IPv4 address, must comply with what applies to hosts in [RFC0791] ("the sender must choose the Identifier to be unique for this source, destination pair and protocol for the time the datagram (or any fragment of it) could be alive in the internet"). Since each BR has its own Datagram-ID generator, BRs should generate Identifications belonging to disjoint sets. This is the reason for having as BR parameter a BR datagram-ID prefix (Section 4.1). Having this parameter, each BR generates Identifications that start with this prefix.

R-30 CEs that are able to route public IPv4 addresses in their sites when assigned IPv4 prefixes (as opposed to always having NAT44s at site entrances), MUST have the same behavior as that described above for BRs.

4.6. TOS and Traffic-Class Considerations

As specified in [RFC3168], the TOS of IPv4 headers and the Traffic class of IPv6 headers must have the same meanings in ECN-capable networks (networks supporting Explicit Congestion Notification). Their first 6 bits are a Differentiated Services CodePoint (DSCP). Their two last bits are an Explicit Congestion Notification (ECN). [RFC6040] details how the ECN field MAY evolve if a packet traverses a router that signals congestion condition before packets are dropped.

R-31 4rd domains MUST support the ECN normal mode of [RFC6040] by default. For this, BRs and CEs MUST copy the IPv4 TOS into the IPv6 Traffic class at Domain entry, and copy back the IPv6 Traffic class, which may have a changed ECN, into the IPv4 TOS at Domain exit.

R-32 A Domain where the ECN normal mode of [RFC6040] is not supported MUST have its Domain-traffic-class parameter set. In this case, BRs and CE's MUST take this parameter as IPv6 Traffic class of Tunnel packets, and MUST keep at Domain exit TOS values that were received at Domain entry. (In the Header-mapping variant, the TOS value is restored as detailed in Section 4.2.)

4.7. Tunnel-Generated ICMPv6 Error Messages

R-33 If an Tunnel packet is discarded on its way across a 4rd domain, possibly because of an unreachable destination, an ICMPv6 error message is returned to the IPv6 source, i.e. to the BR anycast address or to a CE address. For the source of the IPv4 packet to be informed of the packet loss, the ICMPv6 packet MUST be converted to an ICMP packet returned to the IPv4 source. Type, or Type and Code, of the ICMPv4 packet MUST be obtained as specified for error messages in Section 5.2 of [RFC6145]. The ICMP checksum MUST be updated accordingly.

R-34 The IPv4 source address to be used in these ICMP packets MUST be 192.70.192.254 (to be confirmed by IANA - see Section 7). (Note: this value taken in the /24 range proposed in [I-D.xli-behave-icmp-address] for a similar purpose.)

4.8. Provisioning 4rd Parameters to CEs

Parameter listed in Section 4.1 can be announced to CE's in DHCPv6 (ref. [RFC2131]).

At this stage, option formats remain to be defined, one for Domain parameters other than Mapping rules, say OPTION-4RD, and one for mapping rules, say OPTION_4RD_RULE.

5. Use-Case Examples

5.1. How to choose Mapping Rules

As far as mapping rules are concerned, the simplest deployment model is that in which the Domain has only one rule (the Default 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 V octet, a null octet, 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, the same CE prefixes can be used for 4rd and for native IPv6. Also, IPv6 prefixes that have been assigned without previous relationship with IPv4 prefixes can be used for to offer 4rd service. How to choose CE mapping rules for a particular deployment needs not being standardized, but some hints are possible. The following is a pragmatic approach that can be used for various deployment scenarios:

(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 for 4rd. For each one, choose its applicable sharing ratio. Choose the length of CE IPv6 prefixes to be used (in the simplest scenarios, the same for all Mapping rules).

(3) Derive from these data, and for each rule, the length of a "Rule code". This code is that which is appended to the Common prefix to get the Rule IPv6 prefix (Figure 10). For Mapping rule i, its length is L(Rule code(i) = L(CE IPv6 prefix(i)) - L(Common_IPv6_prefix) - (32 - L(Rule IPv4 prefix(i))) - L(PSID), where L(PSID) = k is the sharing ratio is 2^k:

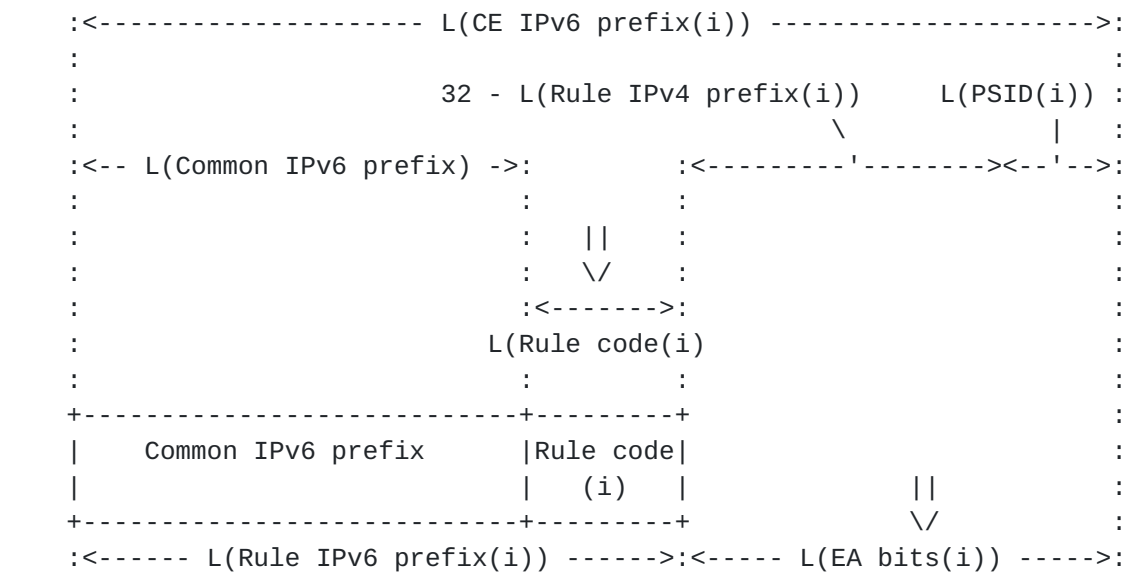


Figure 10

(4) For each rule successively, take as Rule code the prefix which has the obtained length, which does not overlap with previously chosen Rule codes, and which, to make a systematic choice, has the lowest value if interpreted as fractional part of a binary number. (For example with successive lengths 4, 3 , 5, and 2, this gives, in binary notation, 0000, 001, 00010, and 01)

For textual representation of CE mapping rules in the next sections, we will use {Rule IPv4 prefix, EA-bits length, Rule IPv6 prefix}. Example: {198.16.0.0/14, 22, 2001:db8:4000::/34}.

5.2. Adding Shared IPv4 Addresses to an IPv6 Network

5.2.1. IPv6 network of the ISP

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, but does not have 2^20 IPv4 addresses. IPv4 prefixes it can use are only 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 must 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 Section 5.1 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

Table 5

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, 2001:0db8:0000:0001:3000::/80}
```

The CE of whose IPv6 prefix is, for example, 2001:db8:0bbb:bb00::/56, derives its IPv4 address and its port set as follows (Section 4.3):

```
IPv6 prefix         : 2001:0db8:0bbb:bb00::/56

Rule IPv6 prefix(i) : 2001:0db8:0800::/38 (longest match)
Rule IPv4 prefix(i) : 192.4.0.0/16
EA-bits length(i)   : 18

PSID length(i)      : 16 + 18 - 32 = 2
EA bits             :   11 1011 1011 1011 1011
Rule IPv4 prefix(i) : 1100 0000 0000 0100

IPv4 address        : 1100 0000 0000 0100 1110 1110 1110 1110

PSID                : 10

IPv4 address        : 192.4.238.238  (1110 1110 = 238)
Ports               : yyyy 10xx xxxx xxxx
                        with y..y > 0, and x...x any value
```

In the Mesh variant, a packet sent to port IPv4 address 192.4.238.238 and port 7777 is tunneled to IPv6 address obtained as follows (Section 4.4.1):

```
  IPv4 address         : 192.4.238.238
  Port field           : 6666

  Rule IPv4 prefix (i) : 192.4.0.0/16  (longest match)
  EA-bits length (i)   : 18
  Rule IPv6 prefix (i) : 2001:0db8:0800::/38

  IPv4 address         : 1100 0000 0000 0100 1110 1110 1110 1110
  EA bits
  Rule IPv4 prefix (i) : 1100 0000 0000 0100
  IPv4 suffix          :                     1110 1110 1110 1110

  PSID length (i)      : 19 - 17 = 2
  Port field           : 0001 1110 0110 0001  (7777)
  PSID                 :      11

  EA bits              :                     1110 1110 1110 1110 11
                       :                      11 1011 1011 1011 1011
  Derived IPv6 prefix  : 2001:0db8:0bbb:bb00::/56
  IPv6 address         : 2001:0db8:0bbb:bb00:3000:192.4.238.238:yyyy
     with yyyy = CNP in the Header-mapping variant
               = 0 in the Encapsulation variant
```

5.2.2. IPv6 Network of a Third-Party Provider

We now consider an ISP that has the same need as in the previous section except that, instead of using its own IPv6 infrastructure, it uses that of a third party whose CPEs are imposed in customer sites. These CPEs use a suffix = 0xF to route IPv6 packets to physical ports to which CEs are attached. It is supposed to have the same IPv4 prefixes (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).

By adding a Domain IPv6 suffix = 0xF in parameters sent to CE's, the same Mapping rules can be used, i.er. :

```
          {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, 2001:0db8:0000:0001:3000::/80}
```

CEs derive their own IPv4 addresses and port sets as in the previous section, except that they ignore the Domain IPv6 suffix at the end of their delegated IPv6 prefixes.

The IPv6 address derived from IPv4 address 192.4.238.238 and port 7777 is derived as in the previous section except that, after the step that obtains EA bits = 11 1011 1011 1011 1011, we now have:

```
Derived IPv6 prefix  : 2001:0db8:0bbb:bbf0::/60 (suffix added)
IPv6 address         : 2001:0db8:0bbb:bbf0:3000:xxxx:xxxx:yyyy
```

5.3. 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, at least for its chosen variants. 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 Section 5.1. 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:db8:0:1:3000::/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:db8:0:1:3000:x.x.x.x::/(80+n)

- In router where IPv4 address x.x.x.x was assigned to a CPE, now delegate IPv6 prefix 2001:db8:0:1:3000: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 Section 5.2.

5.4. Adding IPv6 and 4rd Service to a Net-10 network

In this use case, we consider an ISP that, possibly because some of its network devices are not yet IPv6 capable, has only deployed IPv4 and that, because it did not have enough IPv4 addresses did it with private IPv4 addresses of [RFC1918], say 10.x.x.x to support up to 2^24 customers (a so called Net-10 network). It wishes to add IPv6 service without delay to this network, using for this 6rd [RFC5969], and also wishes to offer incoming IPv4 connectivity to its customers with a simpler solution than that of PCP [I-D.ietf-pcp-base]. 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 192.16.0.0/16 (0xc610). The resulting sharing ratio is 2^24 / 2^(32-16) = 256, giving a PSID length of 8.

The ISP installs or several BRs, at the its border to the public IPv4 Internet, to support both 6rd and 4rd BR functions (4rd function on top of 6rd function). 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:

{192.16.0.0/16, 24, 2001:db8:0:0:3000::/80}
{0.0.0.0/, 32, 2001:db8:0:100:3000:/80,}

Once this is done, any customer device that supports 4rd can use its assigned IPv4 address and port set of 240 ports.It can thus avoid cascading its NAT44 with the NAT44 carrier-grade NAT44 of the ISP. Its site can get back the port-forwarding function of its NAT44, which is lost in net-10 networks.

Public-address IPv4 packets have their 4rd headers (48 or 60 octets depending on which 4rd Header variant is used) preceded by the Net-10 IPv4 header (20 octets)

6. Security Considerations

Spoofing attacks

R-35 Domain-exit nodes MUST check consistency of IPv6 source address with the 4rd source addresses, and if the check fails MUST silently discard the packet.

This is needed because IPv6 ingress filtering only guarantees that CE-provided source addresses do belong to the right CEs (ref. [RFC3704]). It does not guarantee that the Tunnel packets are built in compliance with rules of the present specification.

The check can be performed in two steps. First, the Domain-exit node derives an IPv6 prefix from the source 4rd address according to Section 4.4.1. It then verifies that this prefix is present at the beginning of the IPv6 source address.

With this precaution, and provided IPv6 ingress filtering is effective in the Domain, no opportunity for spoofing attacks 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 introduced by 4rd has been identified.

Fragmentation-related attacks

As discussed in Section 4.5, BRs of Domains that assign shared IPv4 addresses to CEs maintain dynamic tables for fragmented datagrams that go from the IPv4 Internet to these CEs and that go from these CEs to the IPv4 Internet.

* In the CE to Internet direction, this does not open a vulnerability to DOS-attacks because the table has at most one entry per CE.

* In the Internet to CE direction, this opens a vulnerability any remote host can send a large number of first datagram fragments without sending any following fragment, thus occupying many table entries. This vulnerability has no effect on non-fragmented datagrams but is unavoidable in any network that shares IPv4 addresses between customers, be its statically or dynamically. The obvious way to eliminate this vulnerability is to use IPv6.

7. IANA Considerations

IANA is requested to allocate the following:

- An IPv6 Interface-ID type reserved for 4rd (the V octet of Section 4.4.2). Its proposed value is 0x03. A registry for Interface-ID types that have neither local scope nor Modified EUI-64 format of [RFC4291]) could be created on this occasion. It would be available to serve other needs that may exist in the future.

- A reserved IPv4 address to be used as source of ICMPv4 messages due to ICMPv6 error messages. Its proposed value is 192.70.192.254 (Section 4.7).

- Two DHCPv6 option codes, to be defined, for the OPTION_4RD and OPTION_4RD_RULE options of Section 4.8.

8. 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):

- 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)
- chen-softwire-4v6-add-format (2011-10-2)
- boucadair-softwire-stateless-requirements (2011-09-08)
- mawatari-softwire-464xlat (2011-10-16)
- mdt-softwire-mapping-address-and-port (2011-11-25)

9. Acknowledgements

This specification has benefited over several years from independent proposals, questions, comments, constructive suggestions, and useful criticisms, from numerous IETF contributors. The author would like to thank all of them, but more particularly, in first name alphabetical order, Brian Carpenter, Behcet Sarikay,a Congxiao Bao, Dan Wing, Francis Dupont, Gabor, Bajko, Gang Chen, Hui Deng, Jacni Qin, Jan Zorz, Jaro Arkko, Laurent Toutain, Leaf Yeh, Mark Townsley, Maoke Chen, 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, Yiu Lee.

10. References

10.1. 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.

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, October 1996.

[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.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, September 2001.

[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.

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

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

10.2. 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.xli-behave-icmp-address]
Li, X., Bao, C., Wing, D., Vaithianathan, R., and G. Huston, "Stateless Source Address Mapping for ICMPv6 Packets", draft-xli-behave-icmp-address-04 (work in progress), May 2011.

[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.

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

[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.

[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.

[RFC6324] Nakibly, G. and F. Templin, "Routing Loop Attack Using IPv6 Automatic Tunnels: Problem Statement and Proposed Mitigations", RFC 6324, August 2011.

Author's Address

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

Email: despres.remi@laposte.net