Mapping of Address and Port with Encapsulation (MAP)
draft-ietf-softwire-map-10

Abstract

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

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

Mapping of IPv4 addresses in IPv6 addresses has been described in
numerous mechanisms dating back to 1996 [RFC1933]. The Automatic
tunneling mechanism described in RFC1933 assigned a globally unique
IPv6 address to a host by combining the host's IPv4 address with a
well-known IPv6 prefix. Given an IPv6 packet with a destination
address with an embedded IPv4 address, a node could automatically
tunnel this packet by extracting the IPv4 tunnel end-point address
from the IPv6 destination address.

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

The commonalities of all these IPv6 over IPv4 mechanisms are:

o Automatically provisions an IPv6 address for a host or an IPv6
   prefix for a site

o Algorithmic or implicit address resolution of tunnel end point
   addresses. Given an IPv6 destination address, an IPv4 tunnel
   endpoint address can be calculated.

o Embedding of an IPv4 address or part thereof into an IPv6 address.

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

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

o Provisioning an IPv4 prefix, an IPv4 address or a shared IPv4
   address.

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

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

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

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

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

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

2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

3. Terminology

MAP domain: One or more MAP CEs and BRs connected to the same virtual link. A service provider may deploy a single MAP domain, or may utilize multiple MAP domains.

MAP rule   A set of parameters describing the mapping between an IPv4 prefix, IPv4 address or shared IPv4 address and an IPv6 prefix or address. Each domain uses a different mapping rule set.

MAP node   A device that implements MAP.

MAP Border Relay (BR):   A MAP enabled router managed by the service provider at the edge of a MAP domain. A Border Relay router has at least an IPv6-enabled interface and an IPv4 interface connected to the native IPv4 network. A MAP BR may also be referred to simply as a "BR" within the context of MAP.

   MAP Customer Edge (CE): A device functioning as a Customer Edge
                           router in a MAP deployment. A typical MAP CE
                           adopting MAP rules will serve a residential
                           site with one WAN side interface, and one or
                           more LAN side interfaces. A MAP CE may also
                           be referred to simply as a "CE" within the
                           context of MAP.

   Port-set: The separate part of the transport layer port
                           space; denoted as a port-set.

   Port-set ID (PSID):   Algorithmically identifies a set of ports
                           exclusively assigned to a CE.

   Shared IPv4 address: An IPv4 address that is shared among multiple
                           CEs. Only ports that belong to the assigned
                           port-set can be used for communication. Also
                           known as a Port-Restricted IPv4 address.

   End-user IPv6 prefix: The IPv6 prefix assigned to an End-user CE by
                           other means than MAP itself. E.g.
                           Provisioned using DHCPv6 PD [RFC3633],
                           assigned via SLAAC [RFC4862], or configured
                           manually. It is unique for each CE.

   MAP IPv6 address: The IPv6 address used to reach the MAP
                           function of a CE from other CEs and from BRs.

   Rule IPv6 prefix: An IPv6 prefix assigned by a Service Provider
                           for a mapping rule.

   Rule IPv4 prefix: An IPv4 prefix assigned by a Service Provider
                           for a mapping rule.

   Embedded Address (EA) bits: The IPv4 EA-bits in the IPv6 address
                           identify an IPv4 prefix/address (or part
                           thereof) or a shared IPv4 address (or part
                           thereof) and a port-set identifier.

4. Architecture

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

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

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

The MAP architecture described here restricts the use of the shared
IPv4 address to only be used as the global address (outside) of the
NAPT [RFC2663] running on the CE. A shared IPv4 address MUST NOT be
used to identify an interface. While it is theoretically possible to
make host stacks and applications port-aware, it would be a drastic
change to the IP model [RFC6250].

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

This architecture is illustrated in Figure 1.

```
       User N
     Private IPv4
     |  Network
     |
 O--+---------------O
 |  |  MAP CE       |
 | +-----+--------+ |
 | NAPT44|  MAP   | |
 | +-----+        | |\      ,-------.                        .------.
 |      +--------+ | \ ,-'          `-.              ,-'         `-.
 O----------------O  /                \   O---------O  /   Public    \
                    /    IPv6 only  \  |  MAP    | /      IPv4        \
                   (     Network      --+  Border +-      Network     )
                    \  (MAP Domain) /  |  Relay  | \                  /
 O----------------O  \              /   O---------O  \                /
 |    MAP   CE     | /".         ,-'                  `-.         ,-'
 | +-----+--------+ | /    `----+--'                      ------'
```

```
| NAPT44|  MAP   | |/
| +-----+        | |
|   |   +--------+ |
O---+-------------O
    |
      User M
   Private IPv4
      Network
```
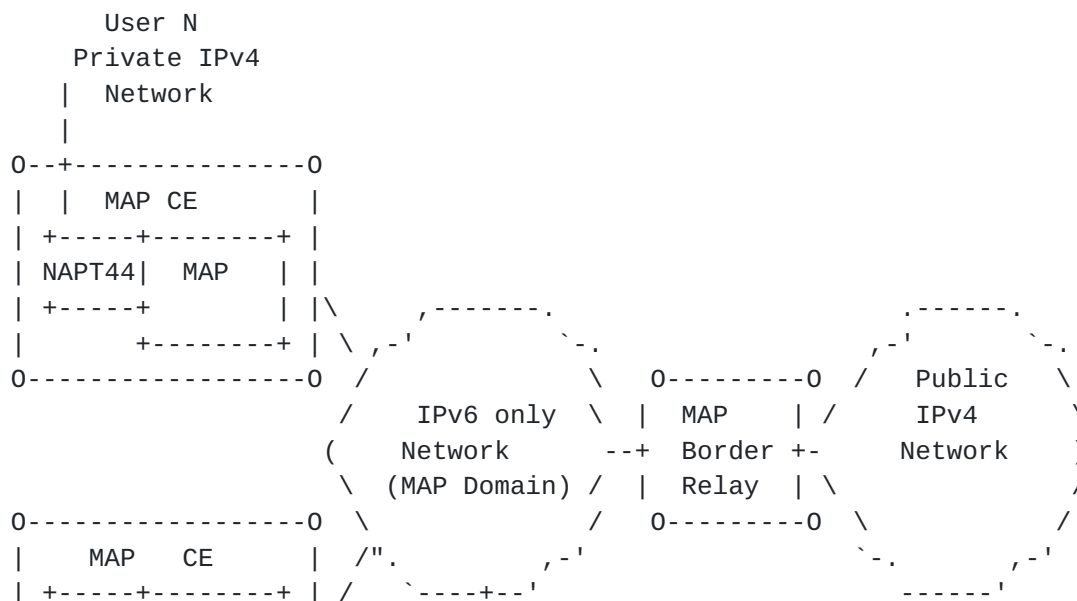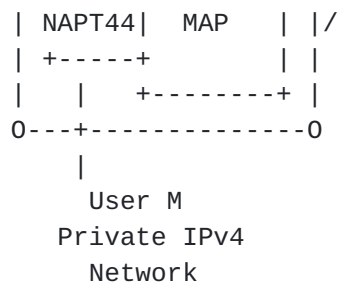
Figure 1: Network Topology

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

5. Mapping Algorithm

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

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

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

There are two types of mapping rules:

1. Basic Mapping Rule (BMR) - mandatory. A CE can be provisioned
    with multiple End-user IPv6 prefixes. There can only be one
    Basic Mapping Rule per End-user IPv6 prefix. However all CE's
    having End-user IPv6 prefixes within (aggregated by) the same
    Rule IPv6 prefix may share the same Basic Mapping Rule. In
    combination with the End-user IPv6 prefix, the Basic Mapping Rule
    is used to derive the IPv4 prefix, address, or shared address and
    the PSID assigned to the CE.

2. Forwarding Mapping Rule (FMR) - optional, used for forwarding.
    The Basic Mapping Rule may also be a Forwarding Mapping Rule.
    Each Forwarding Mapping Rule will result in an entry in the Rules
    table for the Rule IPv4 prefix. Given a destination IPv4 address
    and port within the MAP domain, a MAP node can use the matching

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

Both mapping rules share the same parameters:

o Rule IPv6 prefix (including prefix length)

o Rule IPv4 prefix (including prefix length)

o Rule EA-bits length (in bits)

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

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

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

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

5.1. Port mapping algorithm

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

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

When the PSID is embedded in the End-user IPv6 prefix, then to
minimise dependencies between the End-user IPv6 prefix and the
assigned port-set, it is desirable to minimize the restrictions of
possible PSID values. This is achieved by using an infix
representation of the port value. Using such a representation, the

well-known ports are excluded by restrictions on the value of the
high-order bitfield (A) rather than the PSID.

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

```
                     0                       1
                     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
                    +-----------+-----------+-------+
       Ports in     |     A     |    PSID   |   M   |
    the CE port-set  |    > 0    |           |       |
                    +-----------+-----------+-------+
                    |   a bits  |   k bits  |m bits |
```
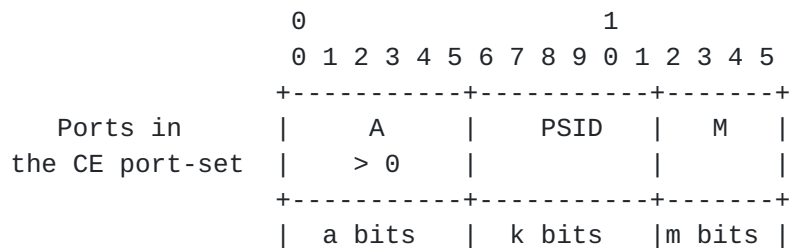
            Figure 2: Structure of a port-restricted port field

a bits: The number of offset bits. 6 by default as this excludes the
   system ports (0-1023).

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

PSID: The Port-Set Identifier (PSID). Different PSIDs guarantee
   non-overlapping port-sets.

k bits: The length in bits of the PSID field. The sharing ratio is
   2^k. The number of ports assigned to the user is $2^{(16-k)} - 2^m$
   (excluded ports)

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

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

5.2. Basic mapping rule (BMR)

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

o   Rule IPv6 prefix (including prefix length)

o Rule IPv4 prefix (including prefix length)

o Rule EA-bits length (in bits)

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


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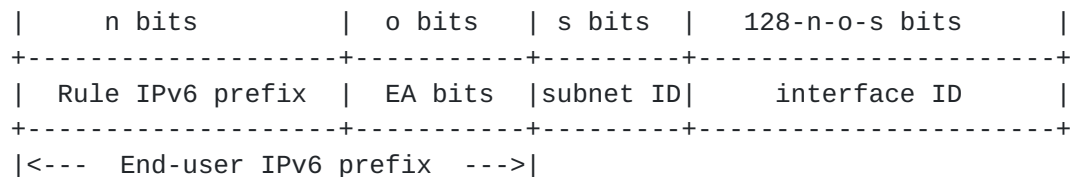

                     Figure 3: MAP IPv6 Address Format

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

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

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

Define:

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

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

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

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

create a complete IPv4 address (or prefix), the IPv4 address suffix
(p) from the EA bits, is concatenated with the Rule IPv4 prefix (r
bits).

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

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

```
                    IPv4 prefix:


             |   r bits    |  o bits =  p bits   |
             +-------------+---------------------+
             |  Rule IPv4  | IPv4 Address suffix |
             +-------------+---------------------+
             |              < 32 bits            |
```
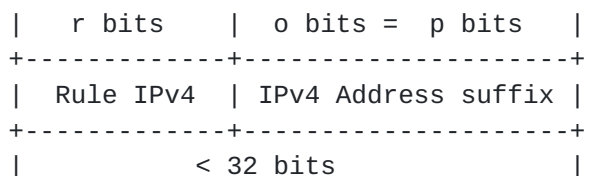
                    Figure 4: IPv4 prefix

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

```
                  Complete IPv4 address:


             |   r bits    |  o bits = p bits    |
             +-------------+---------------------+
             |  Rule IPv4  | IPv4 Address suffix |
             +-------------+---------------------+
             |              32 bits              |
```
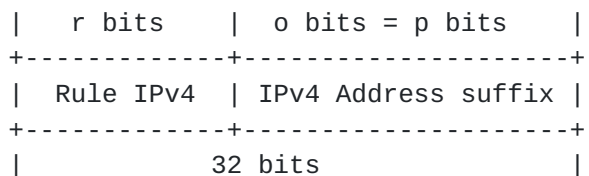
                 Figure 5: Complete IPv4 address

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

Shared IPv4 address:

```
| r bits   |           p bits        |       |  q bits  |
+-------------+---------------------+       +------------+
|  Rule IPv4  | IPv4 Address suffix |       |Port-Set ID |
+-------------+---------------------+       +------------+
|              32 bits              |
```
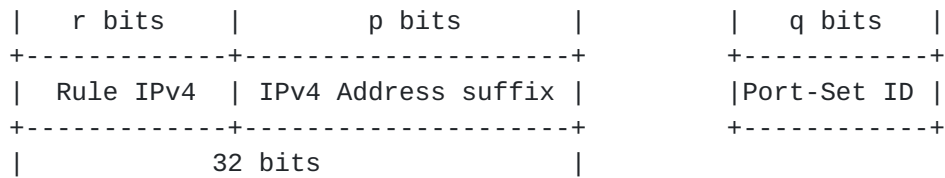
Figure 6: Shared IPv4 address

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

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

5.3. Forwarding mapping rule (FMR)

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

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

```
      |        32 bits        |        |      16 bits       |
      +------------------------+        +-------------------+
      | IPv4 destination address |      |   IPv4 dest port  |
      +------------------------+        +-------------------+
                  :         :          ___/        :
                  | p bits  |         /  q bits   :
                  +---------+        +-----------+
                  |IPv4  sufx|       |Port-Set ID |
                  +---------+        +-----------+
                 \         /   ____/    _____/
                  \      :  __/    ____/
                   \    : /      /
      |    n bits       | o bits | s bits |   128-n-o-s bits      |
      +-----------------+--------+--------+----------------------+-----------+
      |  Rule IPv6 prefix  | EA bits |subnet ID|    interface ID      |
      +-----------------+--------+--------+----------------------+-----------+
      |<---  End-user IPv6 prefix  --->|
```
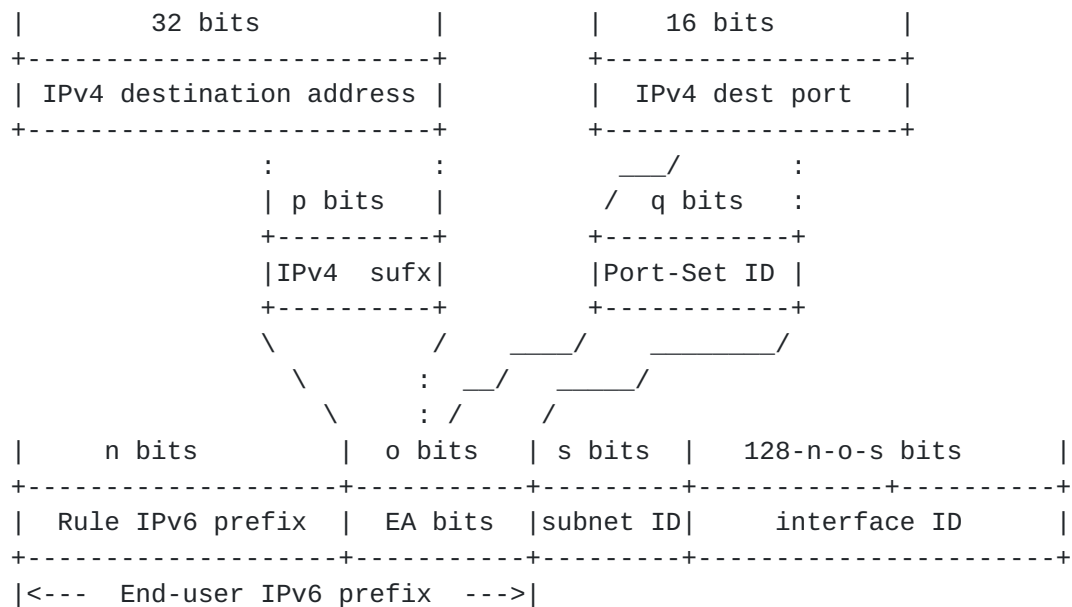
Figure 7: Deriving of MAP IPv6 address

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

5.4. Destinations outside the MAP domain

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

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

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

6. The IPv6 Interface Identifier

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

```
      |          128-n-o-s bits        |
      | 16 bits|    32 bits     | 16 bits|
```
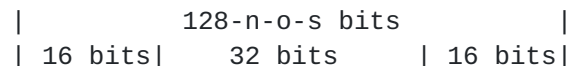
```
+--------+----------------+--------+
|   0    |  IPv4 address  |  PSID  |
+--------+----+-----------+--------+
```

Figure 8

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

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

7. MAP Configuration

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

o The Basic Mapping Rule and optionally the Forwarding Mapping
   Rules, including the Rule IPv6 prefix, Rule IPv4 prefix, and
   Length of EA bits

o Hub and spoke mode or Mesh mode. (If all traffic should be sent
   to the BR, or if direct CE to CE traffic should be supported).

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

7.1. MAP CE

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

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

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

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

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

7.2. MAP BR

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

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

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

8. Forwarding Considerations

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

MAP uses Encapsulation mode as specified in [RFC2473].

For a shared IPv4 address, a MAP CE forwarding IPv4 packets from the
LAN performs NAT44 functions first and creates appropriate NAT44
bindings. The resulting IPv4 packets MUST contain the source IPv4
address and source transport identifiers specified by the MAP
provisioning parameters. The IPv4 packet is forwarded using the CE's

MAP forwarding function. The IPv6 source and destination addresses
MUST then be derived as per Section 5 of this draft.

8.1. Receiving Rules

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

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

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

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

8.2. ICMP

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

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

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

8.3. Fragmentation and Path MTU Discovery

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

8.3.1. Fragmentation in the MAP domain

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

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

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

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

8.3.2. Receiving IPv4 Fragments on the MAP domain borders

Forwarding of an IPv4 packet received from the outside of the MAP
domain requires the IPv4 destination address and the transport
protocol destination port. The transport protocol information is
only available in the first fragment received. As described in
section 5.3.3 of [RFC6346] a MAP node receiving an IPv4 fragmented

packet from outside has to reassemble the packet before sending the
packet onto the MAP link. If the first packet received contains the
transport protocol information, it is possible to optimize this
behavior by using a cache and forwarding the fragments unchanged.
Implementers of MAP should be aware that there are a number of well-
known attacks against IP fragmentation; see [RFC1858] and [RFC3128].
Implementers should also be aware of additional issues with
reassembling packets at high rates, as described in [RFC4963].

8.3.3. Sending IPv4 fragments to the outside

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

9. NAT44 Considerations

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

10. IANA Considerations

This specification does not require any IANA actions.

11. Security Considerations

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

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

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

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

[RFC6269] outlines general issues with IPv4 address sharing.

12. Contributors

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

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

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

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

    Yu Zhai
    CERNET Center/Tsinghua University
    Room 225, Main Building, Tsinghua University
    Beijing 100084

CN
Email: jacky.zhai@gmail.com

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

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

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

13. Acknowledgements

14. References

14.1. Normative References

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

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

14.2. Informative References

[I-D.ietf-softwire-map-deployment]

Qiong, Q., Chen, M., Chen, G., Tsou, T., and S. Perreault, "Mapping of Address and Port (MAP) - Deployment Considerations", draft-ietf-softwire-map-deployment-03 (work in progress), October 2013.

[I-D.ietf-softwire-map-dhcp]
          Mrugalski, T., Troan, O., Dec, W., Bao, C., leaf.yeh.sdo@gmail.com, l., and X. Deng, "DHCPv6 Options for configuration of Softwire Address and Port Mapped Clients", draft-ietf-softwire-map-dhcp-06 (work in progress), November 2013.

[I-D.ietf-softwire-stateless-4v6-motivation]
          Boucadair, M., Matsushima, S., Lee, Y., Bonness, O., Borges, I., and G. Chen, "Motivations for Carrier-side Stateless IPv4 over IPv6 Migration Solutions", draft-ietf-softwire-stateless-4v6-motivation-05 (work in progress), November 2012.

[I-D.ietf-tsvwg-iana-ports]
          Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", draft-ietf-tsvwg-iana-ports-10 (work in progress), February 2011.

[RFC0897] Postel, J., "Domain name system implementation schedule", RFC 897, February 1984.

[RFC1858] Ziemba, G., Reed, D., and P. Traina, "Security Considerations for IP Fragment Filtering", RFC 1858, October 1995.

[RFC1933] Gilligan, R. and E. Nordmark, "Transition Mechanisms for IPv6 Hosts and Routers", RFC 1933, April 1996.

[RFC2529] Carpenter, B. and C. Jung, "Transmission of IPv6 over IPv4 Domains without Explicit Tunnels", RFC 2529, March 1999.

[RFC2663] Srisuresh, P. and M. Holdrege, "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

   [RFC3056]   Carpenter, B. and K. Moore, "Connection of IPv6 Domains
               via IPv4 Clouds", RFC 3056, February 2001.

   [RFC3128]   Miller, I., "Protection Against a Variant of the Tiny
               Fragment Attack (RFC 1858)", RFC 3128, June 2001.

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

   [RFC4459]   Savola, P., "MTU and Fragmentation Issues with In-the-
               Network Tunneling", RFC 4459, April 2006.

   [RFC4632]   Fuller, V. and T. Li, "Classless Inter-domain Routing
               (CIDR): The Internet Address Assignment and Aggregation
               Plan", BCP 122, RFC 4632, August 2006.

   [RFC4787]   Audet, F. and C. Jennings, "Network Address Translation
               (NAT) Behavioral Requirements for Unicast UDP", BCP 127,
               RFC 4787, January 2007.

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

   [RFC4953]   Touch, J., "Defending TCP Against Spoofing Attacks", RFC
               4953, July 2007.

   [RFC4963]   Heffner, J., Mathis, M., and B. Chandler, "IPv4 Reassembly
               Errors at High Data Rates", RFC 4963, July 2007.

   [RFC5214]   Templin, F., Gleeson, T., and D. Thaler, "Intra-Site
               Automatic Tunnel Addressing Protocol (ISATAP)", RFC 5214,
               March 2008.

   [RFC5382]   Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P.
               Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142,
               RFC 5382, October 2008.

   [RFC5508]   Srisuresh, P., Ford, B., Sivakumar, S., and S. Guha, "NAT
               Behavioral Requirements for ICMP", BCP 148, RFC 5508,
               April 2009.

   [RFC5961]   Ramaiah, A., Stewart, R., and M. Dalal, "Improving TCP's
               Robustness to Blind In-Window Attacks", RFC 5961, August
               2010.

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

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

   [RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-
              Protocol Port Randomization", BCP 156, RFC 6056, January
              2011.

   [RFC6250] Thaler, D., "Evolution of the IP Model", RFC 6250, May
              2011.

   [RFC6269] Ford, M., Boucadair, M., Durand, A., Levis, P., and P.
              Roberts, "Issues with IP Address Sharing", RFC 6269, June
              2011.

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

   [RFC6333] Durand, A., Droms, R., Woodyatt, J., and Y. Lee, "Dual-
              Stack Lite Broadband Deployments Following IPv4
              Exhaustion", RFC 6333, August 2011.

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

Appendix A. **Examples**

   Example 1 - Basic Mapping Rule

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

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

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

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

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

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

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

Example 2 - BR:

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

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

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

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

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

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

The resulting IPv6 packet will have the following key fields:

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

Example 3 - FMR:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

   o In keeping with the general MAP algorithm the port-set MUST be
      derivable from a port-set identifier (PSID) that can be embedded
      in the End-user IPv6 prefix.

o The mapping MUST be reversible, such that, given the port number,
   the PSID of the port-set to which it belongs can be quickly
   derived.

o The algorithm MUST allow a broad range of address sharing ratios.

o It SHOULD be possible to exclude subsets of the complete port
   numbering space from assignment. Most operators would exclude the
   system ports (0-1023). A conservative operator might exclude all
   but the transient ports (49152-65535).

o The effect of port exclusion on the possible values of the End-
   user IPv6 prefix (i.e., due to restrictions on the PSID value)
   SHOULD be minimized.

o For administrative simplicity, the algorithm SHOULD allocate the
   the same or almost the same number of ports to each CE sharing a
   given IPv4 address.

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

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

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

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

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

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

(2) ... $PSID = trunc((P \bmod (R * M)) / M)$

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

Formula (1) can be interpreted as follows. First, the available port
space is divided into blocks of size R * M. Each block is divided

into R individual ranges of length M. The index i in formula (1)
selects a block, PSID selects a range within that block, and the
index j selects a specific port value within the range. On the basis
of this interpretation:

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

o PSID ranges from 0 to R - 1;

o j ranges from 0 to M - 1.

B.1. Bit Representation of the Algorithm

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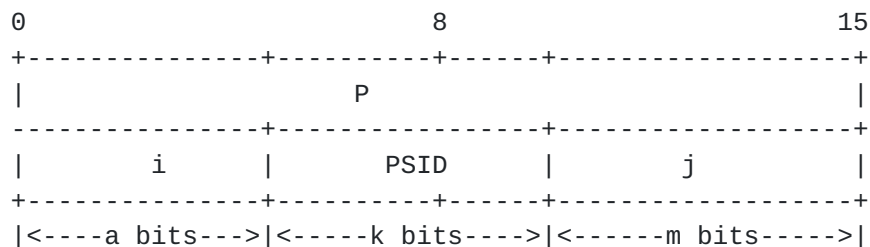Figure 9.

```
0                              8                          15
+---------------+----------+------+------------------+
|                         P                          |
+---------------+----------------+------------------+
|       i       |      PSID      |        j         |
+---------------+----------+------+------------------+
|<----a bits--->|<-----k bits---->|<------m bits----->|
```

                Figure 9: Bit Representation of a Port Number

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

For any port number, the PSID can be obtained by a bit mask
operation.

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

   if block size R * M is a factor of N, the number of ports to be
   excluded.

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

B.2. GMA examples

   For example, for R = 256, PSID = 0, offset: a = 6 and PSID length: k
   = 8 bits

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

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

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

   Available ports (1 range) : 0-1023

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

Authors' Addresses

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

   Email: ot@cisco.com

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

   Email: wdec@cisco.com

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

   Email: xing@cernet.edu.cn

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

   Email: congxiao@cernet.edu.cn

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

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

   Tetsuya Murakami
   IP Infusion
   1188 East Arques Avenue
   Sunnyvale
   USA

   Email: tetsuya@ipinfusion.com

    Tom Taylor (editor)
    Huawei Technologies
    Ottawa
    Canada

    Email: tom.taylor.stds@gmail.com