Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 16, 2014

Q. Sun
China Telecom
M. Chen
FreeBit
G. Chen
China Mobile
T. Tsou
Huawei Technologies
S. Perreault
Viagenie

July 15, 2013

Mapping of Address and Port (MAP) - Deployment Considerations
draft-ietf-softwire-map-deployment-02

Abstract

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

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

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

Mapping of Address and Port (MAP) [I-D.ietf-softwire-map] is designed in response to the requirement of stateless residual deployment. The term "residual deployment" refers to utilizing not-yet-assigned or recalled IPv4 addresses for IPv4 communications going across the IPv6 domain backbone. MAP assumes the IPv6-only backbone as the prerequisite of deployment so that native IPv6 services and applications are fully supported and encouraged. The statelessness of MAP ensures only moderate overhead is added to part of the network devices.

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

1. What are the requirements of MAP deployment ?

2. What technical options needs to be considered when deploying MAP, and how?

3. How does MAP impact on the address planning for both IPv6 and IPv4 pools?

4. How does MAP impact on daily network operations and administrations?

5. How do we migrate to IPv6-only network with the help of MAP?

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

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

2. Conventions

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

3. Case Studies

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

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

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

```
                     +------------------------------+
                     |        MAP Domain            |
                 +---+---------------+--------------|
     +--------+  +                   |              |
     |        |  +---------+      +--+--+           |
     |  Host  |--|   CPE   |      |     |           |
     |        |  |(MAP CE) |======| BNG | ======+---------+   +-----------+
     +--------+  +---------+      +--|--+       |  Core   |   |   IPv4    |
     +--------+      +---------------+          | Router  |---| Internet  |
     |        |  +---|-----+      +--+--+       |(MAP BR) |   |           |
     |  Host  |--|   CPE   |======|     | ======+---------+   +-----------+
     |        |  |(MAP CE) |      | BNG |           |
     +--------+  +---------+      +--+--+           |
                 +                   |              |
                 +-------------------+--------------+
```

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

4. Deployment Consideration

4.1. Network Models

A MAP domain connects IPv4 subnets, including home networks, enterprise networks, and collective residence faclities, with the IPv4 Internet through the IPv6 routing infrastructure.

In home network, three network models are defined in [I-D.ietf-homenet-arch]: A. single ISP, single customer edge router (CER), internal routers; B. two ISPs, two CERs, shared subnet; C. two ISPs, one CER, shared subnet. Models A and B are different from model C when using MAP. For models A and B, the CE (=CER) needs to correspond with only one BR, while in model C one CE has to correspond with multiple BRs. Figure 2 illustrates a typical case, where the home network has multiple connections to multiple providers or multiple logical connections to the same provider. In general, a CE may have different paths towards multiple MAP border relays.

```
            +-------+-------+     +-------+-------+         \
            |   Service     |     |   Service     |          \
            |  Provider A   |     |  Provider B   |           | Service
            |    Router     |     |    Router     |           | Provider
            +-------+-------+     +-------+-------+           | network
                     |                 |                     /
                     |    Customer     |                   /
                     |    Internet     |                  /
                     |   connections   |                 |
                    +---------+---------+                 \
                    |       IPv6        |                   \
                    |   Customer Edge   |                    \
                    |      Router       |                    /
                    +---------+---------+                   /
                              |                            /
                              |                            | End-User
      ---+------------+-------+--------+-------------+---  | network(s)
         |            |                |             |      \
    +----+-----+ +----+-----+     +----+-----+ +-----+----+  \
    |IPv6 Host | |IPv6 Host |     | IPv6 Host| |IPv6 Host |  /
    |          | |          |     |          | |          | /
    +----------+ +----------+     +----------+ +----------+
```

Figure 2: Relations between home networking and MAP domain

4.2. Building the MAP Domain

When deploying stateless MAP in an operational network, a provider should firstly do MAP domain planning based on that existing network. According to the definition of [I-D.ietf-softwire-map], a MAP domain is a set of MAP CEs and BRs connected to the same virtual link. One MAP domain shares a common BR. When multiple IPv4 subnets are deployed in one MAP domain, it is recommanded to further divide the MAP domain into mutiple sub-domains, each with only one IPv4 subnet. This can simplify the MAP domain planning. All CEs in the MAP domain are provisioned with the same set of MAP rules by MAP DHCPv6 server [I-D.ietf-softwire-map-dhcp]. There might be multiple BMRs in one MAP domain, and CE would pick up its own BMR by longest prefix matching lookup. However, all CEs within the sub-domain will have the same BMR. In hub and spoke mode, CE would use DMR as its only FMR for outbound traffic; while in mesh mode, a longest-matching prefix lookup is done in the IPv4 routing table and the correct FMR is chosen.

[Note:Currently, there is no DMR in MAP-E. The IPv6 address of the BR could be provisioned by the DS-Lite AFTR Name option. But the DMR is still in use in MAP-T. Is this the final decision ?]

Basically, operator should firstly determine its own deployment topology for MAP domain as described in Section 4.2.1, as different considerations apply for different deployment models. Next, MAP domain planning, MAP rule provision, addressing and routing, etc., for a MAP domain should be taken into consideration, as discussed in the sections following Section 4.2.1.

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

4.2.1. MAP Deployment Model Planning

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

architecture.

In mesh topology, CE to CE traffic flows are optimized since they pass directly between the two nodes. Mesh topology is recommended when CE to CE traffic is high and there are not too many MAP rules, say fewer than 10 MAP rules, in the given domain.

4.2.2. MAP Domain Planning

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

MAP subdomains can be defined to support provision of differentiated service. Different subdomains could be distinguished by different Rule IPv4 prefixes. As stated previously, all CEs within the same MAP subdomain will have the same Rule IPv4 prefix, Rule IPv6 prefix and PSID parameters.

4.2.3. MAP Rule Provisioning

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

4.2.3.1. Full Hub and Spoke Communication among CEs

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

4.2.3.2. Full Mesh Communication among CEs

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

destination address. If the FMR is found (destination is one of the CEs in the MAP domain), the packet will be forwarded to associated CE directly without going though BR. If the FMR is not found (destination is out of the MAP domain), the DMR will be selected as FMR, the CE then forwards the packet to the associated BR.

4.2.3.3. Mesh or Hub/Spoke communication among some CEs

Mesh communications among some CEs along with Hub/Spoke communications among some other CEs can be achieved by which differentiated FMRs are assigned to CEs. For instance, as Figure 3 shown, Mapping rule 1, Mapping rule 2, Mapping rule 3 is provisioned to CE1, CE2, CE3 respectively as BMR, and rule 1 and rule2, and rule 1 and rule 2 and rule 3, and rule 2 and rule 3 are assigned to CE1, CE2, CE3 respectively, then CE1 and CE2, CE2 and CE3 communicate directly without going though associated BR (Mesh topology), the communication between CE1 and CE3 must go though BR before reaching peer each other (Hub/Spoke topology).

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

Figure 3: Mapping rules assigned to CEs in example

4.2.4. MAP DHCPv6 server deployment consideration

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

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

MAP domain : DHCPv6 server = 1:1 This is the ideal solution that each MAP domain would have its own MAP DHCPv6 server. In this case, MAP DHCPv6 server only needs to configure parameters for the specific MAP domain. It is highly recommended to adopt this deployment model in stateless MAP.

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

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

Besides, the situation of remaining IPv4 address prefixes may have big impact on MAP rule planning, especially for service operators who only have rather scattered address space. Since the number of scattered IPv4 address prefixes would be equal to the number of FMR rules within a MAP domain, one should choose as large IPv4 address pool as possible to reduce the number of FMR rules.

4.2.5. **PSID Consideration**

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

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

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

Note that some ISPs may need to offer services in a MAP domain with a

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

4.2.6. Addressing and Routing

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

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

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

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

4.2.7. MAP vs. MAP-T vs. 4rd

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

1. Option header

There may be some options in the IPv4 header, and some of them may not be able to mapped to IPv6 option headers accurately [RFC791][RFC2460]. If translation or 4rd 'reversible translation' is applied, those options can not be supported, and packets with those

options SHOULD be dropped. Encapsulation does not have this problem.

2. ICMP

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

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

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

3. PMTU and fragmentation

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

If the fragment happens in the IPv6 stack, then only the first fragement contains full IPv4 destination address so that BR cannot do the decapsulation well until all fragments has been received. This disables the funtionality of anycast BR. To prevent this problem, MAP require the fragmentation is done in the IPv4 stack to fit the IPv6 domain path MTU. MAP-T and 4rd has not this problem as every IPv6 packet contains the full IPv4 address embedded into the IPv6 address and end-point reassembly is ensured.

4.3. BR Settings

1. BR placement

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

A first concern should be the avoidance of "triangle routing". That is, the path from the CE to an IPv4 peer via the BR should be closer than that would be taken if the CE had native IPv4 connectivity. This can be accomplished easily by placing the BR close to the CE, such that the length of the path from the CE to the BR is minimized.

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

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

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

Note also that MAP being stateless, asymmetric routing is possible, meaning that separate BRs can be used for traffic entering and exiting a MAP domain. This option can be considered for its effects on traffic engineering.

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

2. Reliability Considerations

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

Anycast, already mentioned in section 4.2.1, can be used to ensure reliability of traffic from CE to BR. Since there can be only one Default Mapping Rule per MAP domain, traffic from CE to BR will always use the same destination address. When this address is anycast, reliability is greatly increased. If a BR goes down, it stops advertising the IPv6 anycast address, and traffic is automatically re-routed to other BRs. For this mechanism to work correctly, it is crucial that the anycast route announcement be very closely tied to BR availability. See [RFC4786] for best current practices on the operation of anycast services.

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

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

3. MTU/Fragmentation

If the MTU is well-managed such that the IPv6 MTU on the CE WAN side interface is set so that no fragmentation occurs within the boundary of the MAP domain, then the Tunnel MTU can be set to the known IPv6 MTU minus the size of the encapsulating IPv4 header (40 bytes). For example, if the IPv6 MTU is known to be 1500 bytes, the Tunnel MTU might be set to 1460 bytes. Without more specific information, the Tunnel MTU SHOULD default to 1280 bytes.

It is important that fragments of a MAP packet sent according to the Default Mapping Rule be handled by the same BR. This can be a problem when using an anycast BR address and routing fluctuations cause fragments of a packet to be routed to multiple BRs.

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

completely, it is recommended to increase the MTU of the IPv6 network so that no fragmentation and Too Big ICMP error occurs.

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

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

Another peculiarity of shared IPv4 addresses is that, without precaution, a destination could simultaneously receive from different sources fragmented datagrams that have the same Datagram ID (the Identification field of [RFC0791]). This would disturb the reassembly process. To eliminate this risk, CE MUST rewrite the datagram ID to a unique value among CEs sharing an IPv4 address upon ending the packet over a MAP domain. This value SHOULD be generated locally within the port-range ssigned to a given CE. Note that replacing a Datagram ID in an IPv4 header implies an update of its Header-checksum field, by adding to it the one's complement difference between the old and the new values.

4.4. CE Settings

1. bridging vs. routing

In routing manner, the CE runs a standard NAT44 [RFC3022] using the allocated public address as external IP and ports via DHCPv6 option. When receiving an IPv4 packet with private source address from its end hosts, it performs NAT44 function by translating the source address into public and selecting a port from the allocated port-set. Then it encapsulates/translates the packet with the concentrator's IPv6 address as destination IPv6 address, and forwards it to the concentrator. When receiving an IPv6 packet from the concentrator, the initiator decapsulates/translates the IPv6 packet to get the IPv4 packet with public destination IPv4 address. Then it performs NAT44 function and translates the destination address into private one.

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

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

2. CE-initiated application

CE-initiated case is applied for situations where applications run on CE directly. If the application in CE use the public address directly, it might conflict with other CEs. So it is highly suggested that CE should also run a local NAT to map a private address to public address in CE. In this way, the CE IPv4 address passed to local applications would be conflict with other CEs. Moreover, CE should guarantee that every source port number in the outgoing packets falls into the allocated port-set.

4.5. Supporting System

1. Lawful Intercept

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

2. Traffic Logging

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

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

3. Geo-location aware service

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

4. User Managment

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

5. MAP Address Planning

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

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

Residual deployment starts from IPv6 address planning.

(A) IPv6 considerations

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

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

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

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

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

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

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

(B) IPv4 considerations

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

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

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

If we use both blocks, M = 2^24 + 2^24, and therefore m = 25. Suppose the intended sharing ratio is 8 subscribers per address, resulting in (65536 - 1024)/8 = 8064 ports per subscriber assuming that the well-known ports are excluded. Then the PSID length to achieve this will be log2(8) = 3 bits. Bearing in mind the IPv4 24 bit prefix length for each of our two prefixes, the EA-bit length is (32 - 24) + 3 = 11 bits.

(B3) For each IPv4 prefix, Hi, of length hi, choose an prefix extension, say Ri of length ri = m - (32 - hi).

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

Then we have:

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

Sometimes the IPv4 residual pool is not well aggregated and the contiguous address blocks may have different sizes. For example, in (B1), if we have H1 = 59.112.0.0/13 and H2 = 219.120.0.0/16 as the IPv4 residual pool, then M = 2^19 + 2^16, and in such a case, we must pick m so that m = ceil(log2(M)), where "ceil(x)" means the minimum integer not less than x, i.e., m = 20 in this case. Therefore r1 = 20 - (32 - 13) = 1, while r2 = 20 - (32 - 16) = 4. Several combinations are available for the R1 and R2 and one only needs to pay attention to avoiding overlapping when picking up the values.

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

(C1) Derive the length c of the MAP domain IPv6 prefix, C, that will appear at the beginning of all delegated prefixes (c = x - (n + k)).

(C2) Take any prefix for this C of length c that starts with a RIR-allocated IPv6 prefix.

(C3) For each IPv4 prefix Hi, make the rule, in which the key is Hi and the value is the domain IPv6 prefix C followed by the rule index Ri. Then this i-th rule's Rule IPv6 Prefix will have the length of (c + ri).

Then we can do that:

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

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

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

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

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

5.2. Remarks on Deployment Paradigms

1. IPv6 address planning in residual deployment is independent of the usage of the residual IPv4 addresses. The IPv4 address pool for "residual deployment" contains IPv4 addresses not yet allocated to customers/subscribers and/or those already recalled from ex-customers, re-programmed into relatively well-aggregated blocks.

2. It is recommended to have the number of rule entries as less as possible so that the merit of statelss deployment is reflected in practical performances. However, this effort is often constrained by the condition of an operator whether (a): it holds large-enough contigious IPv4 address block(s) for the residual deployment, and (b): a short-enough IPv6 domain prefix so that the /64 delegation is easily satisfied even the EA-bits is quite long. When condition (a) is not satisfied, sub-domains have to be defined for each relatively small but contigious aggregated block; when condition (b) is not satisfied, one has to devide the IPv4 aggregates into smaller blocks artificially in order to reduce the length of EA-bits. When we have good conditions fitting (a) and (b), it is NOT recommended to define short EA-bits with small length of IPv4 suffix (the value p) nor to increase the number of rule entries (also the number of sub-

domains) unless it really has to.

3. An extreme case is, when EA-bits contain the full IPv4 address while a full IPv4 address is assigned to a CE, i.e., o = p = 32, and q = 0, the MAP address format becomes almost equivalent to RFC6052-format [RFC6052] except the off-domain IPv4 peer's mapped IPv6 address. This frees the domain to distribute rules but the DMR. In such a case, IPv6 addressing is fully dependent of IPv4, which defers from the typical residual deployment case. MAP is mainly designed for residual deployment but also applied for the case of legacy IPv4 networks keeping communication with the IPv4 world over the IPv6 domain without renumbering, as long as the address planning doesn't matter.

4. Another extreme case is, when EA-bits' length becomes to zero, i.e., o = p = q = 0, a rule actually defines a correspondence between an IPv6 address and an IPv4 address (or a prefix), without any algorithmic correlation to each other. Using such a case in practice is not prohibited by the specification, but it is not recommended to deploy null EA-bits in large scale as the concern discussed in the above Remark 2, and as it has the limitation that the PSID must be null (q = 0) and therefore multiple CEs sharing a same IPv4 address is not supported here. It is recommended to apply a stateful solution, like Lightweight 4over6 [I-D.cui-softwire-b4-translated-ds-lite], if a full de-correlation between IPv6 address and IPv4 address as well as port range is demanded.

5. A not-so-extreme case, p = 0, o = q, i.e., only PSID is applied for the EA-bits, is also a case possibly happening in practice. It also potentially generates a huge number of rules and therefore large-scale deployment of this case is not recommended either.

6. For operators who would like to utilize "some bits" of IPv6 address to do service identification, QoS differentiation, etc., it is recommended that these special-purpose bits should be embedded before the EA-bits so as to reduce the possibility of bit-conflict. However, it requires quite shorter IPv6 aggregate prefix of the operator. The bit-conflict is more likely to happen in this case if different domains have different Rule prefix lengths. Operators with this demand should pay attention to the impact on the domain rule planning.

6. Migration Methodology

6.1. Roadmap for MAP-based Solution

6.1.1. Start from Scratch

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

6.1.2. Coexiting Phases

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

6.1.3. Exit Strategy

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

6.2. Migration Mode

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

6.2.1. Passive Transition

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

6.2.2. Active Transition

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

- ISP could install only IPv6 record (i.e. AAAA) in DNS server, which would provide users with IPv6 steering effects. When a host is IPv6-capable and gets IPv6 DNS reply in advance, MAP functionalities would be restricted by IPv6-only record response.

- ISP could retrieve shared IPv4 address by increasing sharing ratio. In this case, number of concurrent IPv4 sessions on MAP CE would be suppressed. It would encourage native IPv6 growth in some extent.

- ISP could allocate a dedicated IPv6 prefix for MAP deployment. The allocation could not only facilitate the differentiation between MAPed traffic and native IPv6 trafffic, but also clearly observe the tendency of MAP traffic. When the traffic is getting down for while, ISP could close the MAP functionalities in some specific area. It would result networks to native IPv6-only capable.

7. IANA Considerations

This specification does not require any IANA actions.

8. Security Considerations

There are no new security considerations pertaining to this document.

9. Contributors

The members of the MAP design team are:

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

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

10. Acknowledgements

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

11. References

11.1. Normative References

[I-D.ietf-softwire-4rd]
Despres, R., Jiang, S., Penno, R., Lee, Y., Chen, G., and M. Chen, "IPv4 Residual Deployment via IPv6 - a Stateless Solution (4rd)", draft-ietf-softwire-4rd-06 (work in progress), July 2013.

[I-D.ietf-softwire-map]
Troan, O., Dec, W., Li, X., Bao, C., Matsushima, S., Murakami, T., and T. Taylor, "Mapping of Address and Port with Encapsulation (MAP)", draft-ietf-softwire-map-07 (work in progress), May 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 Mapping of Address and Port", draft-ietf-softwire-map-dhcp-03 (work in progress), February 2013.

[I-D.ietf-softwire-map-t]
Li, X., Bao, C., Dec, W., Troan, O., Matsushima, S., and T. Murakami, "Mapping of Address and Port using Translation (MAP-T)", draft-ietf-softwire-map-t-03 (work in progress), July 2013.

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

[RFC5342] Eastlake, D., "IANA Considerations and IETF Protocol Usage for IEEE 802 Parameters", BCP 141, RFC 5342, September 2008.

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

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

[RFC6791] Li, X., Bao, C., Wing, D., Vaithianathan, R., and G. Huston, "Stateless Source Address Mapping for ICMPv6 Packets", RFC 6791, November 2012.

11.2. Informative References

[I-D.cui-softwire-b4-translated-ds-lite]
Cui, Y., Sun, Q., Boucadair, M., Tsou, T., Lee, Y., and I. Farrer, "Lightweight 4over6: An Extension to the DS-Lite Architecture", draft-cui-softwire-b4-translated-ds-lite-11 (work in progress), February 2013.

[I-D.ietf-homenet-arch]
Chown, T., Arkko, J., Brandt, A., Troan, O., and J. Weil, "Home Networking Architecture for IPv6", draft-ietf-homenet-arch-08 (work in progress), May 2013.

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

[RFC3194] Durand, A. and C. Huitema, "The H-Density Ratio for Address Assignment Efficiency An Update on the H ratio", RFC 3194, November 2001.

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

Authors' Addresses

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

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

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

Email: fibrib@gmail.com

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

Email: chengang@chinamobile.com

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

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

Simon Perreault
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: simon.perreault@viagenie.ca