Network Working Group   M. Boucadair, Ed.
Internet-Draft   P. Levis
Intended status: Informational   France Telecom
Expires: August 3, 2009   G. Bajko
                                                           T. Savolainen
                                                                   Nokia
                                                        January 30, 2009

IPv4 Connectivity Access in the Context of IPv4 Address Exhaustion
draft-boucadair-port-range-01.txt

Abstract

This memo proposes a solution, based on fractional addresses, to face the IPv4 public address exhaustion. It details the solution and presents a mock-up implementation, with the results of tests that validate the concept. It also describes architectures and how fractional addresses are used to overcome the IPv4 address shortage. A comparison with the alternative Carrier-Grade NAT (CG-NAT) solutions is also elaborated in the document.

Table of Contents

1. Introduction

1.1. Context

It is commonly agreed by the Internet community that the exhaustion of public IPv4 addresses is an ineluctable fact. In this context, the community was mobilized in the past to adopt a promising solution (in particular with the definition of IPv6). Nevertheless, this solution is not globally activated by Service Providers for both financial and strategic reasons. In the meantime, these Service Providers are not indifferent to the alarms recently emitted by the IETF particularly by the reports presented within the GROW working group (Global Routing Operations Working Group) meetings.

G. Huston introduced an extrapolation model to forecast the exhaustion date of IPv4 addresses managed by IANA. This effort indicates that if the current tendency of consumption continues at the same pace, IPv4 addresses exhaustion of IANA's pool would occur in 2011, while RIRs'pool would be exhausted in late 2012. The state of the current consumption of public IPv4 addresses is daily updated and is available at this URL:
http://www.potaroo.net/tools/ipv4/index.html.

1.2. Tentative Solutions: Overview and Limitations

In order to solve this depletion problem, Service Providers need to investigate and enable means to ensure the deployment of their service offerings and their delivery to end users. Two strands may be followed:

(1) Migrate to IPv6:

IPv6 has been introduced for several years as the next version of the IP protocol. This new version offers an abundance of IP addresses as well as several enhancements compared to IPv4 especially with the adoption of hierarchical routing (and therefore allows reducing the routing tables size). IPv6 specifications are mature and current work within the IETF is related to operational aspects. Nevertheless, Service Providers have not largely activated IPv6 in their networks yet.

However, even if a Service Provider activates IPv6, it will be confronted with the problem to ensure a global connectivity towards nowadays Internet v4. Mechanisms such as NAT-PT (Network Address Translation Protocol Translation) were introduced to ensure the interconnection between two heterogeneous realms (i.e. IPv4/IPv6) and to ensure a continuity of IP communications (i.e. end-to-end). It is out of scope of this document to analyze the hurdles of these

solutions.

Despite the current IPv6 deployment situation, IPv6 is a long term and viable alternative to offer IP connectivity services to a large number of customers. From this perspective, Service Providers should avoid introducing new functions and nodes which may be problematic when envisaging migrating to IPv6. This critical requirement should not be taken into account only during the technical engineering phase, but also when elaborating required CAPEX (Capital Expenditure)/OPEX (Operational Expenditure) estimation of activating alternative schemes to solve or to reduce the impact of the IPv4 address exhaustion phenomenon.

Note that this requires deploying interconnection mechanisms with the already existent IPv4 realms. This cost overhead should be considered in transition scenarios.

(2) Enhance current IPv4 architectures and optimize the assignment of IPv4 addresses:

A first example of the implementation of this option is the introduction of a second level of NAT, called Provider-NAT or Carrier Grade NAT (CG-NAT). This node is located in the Service Provider domain. In such option, only private addresses are assigned to end-user home gateways, which still perform their own NAT. The CG-NAT is responsible for translating IP packets issued with private addresses to ones with publicly routable IPv4 addresses when exiting the domain of the Service Provider.

The introduction of the CG-NAT will have an important impact on the applications. Some services will only work in a degraded mode, some will even not work at all (refer to Section 8.1 for more details about encountered hurdles).

Another example of this second option is the proposal that has been made to release IPv4 class E addresses [ID.240space]: concretely to reclassify 240/4 as usable unicast address space. The rationale of this proposal is that since the community has not concluded whether the E block should be considered public or private, and given the current consumption rate, it is clear that the block should not be left unused. This proposal requires updating IP-enabled equipment so as to treat correctly IPv4 addresses belonging to 240/4 blocks. These addresses should be routable and announced for instance between adjacent Autonomous Systems (ASes) through BGP (Border Gateway Protocol) for instance. An exhaustive study should be undertaken to evaluate the economic and technical impact of such new policy. Another alternative is to re-classify class E address as private ones [ID.Eprivate].

1.3. Contribution of this draft

This memo specifies an alternative solution to the Double NAT architecture which aims at solving the depletion problem as encountered by current ISPs. The proposed solution, called Provider-Provisioned CPE is session stateless and does not alter the various offered services. The solution presented in this document does not require severe modifications to current engineering practices as adopted by major Service Providers. Furthermore, the solution is scalable and can be deployed in several variants, especially to prepare the migration towards IPv6.

This draft describes a lightweight architecture that may be deployed by Service Providers to offer IP connectivity services to their already subscribed customers or to new ones. This document provides an implementation scenario. Service Providers are free to enforce their own engineering rules based on their internal policies and available technological means as activated in their IP infrastructure. The solution is flexible enough to be accommodated in various contexts.

The scalability of this solution is similar to current deployed IP architectures. No session-related states are maintained in core nodes operated by a given Service Provider.

This solution can be activated in an end host, CPE (Customer Premises Device), or any other device able to constraint its source port numbers.

This draft is a contribution to the required specification effort mandated in [ID.arkko], especially scenario c.

2. Conventions used in this document

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].

The following abbreviations are used within this document:

- ASN GW: Access Service Network Gateway

- CGN: Carrier Grade Network Address Translator

- CPE: Consumer Premises Equipment, a device that resides between Internet service provider's network and consumers' home network.

- GGSN :Gateway GPRS Support Node

- GPRS: General Packet Radio Service

- PDN GW: Packet Data Network Gateway

- PDSN: Packet Data Serving Node

- PRR: Port Range Router

3. Provider Provisioned CPE: Overall Procedure

3.1. Introduction

As an alternative to the Double NAT solution, which suffers from several drawbacks, a second alternative is proposed within this document. The motivations for introducing this second alternative are as follows:

- Not to alter current (IPv4-based) services delivery and to not impact the introduction of future services;

- Avoid maintaining sessions states at the core network. Stateless solutions are privileged;

- Ease management functions (including provisioning, configuration operations, etc.);

- Optimise CAPEX and OPEX: As shown latter in this draft, the functional requirements to implement the proposed procedure are lightweight. Only slight modifications are required to be brought. Furthermore, the offered services are not impacted. Management practices would remain as today. For example, because the solution described in this memo does not handle dynamic NAT mappings (contrary to the CG-NAT), the planned maintenance operations (replacement of involved network equipment) would not impact the delivered services as a CG-NAT-based solution would do;

- Minor impact on routing and addressing architectures;

- Transparent to end-users: The same practices as today's ones will remain (e.g. Port forwarding on CPE still possible -provided the port is within the allocated Port Range-);

- Usability easiness;

- Facilitate functional separation (Service and Network): For instance, and unlike CG-NAT, the problem to run SIP-based services above a third party IP infrastructure would not be encountered with the proposed solution;

- Ease implementing legal requirements (optimize storage of legal data);

- Ease migration to a long term solution such as IPv6;

This section focuses only on the IPv4 variant of the solution. Other variants have been defined to integrate IPv6 and offer a global IP connectivity services including towards IPv6 realms in a stateless manner. Companion Internet Drafts will be submitted latter.

3.2. Basic Principles

The major idea is to assign the same IP address to several end-users' devices (e.g. Home Gateways (HGW) embedding NAT, but that could be other types of devices embedding NAT) and to constraint the (source) port numbers to be used by each device. In addition to the assigned IP address to access IP connectivity services, an additional parameter, called Port Mask, is also assigned to the customer's device. This mask indicates which Port Range is to be used by the customer's devices.

In the remaining part of this draft, the above mentioned public address is denoted as Primary IP Address.

For outbound communications, a given HGW proceeds to its classical operations except the constraint to control the source port number assignment so as to be within the Port Range assigned by its IP connectivity Service Provider. The traffic is then routed inside the Service Provider's domain and delivered to its final destination (within the service domain or to external domains).

For inbound communications (i.e. Towards customers attached to the Service Provider which has activated the procedure detailed in this memo), the traffic is trapped by a dedicated function called: Port Range Router (PRR). This function may be embedded in current routers or hosted by new nodes to be integrated in the IP infrastructure of these Service Providers. Appropriate routing tuning policies are enforced so as to drive the inbound traffic to cross a PRR (see Section 6.2 for more details). Particularly, each PRR correlate the Primary IP Address and information about the allowed port values with a specific identifier called: routing identifier (e.g. secondary IPv4 address, IPv6 address, point-to-point link identifier, etc). This routing identifier is used to route the packets to the suitable

device among all those owning the same IP address (See Section 6.1).

Note that for some reasons (e.g. Ease implementation of port-driven RPF (Reverse Path Forwarding) checking, anti-spoofing techniques, etc.), outbound traffic may be constrained to invoke the PRR function. This feature for outbound packets is considered as an engineering option. Service Providers are free to enforce it or not.

3.3. Applicability Use Cases

The following sub-sections provide a non exhaustive list of the port range solution applicability use cases. Other scenarios may be envisaged.

3.3.1. CPE

For deployment considerations and reduction of impact on terminals, the recommended scenario (in the context of DSL-type service offerings) of the deployment of the solution is a Provider provisioned CPE. This scenarios hides the connectivity solution and its associated addressing architecture. Machines behind the CPE continue to behave as today. No modification is required on end hosts.

3.3.2. End Host

When a host, which is capable of an IP address and a port range, but some of the applications on the host may have trouble using those addresses (e.g. they require a specific port to operate), as an implementation choice, the host may hide the port restricted nature of the allocated address by implementing an internal NAT as illustrated in the figure:

```
   Host
     +---------------------+
     +  Application        +
     +    |                +
     +    | IPv4p +------+  +  IPv4 address and a port range
     +    |-------| NAT |  +---------------------------------
     +            +------+  +
     +                     +
     +---------------------+
```

Figure 1: Internal NAT in a host

3.3.3. Point-to-point links with L2 IPv4 support

In point-to-point links it can be assumed that there are only two communicating parties on the link, and thus IP address collisions are easy to avoid.

In wireless cellular networks host attached to an access router, such as 3GPP PDN GW or WiMAX ASN GW , over a point-to-point link providing layer 2 IPv4 transport capability.

In order to be able to allocate an IP address together with a port range to a host, the access router needs to implement DHCP server or at least act as a DHCP relay or DHCP proxy , while a DHCP server exists in the backend. These setups are illustrated in the following figure.

```
                              +---------+      |
3GPP  ---Point to Point link--| 3GPP    |------|
host     <-IPv4 EPS bearer--> | PDN GW |       |
                              +---------+      |
                                               | IPv4 Internet
                              +---------+      |
WiMAX ---Point to Point link--| WiMAX   |------|
host     <-----IPv4 CS------> | ASN GW |       |
                              +---------+      |
```

Figure 2: Point-to-point physical links

As each host is attaching to the access router with an individual link, both modified and unmodified hosts can be supported simultaneously. This enables incremental deployment of modified hosts that are supporting public IPv4 address conservation by using DHCP to assign IPv4 address and a port range, while continuing to support the legacy hosts using DHCP as currently specified.

In this scenario, IPv6 addresses can be used in parallel with any IPv4 address, therefore no tunneling is necessary.

3.3.4. Point-to-point tunneled links

From DHCP point of view, tunneled link scenario does not differ very much from L2 point-to-point links as described in the previous section, although there are general concerns regarding tunnels (e.g. decreased MTU).

The tunnel is established between a host (or a CPE) and a tunnel endpoint in the host Operator's network. In different scenarios, the tunnel endpoint may be placed at different locations. The tunnel

endpoint can be at the first hop router such as 3GPP2 PDSN or 3GPP PDN-GW, or farther off in the network. In one scenario, the tunnel endpoint can be the CGN of DS-Lite [ID.durant].

These example setups are illustrated in the following figure.

```
                                Tunnel endpoints,
                               implementing DHCPv4
                               server/relay/proxy

                                 +-------------+
     Host ====IPv4 over IPv6==== | 3GPP2       |      |
          <---PPP & IPv6CP ----> | PDSN        |------|
              (point-to-point)   +-------------+      |
                                                      |
                                 +-------------+      |
     Host ====IPv4 over IPv6==== | 3GPP        |------| IPv4 Internet
          <--IPv6 PDP context--> | GGSN        |      |
             (point-to-point)    +-------------+      |
                                                      |
                                 +-------------+      |
     Host ====IPv4 over IPv6==== | IETF        |------|
          <---- IPv6-only -----> | DS-Lite CGN |      |
                network          +-------------+
```

Figure 3: Point-to-point links as IPv4 over IPv6 tunnels over three different accesses

Having the tunnel endpoint at the first hop router can be beneficial in setups where arrangement of native dual-stack transport for the last mile is not feasible or cost-effective approach. This can be the case e.g. in 3GPP networks, prior 3GPP Release-8, where a PDP context is capable of transporting only IPv4 or IPv6 packets, and for dual-stack access two parallel PDP contexts are required.

For networks which use IP(v6)CP to configure an IPv4 and IPv6 address to the host, allocating an IPv4 address and a port range to the host to prevent running out of available IPv4 addresses, can also be a feasible solution. In these deployment scenarios, IPv6CP would be used to configure an IPv6 address to the host. The host would then set up the tunnel and use the DHCPv4 extensions defined in this document to request an IPv4 address together with a port range. Examples of such networks include 3GPP2 and BRAS.

4. Retrieving IP Configuration Data

4.1. Assumption

In the context of this section, it is assumed that DHCP (Dynamic Host Configuration Protocol, [RFC2131]) is used to convey IP connectivity information. Other alternatives, such as PPP (Point-to-Point Protocol, [RFC1661]), may be used. The procedure described in this section is only an illustration example. It may be adapted so as to be able to apply in other technological contexts.

4.2. Procedure

4.2.1. Overview

At a bootstrapping phase, a given HGW issues a DHCP_DISCOVER message. This message is sent in broadcast. This message can be relayed by a DHCP Client Relay or be received directly by a DHCP Server. Once this message is received by a DHCP Server, this latter answers the requester by a dedicated DHCP_OFFER message containing a configuration offer.

The exchange which intervenes is illustrated in the following figure:

```
+-----+                      +-------------+
| HGW |                      | DHCP Server |
+-----+                      +-------------+
   |                                |
   |      (1) DHCP DISCOVER         |
   |------------------------------->|
   |                                |
   |                                |
   |       (2) DHCP OFFER           |
   |<-------------------------------|
   |                                |
```

Figure 4: DHCP Call Flow

A DHCP OFFER message encloses a set of IP-related information so as to access IP connectivity service. Particularly, it includes an IP address together with a new DHCP option, see: [ID.bajko].

Additional information may be included in the DHCP offer.

The use of Port Mask DHCP sub-option (similar to subnet mask) makes it possible to extend the notion of Port Range with non-continuous values, for the sake of flexibility.

A Port Range is then a set of ports that all have in common a subset of pre-positioned bits.

Once a Port Range information is received by a HGW, it constrains its NAT operations to the provisioned range. The number of customers to which an ISP can assign the same IP address depends on the number of allowed port numbers per user. Thus, if N bits are used to build the Port Mask, 2^N customers can be provided with the same IP address. For example: If N == 3, then the Service Provider multiplies by 8 its capacity in term of number of customers to which the connectivity service may be delivered.

In the remaining part, Port Mask and Port Range are used interchangeably.

5. Required Modifications

5.1. CPE

Above, we have quoted the case of Home Gateway but the solution can fit to any kind of Customer Premises Equipment (CPE).

In order to activate the aforementioned solution, slight modifications are required to be supported by CPEs. Concretely, CPEs MUST be able to constrain their NAT operations and to use only source port numbers within the allocated Port Range. If an IP packet is received by a given Port Range-enabled CPE, with a destination port number outside the assigned Port Range, the packet MUST be discarded.

Moreover, Port Range-enabled CPEs MUST be able to enforce configuration data received from the Service Providers so as to constrain its Port Range. More particularly, if DHCP is used to convey configuration data, a particular DHCP option (to be assigned by IANA) is to be supported by that CPE.

According to the enforced routing identifier mode, a de-encapsulation function MAY be required to be supported.

5.2. End-user Terminals

In some deployment scenarios (e.g. mobile), end-hosts should be updated. Concretely, end-hosts MUST be able to constrain their source port numbers and to use only source port numbers within the allocated Port Range. If an IP packet is received by a given Port Range-enabled end-user terminal, with a destination port number outside the assigned Port Range, the packet MUST be discarded.

Moreover, Port Range-enabled terminals MUST be able to enforce configuration data received from the Service Providers so as to constrain its Port Range.

According to the enforced routing identifier mode, a de-encapsulation function MAY be required to be supported.

5.3. Service Provider Infrastructure

The IP infrastructure of a given IP Service Provider is maintained slightly unchanged when deploying the Provider-Provisioned CPE solution. Only, a new function is introduced. This new function is denoted as PRR. This function is responsible for routing packets to the appropriate end-user's device among those to which the same IP address is assigned by the Service Provider. This operation is denoted as Port-Driven Routing operation since the destination IP address is not sufficient to handle routing operations and the information related to destination port is also required.

Except the PRR, all classical operations and practices remain as today's ones.

A PRR can be stand-alone server, or it can be hosted into other boxes such existing routers, PDN GW, ASN GW, etc.

5.4. DHCP Server Implementations

In case DHCP is used to convey IP connectivity information to customers' devices, DHCP server implementations may be modified accordingly. Indeed, DHCP server implementation should be modified so as to be able to support additional options such as Port Range DHCP option. The DHCP server assignment policy can be tuned by the Service Provider. A given Service Provider can provision its DHCP server with the Port Range to be allocated to end users' devices.

A second alternative to assign Port Ranges is described in Section 11. This alternative does not require any modification of the DHCP Server. Nevertheless, new changes are required to be supported by DHCP proxies.

6. Port Range Router

6.1. Main function

As stated above, the main function implemented by a PRR is a port-driven routing. In order to implement the port-driven routing, the following operations are achieved by a given PRR:

In order to implement the port-driven routing, the following operations are achieved by a given PRR:

1. It retrieves both destination IP address and destination port number.

2. Based on this couple, the PRR consults its binding table and retrieves the routing identifier.

Several modes may be envisaged to assign a routing identifier to be used as a deterministic discriminator to unambiguously identify a device among all those having the same IP address.

Hereafter are provided some implementation alternatives:

1. If a Secondary-IP address is used as the routing identifier: the PRR consults its binding table and retrieves the corresponding Secondary-IP address associated with a (Destination IP, Port Mask). Once retrieved, the PRR encapsulates the original packets in an IPv4 one with a destination IP address equal to Secondary-IP. This packet is then routed according to instantiated IGP (Interior Gateway Protocol) routes. Once received by the CPE, a de-encapsulation operation is achieved. The original packets is then treated and handled locally. If destination port of that packet is within the Port Range of that CPE, and depending on the local NAT implementation, the packet may be accepted and then proceed to classical NAT operation. Otherwise, the packet is dropped. Note that:

    A. The scope of the secondary address is limited to the access segment

    B. The secondary IP address may be an IPv6 address

2. Instead of encapsulation, and if source routing is supported, an explicit route is forced. A loose route is indicated in the packets. This loose route contains at least Secondary-IP. The routing of the resulting packet will be based on that address and not the destination one. The packet will be then received by the CPE with that Secondary-IP address. Then, the CPE will route the packet based on the final destination IP address. Since that address is also an IP address of that CPE, the packet is handled locally. The remaining operations are similar to the ones implemented by current CPEs.

3. If disjoint routes have been pre-installed so as to unambiguously identify the targeted device among all those having the same IP address, the PRR consults its binding tables and retrieves the index of the route corresponding to that (Destination IP, Port Mask) pair. The original packet is then sent over that route. Since the routes are disjoint, the packet will be received by the

targeted CPE. A example is the case where the PRR and the CPEs are directly linked by Ethernet, the route is then identified by the Ethernet MAC address of the CPE.

4. The routing identifier can also be the identifier of the L2 point-to-point link

As for inbound, a new operation is introduced in the path, this operation is a port-driven operation with no modification of the original packet. Further evaluation should be undertaken so as to assess the impact of this operation.

The performance experienced by outbound packets is not impacted since no alteration of the issued packets is to be enforced in the path. The experienced QoS (Quality of Service) is then the same as the currently deployed one.

6.2. Routing Considerations: Focus on IGP

A PRR is inserted in the inbound path in order to execute a port-driven routing. This constraint is translated into an IGP one. Indeed, a given PRR MUST advertise in IGP the primary IP addresses it handles. Doing so, all inbound packets will cross that PRR.

In case IPv4 Secondary-IP addresses are used to uniquely identify a CPE among all those having the same Primary-IP address, IPv4 Secondary-IP addresses MUST NOT be routable addresses inside core network. These addresses MUST NOT be reachable from the Internet. An example of the scope of those addresses is up to the frontier of an IP access POP (Point of Presence).

6.3. Binding Table

In order to implement port-driven routing operations, a PRR maintains a binding table which is a collection of entries correlating (IP address, Port Mask) with a routing identifier.

This table should not be confused with the NAT table as maintained by a CG-NAT.

6.4. Provisioning

6.4.1. Needs

In order to be able to treat received packets and then to proceed to port-driven routing, a PRR MUST be provisioned appropriately. Concretely, and as stated above, a given PRR needs to maintain a binding table which correlates a destination IP address and a Port

Mask with a routing identifier (such as a secondary IPv4 address, IPv6 address, routing index, MAC address, PPP session identifier, etc.). This binding table can be provisioned either by the Service Provider (owing to an internal interface) or by the CPE itself once IP connectivity information has been received from the service platform.

These two options are described hereafter. Service Providers are free to implement the option which meets its internal engineering policies.

6.4.2. Option 1: CPE-Provisioned PRR

Once its IP connectivity configuration is retrieved owing to a dedicated means such as DHCP, a given CPE enforces this new configuration. Particularly, the new received information may contain the following information:

{Primary-IP, Port Mask, Default_PRR, Routing Identifier}

In case the adopted method for the routing identifier (mentioned in Section 3.6.1) is a Secondary-IP address, a message is issued by the CPE towards its Default PRR. This message notifies that PRR about the new association: i.e. (Primary-IP, Port Mask) with Secondary-IP. This notification is achieved owing to a new message denoted as BIND. Once received by the PRR, an ACK message must be sent as response. If no ACK message is received, the CPE re-transmits its BIND message.

The procedure is sketched in the following figure:

```
+------+                        +------+
| HGW |                         | PRR |
+------+                        +------+
   |                               |
   |         (1) BIND              |
   |------------------------------>|
   |                               |
   |                               |
   |           (2) ACK             |
   |<------------------------------|
   |                               |
```

Figure 5: Example of CPE-provisioned PRR

6.4.3. Option 2: Provider-Provisioned PRR

Here, the provisioning of PRR binding table is undertaken by the Service Provider owing to the activation of appropriate management interfaces. These interfaces are internal to Service Provider's domain and are not visible to end-users. Exchanges between the PRR and the management realms are operated by the Service Provider. An implementation scenario of this option, is that once the DHCP server has assigned an IP address together with a Port Range a dedicated message is issued towards a PRR so as to instantiate a new entry in the binding table of that PRR. The entry can be refreshed or dropped once required.

In both options, the structure of the binding tables and the state machine of the PRR are identical.

7. Localization inside a Service Provider's domain

Each service Provider is free to adopt its internal policies for the deployment of PRRs. Nevertheless, we recommend deploying those nodes at access segment in order not to significantly impact end-to-end routing optimization. A PRR function can be embedded in an access router, a DSLAM, etc.

Several engineering options may be enforced:

- A given IP address is shared between several customers located in the same access POP: In this scenario, only access routers should be updated to support a PRR function. Doing so, communication (more precisely IGP routes) between the customers located in the same POP are optimised.

- Re-use the same IP address in several access POP and assign the same port range to all customers of the same POP: In this configuration, a given IP address is assigned to a single customer per POP. For intra-domain communications, and for optimisation purposes, all access routers should enable a PRR function. A far head router in the network should be activated to route inter POP traffic.

8. Fragmentation

In order to deliver a fragmented IP packet to its final destination (among those having the same IP address), the PRR should activated a dedicated procedure which described hereafter:

1. Check if the received packet is a fragment: ((MF == 1 && Fragment Offset == 0) || (Fragment Offset != 0)), else apply the classical PRR routing procedure;

2. Check if this fragment is the first one (MF == 1 && Fragment Offset == 0)

   2.1. In addition to the information retrieved to enforce port range routing, retrieve the source IP address and packet identifier. A fragmentation entry is instantiated. A timer (referred to as fragmentation timer) is associated with this entry. A clean up procedure is achieved when the timer expires.

   2.2. Retrieve the binding entry to be used to route this first fragment. A pointer to this entry is added to the fragmentation entry. A fragmentation entry includes: destination IP address, source IP address, Identifier, binding entry identifier and timer.

3. Check if this fragment is not the first one (Fragment Offset != 0)

   3.1. Retrieve the source IP address, destination IP address and Identifier;

   3.2 Check if an entry having the same source IP address, destination IP address and Identifier is instantiated in the fragmentation table

      3.2.1 If yes, retrieve the binding entry pointer from the fragmentation table. Use the corresponding entry to route the fragment.

      3.2.1 If not (fragments are not received in the order), launch a timer (which value is small than the fragmentation timer). This timer is referred to as fragmentation order timer. Upon expire of this timer, go to Step 3.2. This step is repeated two or three times. If it fails, the fragment is dropped.

Note that it is recommended to use a PMTUD path discovery mechanism (e.g. [RFC1191]).

Security issues related to fragmentation are out of scope of this document. For more details, refer to [RFC1858]

9. Multicast

In the previous sections, only unicast considerations have been elaborated. This section focuses on the impacts on multicast mechanisms and services when a Port Range based solution is activated.

Since the proposed solution does not require any modification on the core network of a given service provider / IP network provider, protocols to build and maintain multicast trees (e.g. PIM-SM [RFC4601], M-OSPF [RFC1584]) can be activated without any modification. Concretely, current multicast configurations on core routers and nodes can be applied without any adaptation.

As far as multicast group membership is concerned, classical procedures, e.g. IGMPv2 [RFC2236], or IGMPv3 [RFC3376], may be impacted.

Concretely:

1. If a secondary IP address (see Section 6.1) is used, the subscription to a multicast group can be done using this address. Thus, IGMP operations to receive traffic (i.e. IGMP requests) are not impacted and multicast traffic can be forwarded to the subscribed hosts;

2. If the shared IP address is used to issue IGMP requests,

    A. If distinct public IP addresses are assigned to each customer which device is attached to the same multicast router: classical IGMP operations are valid. No adaptation is to be enforced. Multicast traffic can be forwarded to each subscribed users without ambiguity.

    B. If a same public IP address is assigned to several customers which devices are attached to the same multicast router: the attached multicast router should correlate the request source with the binding table to unambiguously forward the multicast traffic to the appropriate subscribed user. More precisely, IGMP states should be updated to include the routing identifier to be used to forward traffic to the subscribed host. Appropriate means to uniquely distinguish the source of IGMP request among those having the same IP address should be implemented.

        + To avoid the modification of IGMP, several virtual router instances can be instantiated into the same physical node. Each virtual router manages only distinct IP addresses.

This configuration is similar to the bullet a.

In addition to these considerations, a hurdle can be encountered when using IGMPv3. Indeed, IGMPv3 messages can specify specific sources to be used to be excluded. If a shared IP address is assigned to those sources, traffic issued by other sources having the same IP address can be impacted. This scenario is not viable in current multicast deployments since the source of multicast traffic is under control of a service provider (e.g. head ends in the context of IP TV service offering) and a not shared IP address would be assigned to head ends.

10. IGD 2.0

Version 2.0 of IGP specification recommends the usage of a new method called AddAnyPortMapping() instead of AddPortMapping(). This new specification will ease the deployment of shared IP addresses.

New details will be added in the next version of the draft.

11. An alternative to avoid DHCP Server modifications

To avoid alteration of already in place DHCP servers, this section presents an alternative to implement Port Range assignment procedure. This alternative relies on DHCP Relay Clients or DHCP proxies and not on DHCP servers. These latter are kept unchanged. Their main function is to assign an available IP address. This address is assumed to be routed inside the Service Provider domain.

DHCP proxies, in cooperation with the PRR, maintains a set of pre-assignments based on a pre-provisioned Service Provider policy regarding how to build Port Ranges. As an example, if the implemented policy is to assign the same IP address to 4 customers, then 4 Port Ranges per IP address are statically built and then assigned to customers upon request.

In this context, DHCP proxies do not relay any IP assignment request until all available Port Ranges are allocated.

Figure 6 and Figure 7 provide an example of this option. In this example, CPE-1 and CPE-2 are two CPEs of two distinct customers.

CPE-1 sends first its DHCP DISCOVER message. This message is received by the DHCP proxy. Upon receipt, a lookup on available IP address and Port Range is achieved by the DHCP proxy.

Since no IP address is available, a DHCP DISCOVER message is forwarded to the DHCP Server. A DHCP OFFER is then sent back. This offer is trapped by the DHCP proxy.

The assigned IP address is retrieved and a pre-allocation of a Port Range is achieved. The offer is then updated with the Port Range Information and then relayed to CPE-1.

The remaining operations are the same operations as current DHCP exchanges.

```
   +-----+              +-----+           +--------+            +------+
   |CPE-1|              |DHCP |           |Binding |            |DHCP  |
   |     |              |proxy|           |Table   |            |Server|
   +-----+              +-----+           +--------+            +------+
      |  (1)DHCP DISCOVER  |                  |                    |
      |------------------->|                  |                    |
      |                    |(2) Check if there|                    |
      |                    |  is an available |                    |
      |                    |    IP @ and a    |                    |
      |                    |    Port Range    |                    |
      |                    |----------------->|                    |
      |                    |                  |                    |
      |                    |(3) No Available @|                    |
      |                    |                  |                    |
      |                    |<-----------------|                    |
      |                    |                  |                    |
      |                    | (4) DHCP DISCOVER|                    |
      |                    |-------------------------------------->|
      |                    |              (5) DHCP OFFER(IP-Pub-1) |
      |                    |<--------------------------------------|
      |                    | (6) DHCP REQUEST (IP-Pub-1)           |
      |                    |-------------------------------------->|
      |                    |              (7) DHCP ACK(IP-Pub-1)   |
      |                    |<--------------------------------------|
      |                    |                  |                    |
      |                    |(8)Add IP-Pub-1   |                    |
      |                    |  to Ports Range  |                    |
      |                    |    allocation,   |                    |
      |                    |and pre-assign a                       |
      |                    | Port Range to CPE1                    |
      |                    |----------------->|                    |
      |(9)DHCP OFFER(IP-Pub-1, PR1)           |                    |
      |<-------------------|                  |                    |
      |                    |                  |                    |
      |(10)DHCP REQUEST(IP-Pub-1, PR1)        |                    |
      |------------------->|                  |                    |
      |                    |(11) Assign PR1 to|                    |
      |                    |       CPE1       |                    |
      |                    |----------------->|                    |
      |(10)DHCP ACK(IP-Pub-1, PR1)            |                    |
      |------------------->|                  |                    |
      |                    |                  |                    |
```

Figure 6: First Example

If CPE-2 requests an IP address, it issues a DHCP DISCOVER message. This message is not relayed to the DHCP Server. A lookup request is executed by the DHCP proxy to check if an IP address and a Port Range

are available to be assigned. In this example, a positive answer is sent to the DHCP proxy. An Offer is then sent to CPE-2 as illustrated in Figure 7.

```
+-----+              +-----+           +---------+            +-------+
|CPE-2|              |DHCP |           |Binding |            |DHCP  |
|     |              |proxy|           |Table   |            |Server|
+-----+              +-----+           +---------+            +-------+
   |  (1)DHCP DISCOVER  |                  |                     |
   |------------------->|                  |                     |
   |                    |(2) Check if there|                     |
   |                    |  is an available |                     |
   |                    |    IP @ and a    |                     |
   |                    |    Port Range    |                     |
   |                    |----------------->|                     |
   |                    |                  |                     |
   |                    |(3) OK (IP1)      |                     |
   |                    |                  |                     |
   |                    |<-----------------|                     |
   |                    |                  |                     |
   |                    |                  |                     |
   |                    |(4) Allocate IP1  |                     |
   |                    |and Pre-assign a  |                     |
   |                    |Port Range to CPE2                      |
   |                    |----------------->|                     |
   |(9)DHCP OFFER(IP-Pub-1, PR2)           |                     |
   |<-------------------|                  |                     |
   |                    |                  |                     |
   |(10)DHCP REQUEST(IP-Pub-1, PR2)        |                     |
   |------------------->|                  |                     |
   |                    |(11) Assign PR2 to|                     |
   |                    |       CPE2       |                     |
   |                    |----------------->|                     |
   |(10)DHCP ACK(IP-Pub-1, PR2)            |                     |
   |------------------->|                  |                     |
   |                    |                  |                     |
```

Figure 7: Second Example

12. Comparison with CG-NAT

12.1. Generic Hurdles and Focus on Transparency to applications which enclose IPv4 address in their protocol messages

When deploying a Double NAT scenario, several hurdles will be encountered by Service Providers. Examples of these hurdles are as follows:

- End-users won't be able to configure their own port forwarding policies anymore, whilst with "Provider-Provisioned CPE" solution, the user can still configure port forwarding (provided the port is within the allowed range);

- Need to activate a second ALG (Application Level Gateway) at the core network for some applications (e.g. SIP (Session Initiation Protocol, [RFC3261]);

- Problems to run servers behind middleboxes with private addresses;

- Complication to enable inbound access;

- Performance issues (e.g. maintaining NAT entries by frequent (every 30s for instance) keep-alive messages is a real killer for battery powered devices);

- Interference between the service and network layers: The delivery of some services (e.g. SIP, DNS (Domain Name Service, [RFC1034]), and FTP (File Transfer Protocol, [RFC3659])) will require the knowledge of the underlying network engineering characteristics (i.e. Presence of intermediate CG-NAT boxes). If distinct administrative entities are managing the high-level services and the underlying IP infrastructure, critical problems for the current Internet business model will be raised.

Besides these generic hurdles, let's consider the ones that may arise when delivering SIP-based calls in the presence of CG-NAT boxes. Concretely, the following constraints should be followed:

- The SIP-based Service Provider should be aware about the underlying IP infrastructure so as to implement appropriate ALGs (Application Level Gateway). At least two modifications of SIP messages should be applied: The first one at the Home NAT and the second one at the CG-NAT. If no such ALG is enabled, no communication may be established. This constraint is heavy since it assumes that the same administrative entity administers both service and network infrastructures.

- NAT mapping entries at the CG-NAT should be maintained by keep-alive packets so as to be able to deliver incoming messages to customers' devices located behind the CG-NAT.

- Media flows may encounter some problems to be delivered since RTP (Real Time Transport Protocol, [RFC1889]) ports may not be opened.

The introduction of CG-NAT nodes may impact heavily the delivery of SIP-based services.

With a Port Range approach, nothing is changed with regard to the behavior of a today CPE with NAT: a SIP ALG can be quite easily implemented to take care of swapping the embedded IP address and port number in the messages to reflect the outbound IPv4 address and port of the CPE. On the contrary, running a SIP ALG instance inside the Carrier-Grade NAT for each SIP client may turn out to be very complex. Therefore, with the Port Range approach, SIP-based services are not altered compared to current practices when a CG-NAT is present in the path. The same mechanisms as today have to be deployed without any additional constraint nor impact.

Consequently, SIP-based services are not altered and complexity not increased.

12.2. Focus on Legal Storage

Most National Regulatory Authorities (NRA) require that ISPs provide the identity of a customer upon request of the authorities. This requirement is usually denoted as Legal Storage. In order to implement this requirement, Service Providers have deployed appropriate infrastructures including memory storage and interface to their Information Systems. Due to the continuous increase of traffic exchanged between end users, the amount of data stored by Service Providers would be also impacted if data relevant to all the sessions were to be stored. This is considered as a critical issue by Service Providers.

When deploying a new IP architecture or when modifying the currently deployed ones, Service Providers should be able to assess its impact on their Legal Storage infrastructures. Concretely, and because of the presence of NAPT function the knowledge of the source port number (simply referred to as port number), along with the source public IP address (simply referred to as public IP address), is mandatory to be able to retrieve the appropriate customer (or user) which is concerned by a given flow. This implies that all NAT mapping information is to be stored by a given ISP during the whole legal duration (one year in many countries).

Concretely, and because of the presence of NAPT function (in the CG-NAT), the knowledge of the source port number (simply referred to as port number), along with the source public IP address (simply referred to as public IP address), is mandatory to be able to retrieve the appropriate customer (or user) which is concerned by a given flow. This implies that all NAT mapping information is to be stored by a given ISP during the whole legal duration (one year in many countries).

When a CG-NAT is deployed, a given Service Provider must store legal

information of the mapped addresses in form of the following tuple:

{Public IP address - Public Port - Private IP address - Private port - protocol - date and hour of the beginning of address/port allocation - duration of this allocation (or date and hour of the allocation end)}.

Note that to actually find the identity of the appropriate customer which is concerned by a given IP flow, a given ISP must also store the mapping between the private IP address and the customer identification.

As for the Provider-Provisioned CPE approach, the required information to be stored is the following tuple (called in the remaining part tuple with Port Range):

{Public IP address - Port Range - protocol - customer identification - date and hour of the beginning of the Public IP address and Port Range allocation - duration of this allocation (or date and hour of the allocation end)}.

The length of this tuple with Port Range is about:

4 + 3 (2 for the Port Range pattern + 1 for the length) + 20 (customer identification) + 8 (date/time begin) + 8 (date/time end) = 43 bytes.

The Port Range is expected to be allocated for the same duration as the IP address, namely for a reasonable term (e.g. more than 24 hours conforming to current practices of IP address assignment). Thus, with regard to the nowadays situation, the additive information to be stored is only the Port Range.

The allocation of Public IP address and Port Range is expected to be made for a reasonable term (e.g. more than 24 hours) as the current practices for the assignment of IP addresses.

In order to illustrate the volume of required data to be stored by Service Providers,let's consider the following figures:

- o 1000 CPEs
- o 100 new sessions per 10 minutes per CPE (optimistic, it may be more)
- o each CPE traffics during 6 hour a day

o the public address and Ports Range change each day (changing these parameters may be even less frequent)

The amount of data to be stored per month when the Provider-Provisioned CPE approach is enabled (i.e. use of a Port Range) is around 1,3 Mbytes. The one for CG-NAT is around 3,1 Gbytes (Gbytes and not Mbytes) per month.

- Provider-Provisioned CPE:

Amount for 1000 CPEs per month = 1000 (CPEs) * 43 (bytes for the tuple with Port Range) * 30 (days in a month) = 1,3 Mbytes

-CG-NAT:

{Public IP address - Public Port - Private IP address - Private port - protocol - date and hour of the beginning of address/port allocation - duration of this allocation (or date and hour of the allocation end)}

= 4 + 2 + + 4 + 2 + 1 + 8 + 8

= 29 bytes.

Note : Storing the customer identification attached to the private address is considered negligible in the calculation.

Amount for 1000 CPEs per month

= 1000 (CPEs) * 100 (number of new sessions in 10 mn) * 36 (number of 10 mn durations in 6 h) * 29 (number of bytes per session) * 30 (days in a month)

= 3,1 Gbytes

Based on this data, a factor of more than 1000 is to be observed between the two solutions (in favor of the Port Range approach).

This factor (i.e. ratio of 1000) is important to be taken into account since CAPEX and OPEX would be impacted drastically for the implementation of this legal requirement. Indeed, a large investment must be forecast(ed) for deploying a suitable infrastructure (e.g. physical nodes and storage capacity). Service Providers should carefully consider this impact on their legal storage infrastructures.

Moreover, as the deployment of the FTTH (Fiber To The Home) will progress it is expected that the number of sessions per user will be

growing which will further increase the amount of data to be stored in CG-NAT but not in the Port Range approach.

12.3. Session Handling in CG-NAT

The complexity of the real-time processing is related to the number of operations to handle the TCP and UDP sessions and associated complexity.

CG-NAT is a NAT and therefore has to monitor dynamically all the sessions in order to identify if a public port number is still in-use or can be released. For this purpose, a CG-NAT needs in particular to handle timeouts and to scrutinize all TCP session states. In addition the entries enclosed in the NAT table maintained by a given CG-NAT is of a much greater complexity than the table in the PRR. The CG-NAT needs to keep all the mappings [Public IP address - Public Port - protocol - Private IP address - Private Port] for each session (UDP or TCP) whilst the PRR has to keep only one entry [Public IP address - Port Range - route to the CPE] per CPE.

For example, if the CPE handles 100 active sessions, the factor is 100 between a CG-NAT and a PRR. For a CPE with 1000 active sessions (which may not be so rare for clients making high use of peer to peer applications) the factor raises to 1000. Again, this is not simply a matter of factor; with CG-NAT, handling a session is complex as already indicated (e.g. timeouts, scrutinizing of session states, NAT entries real time maintenance, etc.).

As for the PRR, it does not handle sessions but simply routes packets (routing based on both IP address and Port Range).

CG-NAT can either be used in a context where the CPE keeps its NAT (yielding a double NAT configuration) or in a configuration where the CPE is a mere router (or bridge) without any NAT. In the first case (i.e. CPE without NAT) there is only one level of NAT in the path (at the CG-NAT level). All the complexity, today distributed among the CPEs, becomes concentrated into CG-NAT equipment. The cost of the CG-NAT is not balanced by a relative simplification of the CPEs (no NAT embedded). In a double NAT configuration the relative simplification of the CPE (no NAT embedded) is not even attained.

12.4. Peer-to-Peer applications

P2P applications can not work at full capabilities when a CG-NAT is in the path. This is because the peers can not initiate communications toward a peer behind a CG-NAT. Consequently the communications must pass through a server which greatly reduces the throughput capabilities of the system. A palliative could be for P2P

applications to use a STUN server so that they can know the public address and port allocated by the CG-NAT and to keep alive the port (by periodical short messages).

There is no such problem with the Port Range approach where the user can still as today set manually the port forwarding policies onto his CPE (e.g. Through WEB page, provided the choice of the port were restricted to the allocated Port Range, etc.).

13. IANA Considerations

TBC.

14. Security Considerations

This section will be completed in the next version of this draft.

15. Contributors

These authors have contributed to this memo:

- o Jean-Luc Grimault (France Telecom, jeanluc.grimault@orange-ftgroup.com)

- o Alain Villefranque (France Telecom, alain.villefranque@orange-ftgroup.com)

16. Acknowledgements

The authors would like to thank Dave THALER and Yoann NOISETTE, for their extensive review and technical input, and Mohammed KASSI LAHLOU for his suggestion regarding the involvement of the DHCP client relay. We would also like to thank Pierrick MORAND and Mohammed ACHEMLAL for their support and suggestions.

17. References

17.1. Normative References

[RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.

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

November 1990.

[RFC1584] Moy, J., "Multicast Extensions to OSPF", RFC 1584, March 1994.

[RFC1661] Simpson, W., "The Point-to-Point Protocol (PPP)", STD 51, RFC 1661, July 1994.

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

[RFC1889] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", RFC 1889, January 1996.

[RFC2026] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

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

[RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[RFC2236] Fenner, W., "Internet Group Management Protocol, Version 2", RFC 2236, November 1997.

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[RFC3659] Hethmon, P., "Extensions to FTP", RFC 3659, March 2007.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

17.2. Informative References

[ID.240space]
Fuller , V., Lear , E., and D. Meyer , "Reclassifying 240/4 as usable unicast address space", March 2008.

[ID.Eprivate]
Savolainen , T., "A way for a host to indicate support for 240.0.0.0/4 addresses", July 2008.

[ID.arkko]
Arkko, A. and M. Townsley, "IPv4 Run-Out and IPv4-IPv6 Co-Existence Scenarios", Work in progress: Internet Drafts draft-arkko-townsley-coexistence-00.txt, September 2008.

[ID.bajko]
Bajko, G., Savolainen , T., Boucadair, M., and P. Levis, "Dynamic Host Configuration Protocol (DHCP) Option for Port Range Assignment", January 2009.

[ID.durant]
Durand , A., Droms, R., Haberman, B., and J. Woodyatt, "Dual-stack lite broadband deployments post IPv4 exhaustion", Internet Draft, draft-durand-softwire-dual-stack-lite-01.txt (work in progress), November 2008.

Appendix A. Illustration Examples

In order to illustrate the procedure detailed above, let's consider the example illustrated Figure 8.

As shown in Figure 8, the same IP address 5.5.5.5 is assigned to the Home NAT of Phone-1, the one of Phone-2 and the one of Phone-3. Three port masks are also assigned to the three users. In this example, we assume that distinct Port Ranges are assigned to each HGW. For example, the Home NAT of Phone-1 can use a range of port numbers up to 8191, the Home NAT of Phone-2 a range of port numbers from 8192 to 16383 and the one of Phone-3 is from 16384 to 24575.

```
+-------+    +---+    +--------------+    +------------+
|       |    |   |    |              |    |            |
|Phone-1|----|HGW|----|              |    |            |
|       |    |   |    |              |    |            |
+-------+    +---+    |              |    |            |
10.0.0.1   5.5.5.5    |              |    |            |
                      |              |    |            |
+-------+    +---+    | Service      +----+ Internet   |
|       |    |   |    |              |    |            |
|Phone-2|----|HGW|----| Provider     |    |            |
|       |    |   |    |              |    |            |
+-------+    +---+    | Domain       |    |            |    +-------+
192.168.1.1 5.5.5.5   |              |    |            |----|Phone-4|
                      |              |    |            |    +       |
+-------+    +---+    |              |    |            |    +-------+
|       |    |   |    |      +-----+ |    |            | 25.25.25.28
|Phone-3+----+HGW+----+      | PRR | |    |            |
|       |    |   |    |      +-----+ |    |            |
+-------+    +---+    +--------------+    +------------+
10.0.45.25  5.5.5.5
```

Figure 8: Reference Architecture

A.1. Outbound communications

When Phone-1 issues an IP packet to Phone-4, the source IP address is equal to 10.0.0.1 and the source port number is 1234 (i.e. Packet Po1 represented in Figure 9).

Once received by the Home NAT, this latter proceeds to its NAT operations and assigns a port number in its provisioned range. In this example, a source port number 9123 is assigned. The packet (i.e. Po2 represented in Figure 9) is then routed until its final destination (i.e. Phone-4).

```
+-------+Po1 +---+Po2 +--------------+    +------------+
|       |===>|   |=======            |    |            |
|Phone-1+----+HGW+----+  \           |    |            |
|       |    |   |    |   \          |    |            |
+-------+    +---+    |    \         |    |            |
10.0.0.1   5.5.5.5    |     \        |    |            |
                      |      ====\   |    |            |
+-------+    +---+    | Service   \  +----+ Internet   |
|       |    |   |    |            \ |    |            |
|Phone-2+----+HGW+----+ Provider    \|    |            |
|       |    |   |    |              \    |            |    +-------+
+-------+    +---+    | Domain       |\====================>|       |
192.168.1.1 5.5.5.5   |              |    |            +----+Phone-4|
                      |              |    |            |    |       |
+-------+    +---+    |              |    |            |    +-------+
|       |    |   |    |      +-----+ |    |            | 25.25.25.28
|Phone-3+----+HGW+----+      | PRR | |    |            |
|       |    |   |    |      +-----+ |    |            |
+-------+    +---+    +--------------+    +------------+
10.0.45.25  5.5.5.5
```

Figure 9: Example of an Inbound Communication

A.2. Inbound communications

Phone-4 can send traffic to 5.5.5.5:9123 (i.e. Ultimately to Phone-1)(see Pi1 traffic of Figure 10). This traffic crosses the Port Range Router which proceeds to a port-driven routing.

Concretely, the PRR retrieves both destination IP address and destination port number from the received packet. Then, it checks its binding table and retrieves the suitable information (i.e. routing identifier) to route the packet towards the appropriate HGW. The initial packet is then routed (e.g. encapsulated towards a private address) and sent to that HGW using the retrieved routing identifier.

Packets are routed up to Home NAT of Phone-1 (see Pi2 traffic of Figure 10) which proceeds to a de-encapsulation operation. At this phase, it retrieves a packet destined to 5.5.5.5:9123. As a final step, it checks its mapping table in order to find which local IP address and port numbers are to be used. In this example, an entry exists: 10.0.0.1 and 1234 are returned and the packet is translated and routed to Phone-1.

All these operations are similar to classical NAT operations except the operations undertaken by the PRR and the conditioned port numbers assignment process in the HGW. This simple example does not take

into account IP addresses which may be involved inside the payload, i.e. those requiring ALG invocation.

```
   +-------+Pi3 +---+    +--------------+    +------------+
   |       |<===|   |<=\ |              |    |            |
   |Phone-1+----+HGW+--|-+              |    |            |
   |       |    |   |  | |              |    |            |
   +-------+    +---+  | |              |    |            |
   10.0.0.1   5.5.5.5  | |              |    |            |
                       | |              |    |            |
   +-------+    +---+  | | Service      +----+ Internet   |
   |       |    |   |  | |              |    |            |
   |Phone-2+----+HGW+--|-+ Provider     |    |            |
   |       |    |   |  | |              |    |    Pi1     |    +-------+
   +-------+    +---+  | | Domain       |/=====================|       |
   192.168.1.1 5.5.5.5  \|              /    |            +----+Phone-4|
                         \             /|    |            |    |       |
   +-------+    +---+    |\           / |    |            |    +-------+
   |       |    |   |    | \Pi2 +-----+ |    |            | 25.25.25.28
   |Phone-3+----+HGW+----+  \===|PRR  | |    |            |
   |       |    |   |    |      +-----+ |    |            |
   +-------+    +---+    +--------------+    +------------+
   10.0.45.25  5.5.5.5
```

Figure 10: Example of an Outbound Communication

Note that the paths shown in the figures above (i.e. Figure 9 and Figure 10) represent a functional invocation path of the PRR function and not real IP routes. Indeed, based on adopted PRR deployment strategy (e.g. PRR embedded in a DSLAM (Digital Subscriber Line Access Multiplexer), PRR embedded in an access router, a centralized PRR per access PoP (Point of Presence), etc.), IP routes may be symmetric or asymmetric at least at access segment.

Moreover, the PRR function may be embedded in an existing router or be hosted by a dedicated node.

Appendix B. Experimentation Results

B.1. Configuration

The main functionalities of the Provider-Provisioned CPE solution have been validated in a proof-of-concept testbed. The goal of this testbed is to assess the validity of the proposed solution and its ability to meet its objectives. Concretely, hereafter are listed two key functionalities which have been implemented:

1. The CPE restricts its source ports to be within its assigned Port Range. By the way, direct communications between two CPEs with the same IP address (but of course with distinct Port Ranges) must be effective and for that purpose the packets must pass through the PRR.

2. A PRR is positioned to be in the path of all inbound packets destined to a shared IP address. This PRR implements a port-driven routing as described in Section 6.

The features relative to the proposed new DHCP options (defined in [ID.bajko]) have not been part of the validation activities which aimed only at validating the fractional address concept and checking its transparency to well-known applications on Internet.

For both the CPEs and the PRR, Linux-based PCs have been used.

As shown in Figure 11, CPEs and the PRR are directly connected via Ethernet. As indicated in Section 6.1, other network configurations are possible. This choice is motivated by its simplicity in the scope of a proof-of-concept testbed.

```
+--------+          +--------+
|        |          |PC Linux|
|  user  |          |        |
|terminal+----Eth-INT|  CPE x |Eth-EXTx----+
|        |          |        |            |
+--------+          +--------+            |
                                          | +--------+    +--------+
                                          | |PC Linux|    |Internet|
                    IPv4_public addr      | |        |    |        |
                    set in NAPTs          | |        |    |        |
                     of CPEs              +-|  PRR   +----+        |
                                          | |        |    |        |
                                          | +--------+    +--------+
+--------+          +--------+            |
|        |          |PC Linux|            |
|  user  |          |        |            |
|terminal+----Eth-INT|  CPE y |Eth-EXTy----+
|        |          |        |
+--------+          +--------+
```

Figure 11: Testbed Configuration

B.2. On the CPE

As shown in Figure 11, each CPE has two Ethernet interfaces, each one being set-up with a private IPv4 address:

- Eth INT: interface towards the LAN the CPE serves (where the user terminal(s) and possibility server(s) lay). The private address [private addr INT CPE] is assigned to this interface.

- Eth EXT: interface towards the PRR; on this interface is set up the private address [private addr EXT CPE].

In the remaining parts of this section, [private addr EXT CPE-x] is used to refer to the private address [private addr EXT CPE] of CPE x.

To force each CPE to send all its outbound IP packets within the assigned Port Range, Netfilter features have been configured. This has consisted to configure through iptables commands the NAT already embedded in the Linux OS of each CPE, as follows (for CPE x):

```
/sbin/iptables -t nat -A POSTROUTING -p tcp -o [Eth_EXT] -j SNAT
--to-source [IPv4-pub1]:[ports-range-x]

-- the same line for UDP --
```

With:

- IPv4-pub1: the public address shared between the CPEs

- ports-range-x: the Port Range assigned to CPE x

Within this testbed, the CPE has none of its two interfaces set-up with the shared IP public address. This latter is ONLY present at the NAPT settings level.

B.3. On the PRR

To enforce a port-driven routing on PRR, Linux Netfilter capabilities have been used. The inbound packets are marked depending of their destination address and destination port. For example for CPE x, the following command is executed:

```
/sbin/iptables -t mangle -A PREROUTING -p tcp --destination [IPv4-
pub1] --dport [ports-range-x] -j MARK --set-mark [x]

-- the same line for UDP --
```

In the Linux Netfilter configuration, this [x] marking is associated

with a routing table dedicated for the [x] marked packets. This table contains only one entry: the one pointing to the private address of the CPE x (private addr EXT CPE-x).

Of course in the PRR, there is a mark setting line (as shown above) along with the corresponding routing table for each of the CPEs the PRR serves.

This kind of marking is purely at Netfilter level and stays within the Linux OS of PRR. It does not entail any marking of IP packets over Ethernet.

Therefore the operations at the PRR are quite simple. Upon an inbound packet coming at the outside interface of the PRR (e.g. Coming from Internet):

o At pre-routing level in the PRR, the packet is marked as shown above. The marking depends on the destination address and on the Port Range in which the destination port falls;

o Owing to this mark (i.e. [x] for CPE x), the packet passes through a routing table which points to the private address of the CPE x (private addr EXT CPE-x). This private address is seen by the PRR as the first hop of the route towards CPE x. The PRR proceeds to an ARP (Address Resolution Protocol) resolution (if not already achieved previously) and matches the [private addr EXT CPE-x] with the MAC (Media Access Control) address of EXT CPE x;

o Then, the packet is encapsulated into an Ethernet frame and transmitted to EXT CPEx.

Inbound packets have not been at all tempered by the PRR. Particularly, the destination IP address is always the shared public IP address of CPE x.

B.4. Main Results

This testbed has been used to conduct various tests. The objectives of those tests were to validate the concept of the Provider-Provisioned CPE solution, in particular its transparency to well-known applications. Indeed, the following applications have been selected and their behaviour evaluated: Web browsing (HTTP), FTP, Email, Instant messaging (two well-known applications have been used), Peer-to-Peer (again two well-known applications) and Voice over IP (an application which does not require an ALG on the CPE has been tested).

Obtained results confirmed the validity of the Provider-Provisioned

CPE solution: Web browsing (HTTP), Email and Instant messaging work normally and no degradation have been experienced.

For P2P (Peer-to-Peer) applications to be fully operational when launched inside terminals behind CPEs, we needed to manually set up a port forwarding at the CPE NAT. This is generally already the case today for users whose machines do not harness UPnP or whose CPE is not UPnP IGD (Internet Gateway Device) enabled. With a "Provider-Provisioned CPE" solution the CPE implementation would need to take care that manual port forwarding be only possible in the allocated Port Range (e.g. through Web settings menus slightly amended). As for UPnP, further considerations are needed to assess whether the future version of UPnP IGD can allow the CPE to allocate a port different from the one the terminal has requested.

For one P2P application tested, we found that two peers each behind a CPE sharing the same public address cannot download a SAME file from a source peer. The reason is certainly that the source peer relies on the IPv4 address and therefore considers the two downloading peers as a unique peer and does not accept parts of a file to be sent to the same peer over two different ports. Such limitation comes from the very principle of sharing an IP address (and not from the "Provider-Provisioned CPE" concept). We may think that other applications on Internet react in such way.

As for FTP, the passive mode works also well. Active mode does not but this is not because of the Provider-Provisioned CPE concept but only because the FTP active mode does not pass naturally well the NAPT (even a plain NAPT without Port Range restriction).

An FTP server has also been installed and launched on a PC behind a CPE. We set up manually port forwarding at the CPE NAT to allow inbound connections. A FTP client (on another machine) succeeded to connect normally to the FTP server provided the client specifies the address AND port of the server when launching the connection. This proves that the solution allows servers behind Port Range restricted CPEs. Further investigation may be undertaken such as using DynDNS and SRV records to retrieve the port number to be used for FTP service.

Tests were also made when the client and the server are each behind a CPE sharing the same address. Again that worked also (the communication passes through the PRR). That shows that in the context of CPEs sharing a same public address, there is solution for allowing communication between the CPEs (namely the shared address acting only at NAT level but not assigned to any interface).

B.5. Conclusion

The conclusion of this implementation is that the two key features of the "Provider-Provisioned CPE" solution (namely: Port Range restriction at CPE and port-driven routing at PRR) are already provided in Linux OS. It is expected that the necessary enhancements on other types of CPE plus the mechanism described in Section 5 should be rather simple modifications in the CPE. This is the same thing for the PRR: deriving a PRR from existing routing equipments should be rather simple. It may be even that, on some existing routers, policies based settings already implemented could perform the port-driven routing.

In addition, the various functional tests we have performed on the testbed have assessed completely the validity of the solution.

Authors' Addresses

Mohamed Boucadair (editor)
France Telecom
42 rue des Coutures
BP 6243
Caen Cedex 4 14066
France

Email: mohamed.boucadair@orange-ftgroup.com

Pierre Levis
France Telecom
42 rue des Coutures
BP 6243
Caen Cedex 4 14066
France

Email: pierre.levis@orange-ftgroup.com

Gabor Bajko
Nokia

Email: gabor.bajko@nokia.com

Teemu Savolainen
Nokia
Hermiankatu 12 D
FI-33720 TAMPERE
Finland

Email: teemu.Savolainen@nokia.com