Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2012

R. Penno
A. Durand
Juniper Networks
A. Clauberg
Deutsche Telekom AG
L. Hoffmann
Bouygues Telecom

March 11, 2012

Stateless DS-Lite
draft-penno-softwire-sdnat-02

Abstract

This memo define a simple stateless and deterministic mode of operating a carrier-grade NAT as a backward compatible evolution of DS-Lite.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 12, 2012.

Table of Contents

1. Introduction

DS-Lite [RFC6333],is a solution to deal with the IPv4 exhaustion problem once an IPv6 access network is deployed. It enables unmodified IPv4 application to access the IPv4 Internet over the IPv6 access network. In the DS-Lite architecture, global IPv4 addresses are shared among subscribers in the AFTR, acting as a Carrier-Grade NAT (CGN).

[I-D.ietf-softwire-public-4over6] extends the original DS-Lite model to offer a mode where the NAT function is performed in the CPE. This simplifies the AFTR operation as it does not have to perform the NAT function anymore, however, the flip side is that the address sharing function among subscribers was no longer available. [I-D.cui-softwire-b4-translated-ds-lite] introduces port restrictions, but does not completely specifies how the CPE acquires the information about its IPv4 address and its port range. More importantly, that draft does not explain how this solution can be deployed in a regular DS-Lite environment. This memo addresses these issues and clarifies the operation model.

Other approaches like variations of 4rd allows also for a full stateless operation of the decapsulation device. By introducing a strong coupling between the IPv6 address and the derived IPv4 address, they get rid of the per-subscriber state on the decapsulation devices. The approach take here argues that such per-subscriber state is not an issue as it is easily replicated among all decapsulation devices. Eliminating the strong coupling between IPv6 and IPv4 derived addresses, the approach presented here enables service providers a greater flexibility on how their limited pool of IPv4 addresses is managed. It also provide greater freedom on how IPv6 addresses are allocated, as sequential allocation is no longer a pre-requisite.

The approach presented here is stateless and deterministic. It is stateless is NAT bindings are maintained on the CPE, not on the AFTR. It is deterministic as no logs are required on the AFTR to identify which subscriber is using an external Ipv4 address and port.

The stateless DS-Lite architecture has the following characteristics:

- Backward compatible with DS-Lite. A mix of regular DS-Lite CPE and stateless DS-Lite CPEs can interoperate with a stateless DS-Lite AFTR.

- Zero log: Because the AFTR relies only on a per-subscriber mapping table that is reversible, the ISP does not need to keep any NAT binding logs.

- Stateless AFTR: There is no per-session state on the AFTRs. By leveraging this stateless and deterministic mode of operation, an ISP can deploy any number of AFTRs to provide redundancy and scalability at low cost. Because there is no per-flow state to maintain, AFTR can implement the functionality in hardware and perform it at high speed with low latency.

- Flexibility of operation: The ISP can add or remove addresses from the NAT pool without having to renumber the access network.

- Leverage IPv6: This stateless DS-Lite model leverage the IPv6 access network deployed by the ISPs.

2. Stateless DS-Lite CPE

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].

A Stateless DS-Lite CPE operates in similar fashion than a regular DS-Lite CPE, where the NAT function is re-introduced in CPE with a modification on how ports are managed.

2.1. Learning external IPv4 address

A stateless DS-Lite CPE MUST implement the DHCPv4 client relay option defined in [I-D.ietf-dhc-dhcpv4-over-ipv6] to learn is external IPv4 address. Other mechanism, such as manual configuration or TR69, MAY be implemented.

2.2. Learning external port range

A stateless DS-Lite CPE MUST implement the ICMP "port restricted" option defined later in this memo.

At boot time and later at intervals of 1h +/- a random number of seconds between 0 and 900), the stateless DS-Lite CPE MUST send packets with source port 0, source IPv4 address of the B4 element, destination IPv4 address 192.0.0.1 (the AFTR well-known IPv4 address) destination port 0, for each of the supported transport protocols (usually TCP and UDP). This will trigger an ICMP "port restricted" message from the AFTR.

After validating the content of the "ICMP port restricted" message, the stateless DS-Lite CPE MUST configure its port pool with it. If existing connections were using source ports outside of that range, the stateless DS-Lite CPE MUST terminate them.

2.3. Stateless DS-Lite CPE operation

The stateless DS-Lite CPE performs IPv4 NAPT from the internal RFC1918 addresses to the IPv4 address configured on the WAN interface, restricting its available ports to the range obtained as described above.

2.4. Host-based Stateless DS-Lite

Any host initiating directly a DS-Lite IPv4 over IPv6 tunnel can benefit from this techniques by implementing a 'virtual' stateless DS-Lite CPE function within its IP stack.

3. Stateless AFTR

3.1. Anycast IPv6 address for Stateless AFTR

All stateless AFTRs associated to a domain (or group of subscribers) will be configured with the same IPv6 address on the interface facing IPv6 subscribers. A route for that IPv6 address will be anycasted within the access network.

3.2. Stateless AFTR IPv4 address pool

All stateless AFTRs associated to a domain (or group of subscribers) MUST be configured with the same pool of global IPv4 addresses.

Routes to the pool of global IPv4 addresses configured on the stateless AFTRs will be anycasted by the relevant AFTRs within the ISP routing domain.

3.3. Stateless AFTR per-subscriber mapping table

Stateless AFTRs associated to a domain (or group of subscribers) MUST be configured with the same per-subscriber mapping table, associating the IPv6 address of the subscriber CPE to the external IPv4 address and port range provisioned for this subscriber.

Because the association IPv6 address --- IPv4 address + port range is not tied to a mathematical formula, the ISP maintains all flexibility to allocate independently IPv6 address and IPv4 addresses. In particular, IPv6 addresses do not have to be allocated sequentially and IPv4 resources can be modified freely.

```
+--------------+------------+----------+
| IPv6 address |IPv4 address|port-range|
+--------------+------------+----------+
|2001:db8::1   |   1.2.3.4  | 1000-1999|
|2001:db8::5   |   1.2.3.4  | 2000-2999|
|2001:db8::a:1 |   1.2.3.4  | 3000-3999|
+--------------+------------+----------+
```

Figure 1: Per-subscriber mapping table example

This per-subscriber mapping table can be implemented in various ways which details are out of scope for this memo. In its simplest form, it can be a static file that is replicated out-of-band on the AFTRs. In a more elaborated way, this table can be dynamically built using radius queries to a subscriber database.

3.4. Stateless AFTR decapsulation rules

Upstream IPv4 over IPv6 traffic will be decapsulated by the AFTR. The AFTR MUST check the outer IPv6 source address belongs to an identified subscriber and drop the traffic if not. The AFTR MUST then check the inner IPv4 header to make sure the IPv4 source address and ports are valid according to the per-subscriber mapping table.

If the inner IPv4 source address does not match the entry in the per-subscriber mapping table, the packet MUST be discarded and an ICMP 'administratively prohibited' message MAY be returned.

If the IPv4 source port number falls outside of the range allocated to the subscriber, the AFTR MUST discard the datagram and MUST send back an ICMP "port restricted" message to the IPv6 source address of the packet.

Fragmentation and reassembly is treated as in DS-Lite [RFC6333].

3.5. Stateless AFTR encapsulation rules

Downstream traffic is validated using the per-subscriber mapping table. Traffic that falls outside of the IPv4 address/port range entries in that table MUST be discarded. Validated traffic is then encapsulated in IPv6 and forwarded to the associated IPv6 address.

Fragmentation and reassembly is treated as in DS-Lite [RFC6333].

3.6. Redundancy and fail over

Because there is no per-flow state, upstream and downstream traffic can use any stateless AFTR.

3.7. SD-AFTR stateless domain

Using the DHCPv6 DS-Lite tunnel-end-point option, groups of subscribers and can be associated to a different stateless AFTR domain. That can allow for differentiated level of services, e.g. number of ports per customer device, QoS, bandwidth, value added services,...

4. Backward compatibility with DS-Lite

A number of service providers are, or are in the process of, deploying DS-Lite in their network. They are interested in evolving their design toward a stateless model. Backward compatibility is a critical issue, as, from an operational perspective, it is difficult to get all CPEs evolve at the same time.

So AFTRs have to be ready to service CPEs that are pure DS-Lite, some that are implementing only DHCPv4 over IPv6 and handle the NAT on the full IPv4 address themselves and some that also implement port restrictions via the ICMP message described here. For this reason, a AFTR operating in backward compatibility mode MAY decide to re-NAT upstream packets which source port number do not fall into the predefined range instead of simply dropping the packets.

The operating model is the following:

- Stateless DS-Lite: for CPEs that pre-NAT and pre-shape the source port space into the range assigned to the subscriber: decapsulate, check per-subscriber mapping, forward.

- B4-translated DS-Lite: for CPEs that performs NAT before encapsulation and are allocated a full IPv4 address: decapsulate, check per-subscriber mapping, forward.

- Re-shaper DS-Lite: for CPEs that pre NAT but fail to restrict the source ports: decapsulate, check per-subscriber mapping, re-NAT statefully the packets into the restricted port range, mark range as 'stateful', forward.

- Regular DS-Lite: for regular DS-Lite CPEs that do not pre-NAT: decapsulate, NAT statefully, forward.

In such a backward compatibility mode, the AFTR is only operating statelessly for the stateless DS-Lite CPEs. It needs to maintain per-flow state for the regular DS-Lite CPEs and the non-ICMP port restricted compliant CPEs. In this legacy mode where per-flow state is required, the simple anycast-based fail-over mechanism is no longer available.

5. ICMP port restricted message

Note: this section may end-up being a separate Internet draft.

5.1. Introduction

In the framework of A+P RFC 6346 [RFC6346], sources may be restricted to use only a subset of the port range of a transport protocol associated with an IPv4 address. When that source transmit a packet with a source outside of the pre-authorized range, the upstream NAT will drop the packet and use the ICMP message defined here to inform the source of the actual port range allocated.

This memo defines such ICMP messages for TCP and UDP and leaves the definition of the ICMP option for other transport protocol for future work.

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].

5.2. Source port restricted ICMP

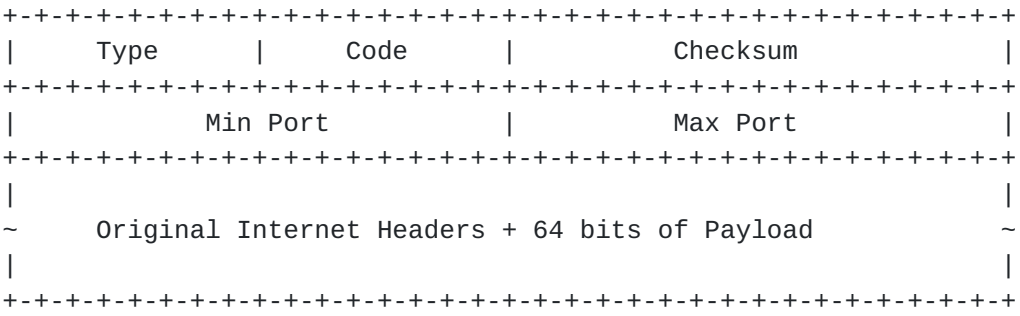

Figure 2: Source Port Restricted ICMP

Type: TBD for Source Port Restricted

Checksum: The checksum is the 16-bit ones's complement of the one's complement sum of the ICMP message starting with the ICMP Type. For

computing the checksum , the checksum field should be zero. This checksum may be replaced in the future.

Code: 6 for TCP, 17 for UDP

Min Port: The lowest port number allocated for that source.

Max Port: The highest port number allocated for that source.

5.3. Host behavior

A host receiving an ICMP type TBD message for a given transport protocol SHOULD NOT send packets sourced by the IP address(es) corresponding to the interface that received that ICMP message with source ports outside of the range specified for the given transport protocol.

Packets sourced with port numbers outside of the restricted range MAY be dropped or NATed upstream to fit within the restricted range.

A host MUST NOT take port restriction information applying to a given IP address and transport protocol and applies it to other IP addresses on other interfaces and/or other transport protocols.

If Min Port = 0 and Max Port = 65535, it indicates that the entire port range for the given transport protocol is available. If such 'full range' messages are received for all transport protocols, the host can take this as an indication that its IP address is probably not shared with other devices.

In order to mitigate possible man in the middle attacks, a host MUST discard ICMP type TBD messages if the associated port range (Max Port - Min Port) is lower than 64.

6. IANA Considerations

IANA is to allocated a code point for this ICMP message type.

7. Security Considerations

This ICMP message type has the same security properties as other ICMP messages such as Redirect or Destination Unreachable. A man-in-the-middle attack can be mounted to create a DOS attack on the source. Ingress filtering on network boundary can mitigate such attacks. However, in case such filtering measures are not enough, the additional provision that a host MUST discard such ICMP message with

a port range smaller than 64 can mitigate even further such attacks.

As described in [RFC6269], with any fixed size address sharing techniques, port randomization is achieved with a smaller entropy.

Recommendations listed in [RFC6302] applies.

8. References

8.1. Normative references

[I-D.ietf-dhc-dhcpv4-over-ipv6]
Lemon, T., Cui, Y., Wu, P., and J. Wu, "DHCPv4 over IPv6 Transport", draft-ietf-dhc-dhcpv4-over-ipv6-00 (work in progress), November 2011.

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

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

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

8.2. Informative references

[I-D.cui-softwire-b4-translated-ds-lite]
Boucadair, M., Sun, Q., Tsou, T., Lee, Y., and Y. Cui, "Lightweight 4over6: An Extension to DS-Lite Architecture", draft-cui-softwire-b4-translated-ds-lite-05 (work in progress), February 2012.

[I-D.ietf-pcp-base]
Cheshire, S., Boucadair, M., Selkirk, P., Wing, D., and R. Penno, "Port Control Protocol (PCP)", draft-ietf-pcp-base-23 (work in progress), February 2012.

[I-D.ietf-softwire-public-4over6]
Cui, Y., Wu, J., Wu, P., Metz, C., Vautrin, O., and Y. Lee, "Public IPv4 over Access IPv6 Network", draft-ietf-softwire-public-4over6-00 (work in progress), September 2011.

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

June 2011.

[RFC6302] Durand, A., Gashinsky, I., Lee, D., and S. Sheppard, "Logging Recommendations for Internet-Facing Servers", BCP 162, RFC 6302, June 2011.

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

Authors' Addresses

Reinaldo Penno
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089-1206
USA

Email: rpenno@juniper.net

Alain Durand
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089-1206
USA

Email: adurand@juniper.net

Alex Clauberg
Deutsche Telekom AG
GTN-FM4
Landgrabenweg 151
Bonn, CA 53227
Germany

Email: axel.clauberg@telekom.de

Lionel Hoffmann
Bouygues Telecom
TECHNOPOLE
13/15 Avenue du Marechal Juin
Meudon 92360
France

Email: lhoffman@bouyguestelecom.fr