

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 15, 2014

V. Bajpai  
J. Schoenwaelder  
Jacobs University Bremen  
July 14, 2013

Measuring the Effects of Happy Eyeballs  
draft-bajpai-happy-01.txt

## Abstract

The IETF has developed solutions that promote a healthy IPv4 and IPv6 co-existence. The happy eyeballs algorithm for instance, provides recommendations to application developers to help prevent bad user experience in situations where IPv6 connectivity is broken. This document describes a metric used to measure the effects of the happy eyeballs algorithm. The insights uncovered by analysing the data from multiple locations is discussed.

## 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 January 15, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. IPv6 Upgrade Policy . . . . .	2
3. Happy Eyeballs . . . . .	3
4. Related Work . . . . .	3
5. Metric . . . . .	4
6. Implementation . . . . .	4
7. Measurement Trials . . . . .	5
8. Data Analysis Insights . . . . .	5
9. Conclusions . . . . .	7
10. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

The function `getaddrinfo(...)` resolves a service name to a list of endpoints in an order that prioritizes an IPv6-upgrade path [RFC6724]. The order can dramatically reduce the application's responsiveness when IPv6 connectivity is broken. The degraded user experience can be subverted by implementing the happy eyeballs algorithm [RFC6555]. The algorithm recommends that a host, after resolving the service name, tries a `TCP connect(...)` to the first endpoint. However, instead of waiting for a timeout, it waits for 300ms, after which it must initiate another `TCP connect(...)` to an endpoint with a different address family and start a competition to pick the one that completes first.

This document describes a metric used to measure the effects of the happy eyeballs algorithm. The insights uncovered by analysing the data from multiple locations is discussed.

## 2. IPv6 Upgrade Policy

The happy eyeballs algorithm as defined in [RFC6555] biases its path selection in favor of IPv6 by design. The connection establishment race has been handicapped for the following reasons:

- o Carrier-grade NATs (CGNs) establish a binding for each connection request. Dual-stack hosts by preferring IPv6 connection routes, reduce their contention towards the critical IPv4 address space.
- o The IPv4 traffic may be billed by Operation Support Systems (OSS) in some networks. Techniques that help move this traffic to IPv6 networks reduce costs.

- o Middleboxes maintain state for each incoming connection request. If the dual-stacked hosts prefer IPv6 path, the load on load balancers and peering links reduces automatically. This reduces the investment on IPv4, and encourages IPv6 migration.

### 3. Happy Eyeballs

The happy eyeballs algorithm defined in [RFC6555] honors the IPv6 upgrade policy. It is therefore not designed to encourage aggressive connection requests over IPv4 and IPv6, but instead to satisfy the following goals:

- o The connection requests must be made in an order that honors the destination-address selection policy as defined in [RFC6724], unless overridden by user or network configuration. The client must prefer IPv6 over IPv4 whenever the policy is not known.
- o The connection initiation must quickly fallback to IPv4 to reduce the wait times for a dual-stack host in situations where the IPv6 path is broken.
- o The network path and destination servers must not be thrashed by mere doubling of traffic by making simultaneous connection requests over IPv4 and IPv6. The connection requests over IPv6 must be given a fair chance to succeed to reduce load on IPv4, before a connection over IPv4 is attempted.

However, applications on top of TCP will not be happy eyeballed only in scenarios where IPv6 connectivity is broken, but also in scenarios where the dual-stack host enjoys comparable IPv6 connectivity. We want to measure how much imposition does such a user experience in reality by measuring the effects of the happy eyeballs timer value.

The recommended timer value is 150-250ms [RFC6555]. However, Chrome uses 300ms. Firefox appears to be using 250ms while an early open-source implementation of happy eyeballs seems to recommend 100ms [Perreault]. We want to affirm the right value by measuring TCP connection establishment times experienced by dual-stacked hosts in real environments over IPv4 and IPv6.

### 4. Related Work

Fred Baker in [RFC6556] describes metrics and testbed configurations to measure how quickly an application can reliably establish connections from a dual-stacked environment. The metrics measure whether the communication establishment time is same regardless of the address family and the routing viability available to a dual-stacked host. The metrics defined in [RFC6556] is different in three ways:

- o DNS is accounted in connection establishment time. Our metric does not take this into account. Accounting DNS resolution may invite multiple input factors (slow resolvers) that may bias our TCP connection establishment time results. In addition, according to [RFC6555], the 300ms advantage applies to the first address family after the `getaddrinfo(...)` call. From a programming perspective, an application calls `getaddrinfo(...)` and that does its job, regardless of which address family is used.
- o The testbed configuration in [RFC6556] is more passive than active. An external analyser is used to passively observe the client's traffic using `tcpdump`. There is no active measurement test, instead the routers along the path are configured to control what connectivity route is taken. We on the other hand, have an active measurement test running on the client. The test is agnostic to network path configuration since it independently tries a TCP connection to each connectivity route. It also actively measures the time taken instead of relying on an external analyser program.
- o The testbed setup in [RFC6556] is designed for a controlled environment. The router in the path is configured to disrupt all but one routes to control the prefix used in the connection. As such, the test is repeated N times with different router configurations to try all possible permutations of route connectivity. Our measurement test is agnostic to the network path and does not require path configuration changes.

## 5. Metric

We have defined a metric that uses the TCP connection establishment times as a parameter to measure the algorithm's effects. The methodology also helps examine the impact of tunneling mechanisms employed by early adopters. The input parameter of the metric is a (IP address, port number) tuple and the output is the connection establishment time, typically measured in microseconds.

## 6. Implementation

We have developed happy, a simple TCP happy eyeballs probing tool that conforms to the definition of our metric. It uses non-blocking connect(...) calls to concurrently establish connections to all endpoints of a service and measures the elapsed time. The tool enforces a small delay between concurrent connect(...) calls to avoid bursty TCP SYN traffic. The initially performed service name resolution is not accounted in the connection establishment elapsed time.

## 7. Measurement Trials

We use Alexa's top 1M service names as input to prepare a top 100 dual-stacked service names list. We run happy on our internal test-bed of multiple measurement agents with different flavors of connectivity ranging from native IPv4, native IPv6, IPv6 tunnel broker endpoints, Teredo and tunnelled IPv4. The list of Measurement Agents (MAs) is shown in Table 1.

MA #	IPv4 AS	IPv6 AS	City	Country	Platform
1	AS680	AS680	Bremen	Germany	Mac OS X
2	AS680	AS680	Braunschweig	Germany	GNU/Linux
3	AS13237	Teredo	Berlin	Germany	GNU/Linux
4	AS31334	AS6939	Bremen	Germany	OpenWrt
5	AS680	AS680	Bremen	Germany	SamKnows
6	AS31334	AS6939	Bremen	Germany	SamKnows
7	AS24956	AS24956	Braunschweig	Germany	SamKnows
8	AS3320	AS3320	Bremen	Germany	SamKnows
9	AS5607	AS5607	London	England	SamKnows
10	AS3269	AS3269	Torino	Italy	SamKnows
11	AS8903	AS8903	Madrid	Spain	SamKnows
12	AS2614	AS2614	Timisoara	Romania	SamKnows
13	AS13030	AS13030	Olten	Switzerland	SamKnows
14	AS2856	AS2856	Ipswich	England	SamKnows

Table 1: A List of Measurement Agents (MAs)

## 8. Data Analysis Insights

The initial results show higher connection times and variations over IPv6 as shown in Figure 1. The services themselves may not be comparable amongst one another due to the sheer nature of different routing paths traversed by the packets.

```

1e+06  +-+-----+-----+-----+-----+-----+-----+-----+-----+
        +mean (v4)  +-----+-----+-----+-----+-----+-----+-----+

```

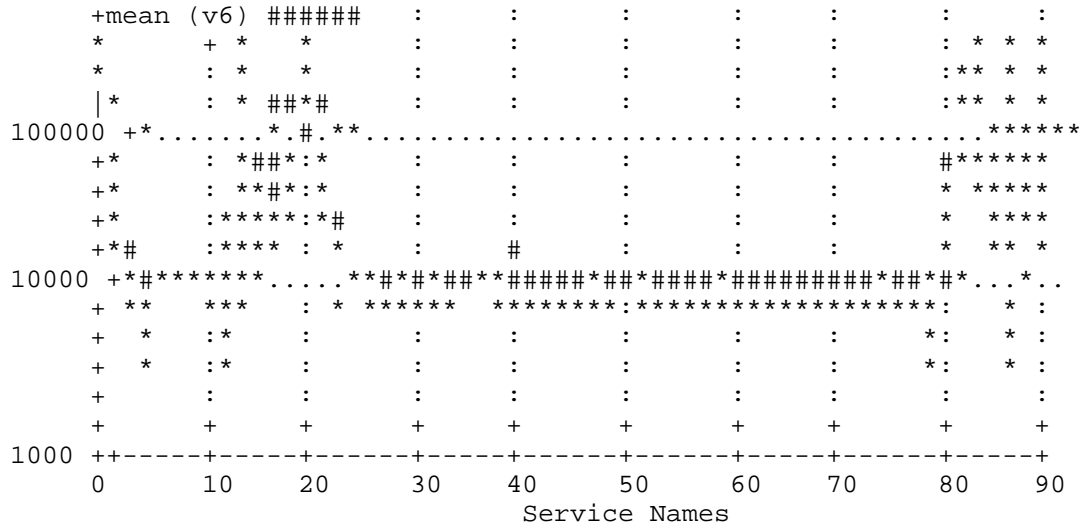


Figure 1: service vs {mean\_v4, mean\_v6}: samsbox1 (30 days, 300ms)

Fig. 1. shows the average TCP connection establishment times for both IPv4 and IPv6. The Measurement Agent (MA) is a SamKnows probe connected at Jacobs University Bremen. It receives IPv4 and IPv6 connectivity via German Research Network (DFN) [AS 680]. A PDF rendering of the plot is available at [mean].

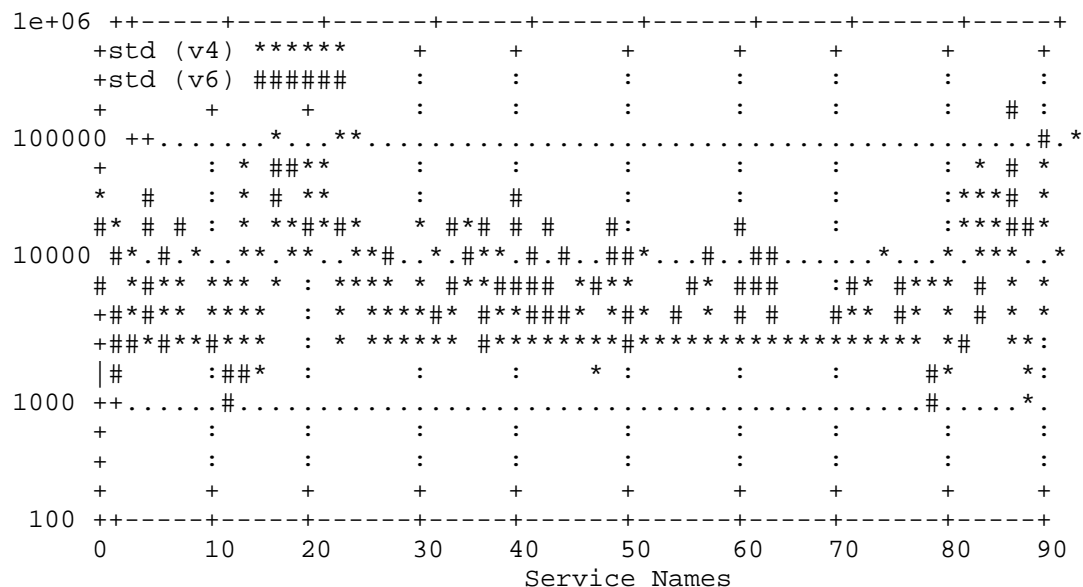


Figure 2: service vs {std\_v4, std\_v6}: samsbox1 (30 days, 300ms)

Figure 2 shows the standard deviation of the TCP connection establishment times for both IPv4 and IPv6. The Measurement Agent (MA) is a SamKnows probe connected at Jacobs University Bremen. It receives IPv4 and IPv6 connectivity via German Research Network (DFN) [AS 680]. A PDF rendering of the plot is available at [std].

It appears that an application never uses IPv6 using Teredo except in situations where IPv4 reachability of the destination service is broken. We noticed, that a 300ms advantage leaves a dual-stacked host only 1% chance to prefer a IPv4 route even though it may be significantly faster than IPv6. We also measured the margin by which happy eyeballs is inhibiting the fastest available route by comparing the slowness of a happy eyeballed winner to that of the loser.

## 9. Conclusions

We have performed a preliminary study on measuring the effects of happy eyeballs. We noticed several cases where the algorithm does not select the best route and instead hampers the user experience. We are working towards running this test on a large-scale measurement platform to develop a more comprehensive picture to help improve the algorithm.

## 10. Informative References

- [I-D.ietf-6man-addr-select-opt]  
Matsumoto, A., Fujisaki, T., and T. Chown, "Distributing Address Selection Policy using DHCPv6", draft-ietf-6man-addr-select-opt-10 (work in progress), April 2013.
- [Perreault]  
Perreault, S., "Happy Eyeballs in Erlang", July 2013, <[http://www.viagenie.ca/news/index.html#happy\\_eyeballs\\_erlang](http://www.viagenie.ca/news/index.html#happy_eyeballs_erlang)>.
- [RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.
- [RFC6556] Baker, F., "Testing Eyeball Happiness", RFC 6556, April 2012.
- [RFC6724] Thaler, D., Draves, R., Matsumoto, A., and T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, September 2012.
- [mean] Bajpai, V., "IPv4 and IPv6 Average Connection Establishment Times", July 2013, <<http://cn.ds.eecs.jacobs-university.de/users/vbajpai/ietf87-v6ops/samsbox1-mean.pdf>>.
- [std] Bajpai, V., "IPv4 and IPv6 Connection Establishment Times Variations", July 2013, <<http://cn.ds.eecs.jacobs-university.de/users/vbajpai/ietf87-v6ops/samsbox1-std.pdf>>.

#### Authors' Addresses

Vaibhav Bajpai  
Jacobs University Bremen  
Campus Ring 1  
28759 Bremen  
Germany

Phone: +49 421 200 3112  
Email: [v.bajpai@jacobs-university.de](mailto:v.bajpai@jacobs-university.de)



Juergen Schoenwaelder  
Jacobs University Bremen  
Campus Ring 1  
28759 Bremen  
Germany

Phone: +49 421 200 3587  
Email: [j.schoenwaelder@jacobs-university.de](mailto:j.schoenwaelder@jacobs-university.de)

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: May 08, 2014

G. Chen  
H. Deng  
China Mobile  
D. Michaud  
Rogers  
J. Korhonen  
Renesas Mobile  
M. Boucadair  
France Telecom  
A. Vizdal  
Deutsche Telekom AG  
C. Byrne  
T-Mobile USA  
November 04, 2013

IPv6 Roaming Behavior Analysis  
draft-chen-v6ops-ipv6-roaming-analysis-02

Abstract

This document intends to enumerate failure cases when a IPv6 subscriber roams into visited network areas. The investigations on those failed cases reveal the causes in order to notice improper configurations, equipment's incomplete functions or inconsistent IPv6 strategy.

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 May 08, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Roaming Architecture Descriptions . . . . .	3
3. Roaming Scenario Overview . . . . .	4
4. Failure Cases Descriptions . . . . .	5
4.1. Failure Case 1: Incompatible with Extended PDP/PDN Type .	5
4.2. Failure Case 2: Splitting Dual-stack Bearer . . . . .	6
4.3. Failure Case 3: Shortage of IPv6 support . . . . .	7
4.4. Failure Case 4: Fallback Incapability . . . . .	7
4.5. Failure Case 5: 464xlat Support . . . . .	7
5. Discussions . . . . .	8
6. IANA Considerations . . . . .	9
7. Security Considerations . . . . .	9
8. Acknowledgements . . . . .	9
9. References . . . . .	9
9.1. Normative References . . . . .	9
9.2. Informative References . . . . .	10
Authors' Addresses . . . . .	11

## 1. Introduction

IPv6 has been deployed globally to overcome the IPv4 depletion. Operators likely start or plan to upgrade the networks that allow IPv6 subscribers to access. As the dramatical uses of Internet services with a mobile access, IPv6 is an essential part to be considered in the mobile network evolution. 3rd Generation Partnership Project (3GPP) published the IPv6 migration guidance [TR23.975], which describes different technical evolution paths. In general, operators may deploy dual-stack or IPv6 single-stack depending on network's conditions. It has been observed that those deployments are rolled out in multiple provisioning domains. In the early IPv6 stage, a mobile subscriber roaming around the different areas may experience service degradations or interruptions due to the inconsistent configurations and incomplete functions in the networks nodes. This memo intends to document the observed failed cases and analyze the causes. It's expected that operators could notice the issues and prevent potential risks.

## 2. Roaming Architecture Descriptions

The roaming process could be triggered in the following scenarios:

- o International roaming: a mobile subscriber may entry a visited network, where different PLMN identity is used. The subscribers could either in an automatic mode or a manual mode to attach a PLMN cell.
- o Intra-PLMN mobility: a subscriber moves to a visited network as that of the Home Public Land Mobile Network (HPLMN). However, the subscriber profiles may not be stored in the area. Once the subscriber attaches to the network, the subscriber profile should be extracted from the home network for the network registration.

When a mobile device is turned on or is transferred via a handover to a visited network, the mobile device will scan all radio channels and find available Public Land Mobile Networks (PLMNs) to attach. Serving GPRS Support Node (SGSN) or Mobility Management Entity (MME) in the visited networks must contact the home Home Location Register(HLR) or Home Subscriber Server(HSS) and obtain the subscriber profile. Once the authentication and registration process is completed, the PDP activation and traffic flows may be operated differently according to the subscriber data configuration. Two modes have been shown at the figure to illustrate, that are "Home routed traffic" and "Local breakout".

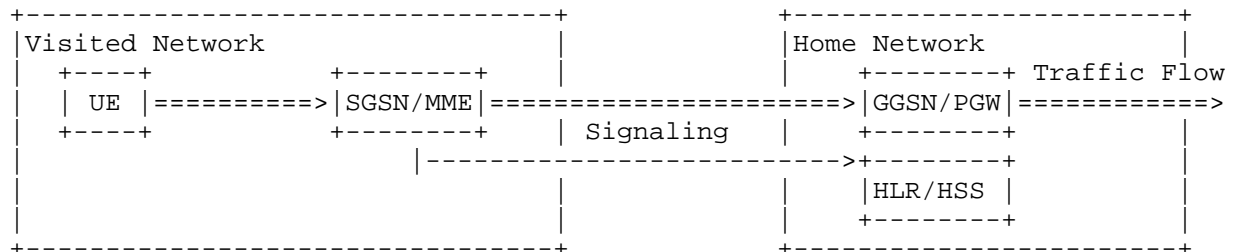
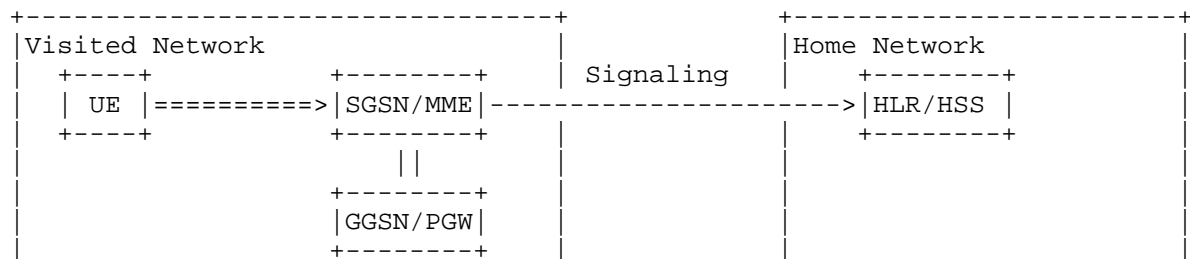


Figure 1: Home Routed Traffic



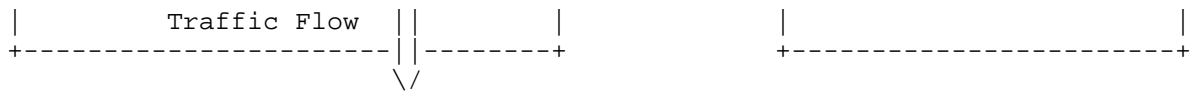


Figure2: Local Breakout

In the home routed mode, subscribers will activate the PDP/PDN context and get address from the home network. All traffic would be routed back to the home networks. That is the default case for an international roaming except for the IP Multimedia Subsystem (IMS) scenario.

In the local breakout mode, the subscriber address will be assigned from the visited network. The traffic flow would directly offloaded locally at a network node close to that device's point of attachment in the visited networks. Therefore, more efficient route would be achieved. The following will describe the cases where there is local breakout mode adopted.

- o Operators may add the APN-OI-Replacement flag defined in 3GPP [TS29.272] into user's subscription-data. The visited network indicates a local domain name to replace the user requested Access Point Name (APN). As the consequence, the traffic would be steered to the visited network. Those functions are normally deployed for the Intra-PLMN mobility cases.
- o Operators could also configure VPLMN-Dynamic-Address-Allowed flag[TS29.272] in the user profile to enable local breakout mode in Visited Public Land Mobile Networks (VPLMNs).
- o 3GPP specified Selected IP Traffic Offload (SIPTO) function[TS23.401] since Release 10 in order to get efficient route paths. It enables an operator to offload certain types of traffic at a network node close to that device's point of attachment to the access network.
- o GSMA has defined RAVEL[IR.65] as IMS international roaming architecture. Local breakout mode has been adopted for the roaming architecture.

### 3. Roaming Scenario Overview

3GPP specified three types of Packet Data Protocol (PDP)/Packet Data Networks (PDN) to describe each connection, i.e. PDP/PDN Type IPv4, PDP/PDN Type IPv6 and PDP/PDN Type IPv4v6. User devices can be set to request a particular PDP/PDN Type. Those PDP/PDN types should also be restored in Home Subscriber Server (HSS) as a part of subscriber profile, as defined in [TS29.272]. When a subscriber

roams to a visited network, the new visited network notices that it is not registered with its own system, and attempts to identify its home network. Afterwards, the visited network will contact the home network and request the subscriber profile from HSS. In this process, service may be provided in a home routed or local breakout mode. The IP address can be allocated from home network or visited network accordingly. There may be a mismatch between the subscriber request and network capability. The following table lists the potential failure cases.

UE Request	Visited Network Capability	Home routed	Local Breakout
Dual stack	IPv4-only	Failure case 1	Failure case 1
Dual stack	IPv4-only/IPv6-only	Failure case 1	Failure case 2
Dual stack	IPv6-only	Failure case 1	Failure case 3
IPv6-only	IPv4-only	OK	Failure case 4
IPv6-only with 464xlat	Dual stack	OK	Failure case 5
IPv6-only with 464xlat	IPv6-only	OK	OK
IPv4-only	Dual stack	OK	OK

Table 1: Roaming Scenario Descriptions

#### 4. Failure Cases Descriptions

##### 4.1. Failure Case 1: Incompatible with Extended PDP/PDN Type

A mobile device in a dual-stack network likely requests PDP/PDN type IPv4v6 to allocate address. Such PDP/PDN type should be understandable in the network nodes, including Serving GPRS Support Node(SGSN), Gateway GPRS Support Node(GGSN), Mobility Management Entity (MME), Serving Gateway(SGW), PDN Gateway(PGW), Home Location Registrar(HLR) and Home Subscriber Server(HSS). When a subscriber roams to the IPv4 network, the visited SGSN or MME has to communicate with HLR/HSS in the home land to retrieve the subscriber profile. The issue we observe is that multiple SGSN/MME will be unable to correctly process a subscriber profile received in the Insert Subscriber Data procedure if it contains an Ext-PDP-Type defined in

3GPP [TS29.002]. Therefore, it will likely refuse the subscriber registration.

Operators may have to remove the PDP/PDN type IPv4v6 from HLR/HSS in home networks, that will restrict UEs only initiates IPv4 PDP or IPv6 PDP activation. In order to avoid this situation, operators should make a comprehensive roaming agreement to support IPv6 and ensure that aligns with GSMA document, e.g [IR.33], [IR.88] and [IR.21]. Since the agreement requires visited operators to upgrade all SGSN nodes, some short- or medium-term solutions have been implemented to fix the issue. There are some specific configurations in HLR/HSS of home network. Multiple PDP/PDN subscription information will be added in the subscriber profiles, for example it may include both PDP/PDN type IPv4 and PDP/PDN type IPv4v6 for a user profile. Once the HLR/HSS receives an Update Location message from visited SGSN/MME, only the subscription data with PDP/PDN type IPv4 will be sent to SGSN/MME in the Insert Subscriber Data procedure. It guarantee the user profile could compatible with visited SGSN/MME capability.

#### 4.2. Failure Case 2: Splitting Dual-stack Bearer

Dual-stack capability can also be provided in a early mobile network(i.e. Pre-Release 8 network) using separate PDP/PDN activations. That means only a single IPv4 and IPv6 PDP/PDN can be initiated to allocate IPv4 and IPv6 address separately. Once a UE with PDP/PDN type IPv4v6 request roams to those networks, same issue described in failure case 1 will be occurred if the UE initiate a network attachment process.

If networks could allow UE to make a success attachment, a roaming subscriber with IPv4v6 PDP/PDN type should change the request to two separated PDP/PDN request with single IP version in order to achieve equivalent results. This restriction may be occurred in the below two cases.

- o The GGSN/PGW preferences dictate the use of IPv4 addressing only or IPv6 prefix only for a specific APN.
- o The SGSN/MME does not set the Dual Address Bearer Flag due to the operator using single addressing per bearer to support interworking with nodes of earlier releases

Above process would likely double PDP/PDN allocation costs. Some operators may only allow one PDP/PDN is alive for each subscriber. For example, IPv6 PDP/PDN would be rejected if the subscriber has an active IPv4 PDP/PDN. Therefore, the subscriber will lost IPv6 connection in the visited network. Even the two parallel PDP/PDN activations are allowed, it will require additional correlation of

those two sessions of single IP version on the charging system. If there are Policy and Charging Rules Function(PCRF)/Policy and Charging Enforcement Function (PCEF) deployed, the system would treat IPv4 and IPv6 session as independent and perform different Quality of Service(QoS) policies. The subscriber may have unstable experiences due to different behaviors on each IP version connection.

#### 4.3. Failure Case 3: Shortage of IPv6 support

Some operators may adopt IPv6-only configuration for the IMS service, e.g. Voice over LTE(VoLTE) or Rich Communication Suite (RCS). Since IMS roaming architecture will offload all traffic in the visited network, a dual-stack subscriber can only be assigned with IPv6 address. There is no IPv4 address returned. It requires all the IMS based applications should be IPv6 enable. A translation-based method, for example Bump-in-the-host (BIH)[RFC6535] and 464xlat[RFC6877] , may help to address the issue if there is IPv6 compatibility problems. Operators may could automatically enable the function in a IPv6-only network and disable in a dual-stack or IPv4 network.

#### 4.4. Failure Case 4: Fallback Incapability

3GPP specified the PDP/PDN type IPv6 as early as PDP/PDN type IPv4. Therefore, the IPv6 single PDP/PDN type has been well supported and interpretable in the 3GPP network nodes. When a subscriber requests PDP/PDN type IPv6, the network should only return the expected IPv6 address. Otherwise, the request should be dropped and the error code should be sent to the user. Roaming to IPv4-only networks with IPv6 PDP/PDN request would fail to get addresses. A proper fallback is desirable however the behavior is implementation specific. There are some UE have the ability to provide a different configuration for home network and visited network respectively. It guarantees UE will always initiate PDP/PDN type IPv4 in the roaming area. Android system solves the issue by setting the roaming Access Point Name(APN). The mobile terminal is allowed to ignore the original requested protocol and always adhere to IPv4 when roaming. Those fallback mechanisms are deserved to be implemented timely.

#### 4.5. Failure Case 5: 464xlat Support

464xlat[RFC6877] is proposed to address IPv4 compability issue in a IPv6 single-stack environment. The function on a mobile terminal likely gets along with PDP/PDN IPv6 type request to cooperate with a remote NAT64[RFC6146] gateway. 464xlat may use the mechanism defined in [I-D.ietf-behave-nat64-discovery-heuristic] to automatically discover NAT64 prefixes. Those behaviors depend on the network deployment. If the DNS64 or NAT64 is not deployed in the visited



networks, 464xlat may be failed to perform. Considering the various network's situations, operators may adopt 464xlat in the home networks and use IPv4-only in the roaming networks with different roaming profile configurations.

As an alternative solution, an AAA Server could be deployed to connect with GGSN/PGW. Once the GGSN/PGW receive the session creation requests, it will initiate an Access-Request to an AAA server via Radius protocol. The Access-Request contains subscriber and visited network information, e.g. PDP/PDN Type, International Mobile Equipment Id (IMEI), Software Version(SV) and visited SGSN/MME location code, etc. The AAA server could take IMEI and SV components to verify if device has 464XLAT support. Combining with the visited network information, the AAA server will ultimately decide to enable 464xlat in an IPv6-only roaming or fallback to IPv4.

## 5. Discussions

The dual-stack deployment is recommended in most cases. However, it may take some times in a mobile environment. 3GPP didn't specify PDP/PDN type IPv4v6 in the early release. Such PDP/PDN type is supported in new-built Long Term Evolution(LTE)/System Architecture Evolution(SAE) network, but didn't support well in the third generation network. The situations may cause the roaming issues dropping the attachment from dual-stack subscribers in the case of LTE to 3G and IPv6-enabled 3G to IPv4 3G. Operators may have to adopt temporary solution unless all the interworking nodes(i.e. SSGN and SGW) in the visited network have been upgraded to support Ext-PDP-Type feature.

As an alternative solution for dual-stack, operators may change a unified PDP/PDN request into two separated single IP version requests. However, this approach is problematic in the Charging records and QoS policy enforcement. In addition, it doubles the PDP resource uses. It may be unappealing for the deployment.

Conversely, some operators may choose PDP/PDN Type IPv6 to start the communications in home networks and use different profile in the roaming area. Since PDP/PDN Type IPv6 has been introduced in 3GPP early release, it didn't require upgrading on the interworking nodes to make compatibility. The proper IPv4 fallback mechanism should be supported either on the mobile terminal or network equipment.

A roaming to IPv6-only network occurs when operators deploy roaming function for IMS service. A dual-stack capable device could implement translation-based function to support the IPv4 applications. Those inserted translation function can be turned off properly when the terminals roam back to dual-stack or IPv4 networks. Operators can also deploy AAA servers to make final decision

## 6. IANA Considerations

This document makes no request of IANA.

## 7. Security Considerations

The draft didn't introduce additional security concerns to the networks.

## 8. Acknowledgements

The authors would like to thank V6ops chairs(Fred Baker and John Brzozowski) to encourage us to continue the work. This document is the result of the IETF V6ops IPv6-Roaming design team effort.

## 9. References

### 9.1. Normative References

- [I-D.ietf-behave-nat64-discovery-heuristic]  
Savolainen, T., Korhonen, J., and D. Wing, "Discovery of the IPv6 Prefix Used for IPv6 Address Synthesis", draft-ietf-behave-nat64-discovery-heuristic-17 (work in progress), April 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", RFC 6146, April 2011.
- [RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 2011.

- [RFC6535] Huang, B., Deng, H., and T. Savolainen, "Dual-Stack Hosts Using "Bump-in-the-Host" (BIH)", RFC 6535, February 2012.
- [RFC6877] Mawatari, M., Kawashima, M., and C. Byrne, "464XLAT: Combination of Stateful and Stateless Translation", RFC 6877, April 2013.

## 9.2. Informative References

- [IR.21] Global System for Mobile Communications Association, GSMA., "Roaming Database, Structure and Updating Procedures", July 2012.
- [IR.33] Global System for Mobile Communications Association, GSMA., "GPRS Roaming Guidelines", July 2012.
- [IR.65] Global System for Mobile Communications Association, GSMA., "IMS Roaming & Interworking Guidelines", May 2012.
- [IR.88] Global System for Mobile Communications Association, GSMA., "LTE Roaming Guidelines", January 2012.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6586] Arkko, J. and A. Keranen, "Experiences from an IPv6-Only Network", RFC 6586, April 2012.
- [TR23.975] 3rd Generation Partnership Project, 3GPP., "IPv6 migration guidelines", June 2011.
- [TS23.401] 3rd Generation Partnership Project, 3GPP., "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access v9.00", March 2009.
- [TS29.002] 3rd Generation Partnership Project, 3GPP., "Mobile Application Part (MAP) specification v9.00", December 2009.
- [TS29.272]

3rd Generation Partnership Project, 3GPP., "Mobility Management Entity (MME) and Serving GPRS Support Node (SGSN) related interfaces based on Diameter protocol v9.00", September 2009.

Authors' Addresses

Gang Chen  
China Mobile  
53A,Xibianmennei Ave.,  
Xuanwu District,  
Beijing 100053  
China

Email: phdgang@gmail.com

Hui Deng  
China Mobile  
53A,Xibianmennei Ave.,  
Xuanwu District,  
Beijing 100053  
China

Email: denghui@chinamobile.com

Dave Michaud  
Rogers

Email: Michaud@rci.rogers.com

Jouni Korhonen  
Renesas Mobile  
Porkkalankatu 24  
FIN-00180 Helsinki, Finland

Email: jouni.nospam@gmail.com

Mohamed Boucadair  
France Telecom  
No.32 Xuanwumen West Street  
Rennes,  
35000  
France

Email: mohamed.boucadair@orange.com

Vizdal Ales  
Deutsche Telekom AG  
Tomickova 2144/1  
Prague 4, 149 00  
Czech Republic

Email: ales.vizdal@t-mobile.cz

Cameron Byrne  
T-Mobile USA  
Bellevue  
Washington 98105  
USA

Email: cameron.byrne@t-mobile.com

INTERNET-DRAFT  
Intended Status: Informational

N. Elkins  
Inside Products  
M. Ackermann  
BCBS Michigan  
W. Jouris  
Inside Products  
K. Haining  
US Bank  
S. Perdomo  
DTCC  
October 3, 2013

Expires: April 2014

End-to-end Response Time Needed for IPv6 Diagnostics  
draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01

Abstract

To diagnose performance and connectivity problems, metrics on real (non-synthetic) transmission are critical for timely end-to-end problem resolution. Such diagnostics may be real-time or after the fact, but must not impact an operational production network. The base metrics are: packet sequence number and packet timestamp. Metrics derived from these will be described separately. This document provides the background and rationale for the requirement for end-to-end response time.

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Background . . . . .	3
1.1	Why End-to-end Response Time is Needed . . . . .	3
1.2	Trending of Response Time Data . . . . .	4
1.3	What to measure? . . . . .	4
1.4	TCP Timestamp not enough . . . . .	5
1.5	Inadequacy of Current Instrumentation Technology . . . . .	5
1.5.1	Synthetic transactions . . . . .	5
1.5.2	PING . . . . .	5
1.5.3	Other Estimates of Network Time . . . . .	6
1.5.4	Server / Client Agents . . . . .	6
2	Solution Parameters . . . . .	6
2.1	Rationale for proposed solution . . . . .	7
2.2	Merits of timestamp in PDM . . . . .	7
2.3	What kind of timestamp? . . . . .	8
3	Backward Compatibility . . . . .	8
4	Security Considerations . . . . .	8
5	IANA Considerations . . . . .	8
6	References . . . . .	8
6.1	Normative References . . . . .	9
6.2	Informative References . . . . .	9
7	Acknowledgments . . . . .	9
	Authors' Addresses . . . . .	10

## 1 Background

To diagnose performance and connectivity problems, metrics on real (non-synthetic) transmission are critical for timely end-to-end problem resolution. Such diagnostics may be real-time or after the fact, but must not impact an operational production network. The base metrics are: packet sequence number and packet timestamp. Metrics derived from these will be described separately. This document provides the background and rationale for the requirement for end-to-end response time.

For background, please see draft-ackermann-tictoc-pdm-ntp-usage-00 [ACKPDM], draft-elkins-v6ops-ipv6-packet-sequence-needed-01 [ELKPSN], draft-elkins-v6ops-ipv6-pdm-recommended-usage-01 [ELKPUSE], draft-elkins-6man-ipv6-pdm-dest-option-02 [ELKPDM] and draft-elkins-ippm-pdm-metrics-00 [ELKIPPM]. These drafts are companions to this document.

As discussed in the above Internet Drafts, current methods are inadequate for these purposes because they assume unreasonable access to intermediate devices, are cost prohibitive, require infeasible changes to a running production network, or do not provide timely data. The IPv6 Performance and Diagnostic Metrics destination option (PDM) provides a solution to these problems. This document will detail the background and need for end-to-end response time.

### 1.1 Why End-to-end Response Time is Needed

The timestamps in the PDM traveling along with the packet will be used to calculate end-to-end response time, without requiring agents in devices along the path. In many networks, end-to-end response times are a critical component of Service Levels Agreements (SLAs).

End-to-end response is what the user of a network system actually experiences. When the end user is an individual, he is generally indifferent to what is happening along the network; what he really cares about is how long it takes to get a response back. But this is not just a matter of individuals' personal convenience. In many cases, rapid response is critical to the business being conducted.

When the end user is a device (e.g. with the Internet of Things), what matters is the speed with which requested data can be transferred -- specifically, whether the requested data can be transferred in time to accomplish the desired actions. This can be important when the relevant external conditions are subject to rapid change.

Response time and consistency are not just "nice to have". On many



networks, the impact can be financial hardship or endanger human life. In some cities, the emergency police contact system operates over IP, law enforcement uses TCP/IP networks, our stock exchanges are settled using IP networks. The critical nature of such activities to our daily lives and financial well-being demand a solution. Section 1.5 will detail the current state of end-to-end response time monitoring today.

## 1.2 Trending of Response Time Data

In addition to the need for tracking current service, end-to-end response time is valuable for capacity planning. By tracking response times, and identifying trends, it becomes possible to determine when network capacity is being approached. This allows additional capacity to be obtained before service levels fall below requirements. Without that kind of tracking, the only option is to wait until there is a problem, and then scramble to get additional capacity on an emergency (and probably high cost) basis.

The documents draft-elkins-v6ops-ipv6-pdm-recommended-usage-01 [ELKPUSE] and draft-elkins-ippm-pdm-metrics-00 [ELKIPPM] will detail use for the PDM for capacity planning purposes.

## 1.3 What to measure?

End to end response time can be broken down into 3 parts:

- Network delay
- Application (or server) delay
- Client delay

Network delay may be one-way delay [RFC2679] or round-trip delay [RFC2681].

Additionally, network delay may include multiple hops. Application and server delay include operating system by stack time. By and large, the three timings are 'good enough' measurements to allow rapid triage into the failing component.

Ways are available (provided by operating systems) to measure Application and Client times. Network time can also be measured in isolation via some of the measurement techniques described in section 1.5. The most difficult portion is to integrate network time with the server or application times. Products exist to do this but are available at an exorbitant cost, require agents, and will likely become more prohibitive as the speed of networks grow and as the world becomes more connected via mobile devices. This is discussed in detail in section 1.5.

Measuring network time requires precise timestamps. Furthermore, those timestamps need to occur at the end-points of the transactions being measured. And they need to be available, regardless of the protocol being used by the transaction. Which is to say, the timestamp has to be available in one of the extensions to the IP header - this is provided by the PDM.

#### 1.4 TCP Timestamp not enough

Some suggest that the TCP Timestamp option might be sufficient to calculate end-to-end response time.

The TCP Timestamp Option is defined in RFC1323 [RFC1323]. The reason for the TCP Timestamp option is to be able to discard packets when the TCP sequence number wraps. (PAWS)

The problems with the TCP Timestamp option are:

1. Not everyone turns this on.
2. It is only available for TCP applications
3. No time synchronization between sender and receiver.
4. No indication of date in long-running connections. (That is connections which last longer than one day)
5. The granularity of the timestamp is at best at millisecond level. In the future, as speeds of devices and networks grow, this level of granularity will be inadequate. Even today, on many networks, the timings are at microsecond level not millisecond.

#### 1.5 Inadequacy of Current Instrumentation Technology

The current technology includes:

1. Synthetic transactions
2. Pings
3. Other Estimates of network time
4. Server / Client Agents

##### 1.5.1 Synthetic transactions

##### 1.5.2 PING An ICMP ping measures network time. First, you can PING the remote device. Then you assume that the time it takes to get a

response to a PING is the same as the time that a transaction (regardless of packet size) would take to traverse the network. However, QoS rules, firewalls, etc. may mean that PING, (and other synthetic transactions) may not be subject to the same conditions.

#### 1.5.3 Other Estimates of Network Time

If a packet trace is done, it is possible to look at the time between when a response was seen to be sent at the packet capture device and when the ACK for the response comes back.

If you assume that the ACK took the same amount of time as the original query, you have the network time. Unfortunately, the time for the ACK may not be the same as the time for a much larger query transaction to traverse the network.

The biggest problem with this method is that of TCP delayed acknowledgements. If the client is doing delayed ACKs, then the ACK will be held until the next request is ready to go out. In this case, the time to receive the ACK has no correlation with network time.

#### 1.5.4 Server / Client Agents

There are also products which claim that they can determine end-to-end response times, integrating server and network times - and indeed they can do so. But they require agents which must be placed at each point which is to be monitored. That is, it is necessary to add those agents EVERYWHERE around the network, at a very high cost. These kind of products can be purchased by only the richest 1% of the corporations. As the speed of networks grow, and as the world becomes more connected via mobile devices, such products will only become more expensive. If, indeed, their technology can keep up.

TCP/IP networks today are used throughout the world. The need for adequate performance will become more and more critical. A method that is scalable and affordable is needed to ensure this growth.

## 2 Solution Parameters

What is needed is:

- 1) A method to identify and/or track the behavior of a connection without assuming access to the transport devices.
- 2) A method to observe a connection in flight without introducing agents.

- 3) a method to observe arbitrary flows at multiple points within a network and correlate the results of those observations in a consistent manner.
- 4) A method to signal and correlate transport issues to application end-to-end behavior.
- 5) A method which does not require changes to a production network in real time.
- 6) Adequate granularity in the measurement technique to provide the needed metrics.
- 7) A method that is scalable to very large networks.
- 8) A method that is affordable to all.

## 2.1 Rationale for proposed solution

The current IPv6 specification does not provide a timestamp number nor similar field in the IPv6 main header or in any extension header. So, we propose the IPv6 Performance and Diagnostic Metrics destination option (PDM) [ELKPDM].

## 2.2 Merits of timestamp in PDM

Advantages include:

1. Less overhead than other alternatives.
2. Real measure of actual transactions.
3. Less cost to provide solutions
4. More accurate and complete information.
5. Independence from transport layer protocols.
6. Ability to span organizational boundaries with consistent instrumentation

In other words, this is a solution to a long-standing problem. The PDM will provide a metric which will allow those responsible for network support to determine what is happening in their network without expensive equipment (agents) at each device.

The PDM does not solve every response time issue for every situation. Network connections with multiple hops will still need more granular

metrics, as will the differentiation between multiple components at each host. That is, TCP/IP stack time vs. applications time will still need to be broken out by client software. What the PDM does provide is triage. That is, to determine quickly if the problem is in the network or in the server or application.

### 2.3 What kind of timestamp?

Questions arise about exactly the kind of timestamp to use. Both the Network Time Protocol (NTP) [RFC5905] and Precision Time Protocol (PTP) [IEEE1588] are used to provide timing on TCP/IP networks.

NTP has evolved within the IETF structure while PTP has evolved within the Institute of Electrical and Electronics Engineers (IEEE) community. By and large, operating systems such as Windows, Linux, and IBM mainframe computers use NTP. These are the source and destination systems for packets. Intermediate nodes such as routers and switches may prefer PTP.

Since we are describing a new extension header for destination systems, the timestamp to be used will be in accordance with NTP. In the documents, draft-ackermann-tictoc-pdm-ntp-usage-00 [ACKPDM] and draft-elkins-v6ops-ipv6-pdm-recommended-usage-01 [ELKPUSE], we will discuss guidelines for implementing NTP for use with the PDM.

## 3 Backward Compatibility

The scheme proposed in this document is backward compatible with all the currently defined IPv6 extension headers. According to RFC2460 [RFC2460], if the destination node does not recognize this option, it should skip over this option and continue processing the header.

## 4 Security Considerations

There are no security considerations.

## 5 IANA Considerations

There are no IANA considerations.

## 6 References

### 6.1 Normative References

- [RFC1323] Jacobson, V., Braden, R., and D. Borman, "TCP Extensions for High Performance", RFC 1323, May 1992.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
- [RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.
- [RFC2681] Almes, G., Kalidindi, S., and M. Zekauskas, "A Round-trip Delay Metric for IPPM", RFC 2681, September 1999.
- [RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.
- [IEEE1588] IEEE 1588-2002 standard, "Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems"

### 6.2 Informative References

- [ACKPDM] Ackermann, M., "draft-ackermann-tictoc-pdm-ntp-usage-00", Internet Draft, September 2013.
- [ELKPSN] Elkins, N., "draft-elkins-v6ops-ipv6-packet-sequence-needed-01", Internet Draft, September 2013.
- [ELKPDM] Elkins, N., "draft-elkins-6man-ipv6-pdm-dest-option-02", Internet Draft, September 2013.
- [ELKPUSE] Elkins, N., "draft-elkins-v6ops-ipv6-pdm-recommended-usage-01", Internet Draft, September 2013
- [ELKIPPM] Elkins, N., "Draft-elkins-ippm-pdm-metrics-00", Internet Draft, September 2013.

## 7 Acknowledgments

The authors would like to thank Rick Troth, David Boyes, and Fred Baker for their comments.

Authors' Addresses

Nalini Elkins  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States  
Phone: +1 831 659 8360  
Email: [nalini.elkins@insidethestack.com](mailto:nalini.elkins@insidethestack.com)  
<http://www.insidethestack.com>

Michael S. Ackermann  
Blue Cross Blue Shield of Michigan  
P.O. Box 2888  
Detroit, Michigan 48231  
United States  
Phone: +1 310 460 4080  
Email: [mackermann@bcbsmi.com](mailto:mackermann@bcbsmi.com)  
<http://www.bcbsmi.com>

Keven Haining  
US Bank  
16900 W Capitol Drive  
Brookfield, WI 53005  
United States  
Phone: +1 262 790 3551  
Email: [keven.haining@usbank.com](mailto:keven.haining@usbank.com)  
<http://www.usbank.com>

Sigfrido Perdomo  
Depository Trust and Clearing Corporation  
55 Water Street  
New York, NY 10055  
United States  
Phone: +1 917 842 7375  
Email: [s.perdomo@dtcc.com](mailto:s.perdomo@dtcc.com)  
<http://www.dtcc.com>

William Jouris  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States  
Phone: +1 925 855 9512  
Email: [bill.jouris@insidethestack.com](mailto:bill.jouris@insidethestack.com)  
<http://www.insidethestack.com>





v6ops Working Group  
Internet Draft  
Intended status: Standards track  
Expires: September, 2013

N. Elkins  
Inside Products  
L. Kratzke  
IBM  
M. Ackermann  
BCBS of Michigan  
K. Haining  
US Bank

April 2013

IPv6 IPID Needed  
draft-elkins-v6ops-ipv6-ipid-needed-01.txt

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), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on October 4, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

## Abstract

The IPv4 main header contained a 16-bit IP Identification (IPID) field used for fragmentation and reassembly. In practice, this field was commonly used by network diagnosticians for tracking packets. In IPv6, the IPID has been moved to the Fragment header, and would only be used when fragmentation is required. Thus, the IPID field in IPv6, is no longer able to be utilized in the valuable role it played in IPv4, relative to diagnostics and problem resolution. This causes great concern in particular for end users and large enterprises, for whom Network/Application availability and performance can directly and profoundly affect bottom line financials. Several viable solutions to this situation exist.

## Table of Contents

1. Introduction .....	4
2. Conventions used in this document .....	5
3. Applicability .....	6
6. Security Considerations .....	7
7. IANA Considerations .....	7
10. References .....	7
10.1. Normative References .....	8
11. Acknowledgments .....	8

## 1. Introduction

In IPv4, the 16 bit IP Identification (IPID) field is located at an offset of 4 bytes into the IPv4 header and is described in RFC791 [RFC791]. In IPv6, the IPID field is a 32 bit field contained in the Fragment Header defined by section 4.5 of RFC2460 [RFC2460]. Unfortunately, unless fragmentation is being done by the source node, the packet will not contain this Fragment Header, and therefore will have no Identification field.

The intended purpose of the IPID field is to enable fragmentation and reassembly, and as currently specified is required to be unique within the maximum segment lifetime (MSL) on all datagrams. The MSL is often 2 minutes.

In Large Enterprise Networks, the IPID field is used for more than fragmentation. During network diagnostics, packet traces may be taken at multiple places along the path, or at the source and destination. Then, packets can be matched by looking at the IPID.

Obviously, the time at each device will differ according to the clock on that device; so another metric is required. This method of taking multiple traces along the path is of special use on large multi-tier networks to see where the packet loss or packet corruption is happening. Multi-tier networks are those which have multiple routers or switches on the path between the sender and the receiver.

The inclusion of the IPID makes it easier for a device(s) in the middle of the network, or on the receiving end of the network, to identify flows belonging to a single node, even if that node might have a different IP address. For example, if the sending node is a mobile laptop with a wireless connection to the Internet.

For its de-facto diagnostic mode usage, the IPID field needs to be available whether or not fragmentation occurs. It also needs to be unique in the context of the entire session, and across all the connections controlled by the stack.

This document will present information that demonstrates how valuable and useful the IPID field has been (in IPv4) for diagnostics and problem resolution, and how not having it available (in IPv6), could be a major detriment to new IPv6 deployments and contribute to protracted downtimes in existing IPv6 operations.

As network technology has evolved, the uses to which fields are put can change as well. De-facto use is powerful, and should not be lightly ignored. In fact, it is a testament to the power and pervasiveness of the protocol that users create new uses for the original technology.

For example, the use of the IPID goes beyond the vision of the original authors. This sort of thing has happened with numerous other technologies. It is similar to the ways in which cell phones have evolved to be more than just a means of vocal communication, including Internet communications, photo-sharing, stock exchange transactions, etc. Or the way that the bicycle, originally intended merely as a means of fashionable transportation for a single individual, developed into a replacement for the horse in hauling materials. Or the way that the automobile went from being a means of transport for people to a truck, for transport of materials on a large scale. Indeed, the Internet itself has evolved, from a small network for researchers and the military to share files into the pervasive global information superhighway that it is today.

## 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 RFC2119 [RFC2119].

### 3. Applicability

The ability to utilize the IPID has enhanced problem diagnosis efforts and significantly reduced problem resolution time.

Several actual use case examples are shown below. These demonstrate how use of the IPID has reduced problem resolution time in very valuable production networks of Large Enterprises/End Users. In general, if a problem or performance issue with an application or network component can be fixed in minutes, as opposed to hours, this can mean significant dollar savings to large enterprises. The IPID can be used extensively when debugging involves traces or packet captures. Its absence in IPv6 may lead to protracted problem diagnosis and extended problem resolution time.

This value/perspective may be unique to tech support organizations of large enterprises. Other functional areas may not share this concern/perspective, as packets could continue to flow, but service levels may not be acceptable to end users during the extended problem resolution time.

Although very situation dependent, the use cases below clearly illustrate the value of network availability, and the need to keep problem resolution time to an absolute minimum.

Another benefit of using the IPID to expedite problem resolution is reducing the cost of associated resources being consumed during extended problem resolution, such as storage, CPU and staff time.

Will IPID be critical in most problem resolutions? NO! But if it even helps in a few per year, significant money and/or lost business could be saved.

A facility such as IPID, that has proven field value, should not be eliminated as an effective diagnosis tool!

#### USE CASE EXAMPLES:

##### USE CASE #1 --- Large Insurance Company

- (estimated time saved by use of IPID: 7 hours)

##### PERFORMANCE TOOL PRODUCES EXTRANEIOUS PACKETS?

- Issue was whether a performance tool was accurately replicating session flow during performance testing?
- Trace IPIDs showed more unique packets within same flow from performance tool compared to IE Browser.
- Having the clear IPID sequence numbers also showed where and why the extra packets were being generated.
- Solution: Problem rectified in subsequent version of performance tool.
- Without IPID, it was not clear if there was an issue at all.

##### USE CASE #2 --- Large Bank

- (estimated time saved by use of IPID: 4 hours)

##### BATCH TRANSFER DURATION INCREASES 12X

- A 30 minute data transfer started taking 6-8 hours to complete.
- Possible packet loss? All vendors said no.
- Other Apps were working OK.
- 4 trace points used, and then IPIDs compared.
- Showed 7% packet loss.
- Solution: WAN hardware was replaced and problem fixed.
- Without IPID, no one would agree a problem existed

##### USE CASE #3 --- Large Bank

- (estimated time saved by use of IPID: 6 hours)

##### VERY SLOW INTERACTIVE PERFORMANCE.

- All network links looked good.
- Traces showed duplicated small packets (which can be OK).
- Saw that IPID was equal but TTL was always + 1.
- Network device was "Splitting" small packets only.(2 interfaces).
- The small packets were control info, telling other side to slow down.
- Erroneously looked like network congestion.
- Solution: Network Device replaced and good interactive performance restored.
- Without IPID, flows would have appeared OK.

## USE CASE #4 --- Large Government Agency

- (estimated time saved by use of IPID: 9 hours)

## VPN DROPS

- Cell phone connections to law enforcement were being dropped. Going through a VPN.
- All parties (both sides of VPN connection, application, etc.) said it was not their problem. Problem went on for weeks.
- Finally, when we were called in as consultants, we took a trace which showed packet with IPID and TTL that did not match others in the flow AT ALL was coming from router nearest application server end of VPN.
- Solution: Provider for VPN for application server changed. Problem resolved.
- Without IPID, much harder to diagnose problem.
- (Same case also happened with large corporation. Again, all parties saying not their fault until proven via packet trace.)

The IPID is very valuable to large enterprises and Data Center Operators (EDCO) in trace analysis, specifically in reducing problem diagnosis and resolution time. As such, IPID or something equivalent, should be part of IPv6 for all situations where it can provide value. (As it is IPv4.) Not just where fragmentation is required.

## 6. Security Considerations

There are no security considerations.

## 7. IANA Considerations

There are no IANA considerations.

## 10. References

## 10.1. Normative References

[RFC791] Postel, J., "Internet Protocol", RFC 791 / STD 5, September 1981.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.



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

## 11. Acknowledgments

The authors would like to thank Fred Baker, Bill Jouris, Jose Isidro, R. J. Atkinson, James Ashton, Sigfrido Perdomo and Neil Wasserman for their reviews and suggestions that made this document better.

This document was prepared using 2-Word-v2.0.template.dot.

### Authors' Addresses

Nalini Elkins  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States

Phone: +1 831 659 8360  
Email: [nalini.elkins@insidethestack.com](mailto:nalini.elkins@insidethestack.com)

Lawrence Kratzke  
IBM  
8121 Glenbrittle Way  
Raleigh, NC 27615  
United States

Phone: +1 800-876-8801  
Email: [kratzke@us.ibm.com](mailto:kratzke@us.ibm.com)

Internet-Draft

IPv6 IPID Needed

April 2013

Michael Ackermann  
Blue Cross Blue Shield of Michigan  
P.O. Box 2888  
Detroit, Michigan 48231  
United States

Phone: +1 310 460 4080  
Email: mackermann@bcbsmi.com

Keven Haining  
US Bank  
16900 W Capitol Drive  
Brookfield, WI 53005

Phone: +1 262-790-3551  
Email: keven.haining@usbank.com

Elkins

Expires October 4, 2013

[Page 10]

INTERNET-DRAFT  
Intended Status: Informational

N. Elkins  
Inside Products  
M. Ackermann  
BCBS Michigan  
W. Jouris  
Inside Products  
K. Haining  
US Bank  
S. Perdomo  
DTCC  
October 3, 2013

Expires: April 2014

IPv6 Packet Sequence Number Needed  
draft-elkins-v6ops-ipv6-packet-sequence-needed-01

Abstract

To diagnose performance and connectivity problems, metrics on real (non-synthetic) transmission are critical for timely end-to-end problem resolution. Such diagnostics may be real-time or after the fact, but must not impact an operational production network. The base metrics are: packet sequence number and packet timestamp. Metrics derived from these will be described separately. This document provides the background and rationale for the packet sequence number which is a part of the IPv6 Performance and Diagnostic Metrics Destination Option (PDM).

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>

## Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Background . . . . .	3
1.1	Why Packet Sequence Number . . . . .	3
1.2	IPv4 IPID : DeFacto Sequence Number . . . . .	4
1.2.1	Description of IPID in IPv4 . . . . .	4
1.2.2	DeFacto Use of IPID . . . . .	4
1.2.3	Merits of DeFacto Usage . . . . .	5
1.2.4	Use Cases of IPv4 IPID in Diagnostics . . . . .	5
1.3	TCP sequence number is not enough . . . . .	6
1.4	Inadequacy of current measurement techniques . . . . .	7
1.4.1	SNMP / CMIP Counters . . . . .	7
1.4.2	Router / Firewall Logs . . . . .	7
1.4.3	Netflow . . . . .	7
1.4.4	Access to Intermediate Devices . . . . .	8
1.4.5	Modifications to an Operational Production Network . . . . .	8
2	Solution Parameters . . . . .	9
2.1	Packet Trace Meets Criteria . . . . .	9
2.1.1	Limitations of Packet Capture . . . . .	9
2.1.2	Problem Scenario 1 . . . . .	9
2.1.3	Problem Scenario 2 . . . . .	11
3	Rationale for Proposed Solution (PDM) . . . . .	11
4	Backward Compatibility . . . . .	11
5	Security Considerations . . . . .	12
6	IANA Considerations . . . . .	12
7	References . . . . .	12
7.1	Normative References . . . . .	12
8	Acknowledgments . . . . .	12
	Authors' Addresses . . . . .	13

## 1 Background

To diagnose performance and connectivity problems, metrics on real (non-synthetic) transmission are critical for timely end-to-end problem resolution. Such diagnostics may be real-time or after the fact, but must not impact an operational production network. The base metrics are: packet sequence number and packet timestamp. Metrics derived from these will be described separately.

For background, please see draft-ackermann-tictoc-pdm-ntp-usage-00 [ACKPDM], draft-elkins-6man-ipv6-pdm-dest-option-02 [ELKPDM], draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01 [ELKRSP], draft-elkins-v6ops-ipv6-pdm-recommended-usage-01 [ELKUSE] and draft-elkins-ippm-pdm-metrics-00 [ELKIIPPM]. These drafts are companions to this document.

As discussed in the above Internet Drafts, current methods are inadequate for these purposes because they assume unreasonable access to intermediate devices, are cost prohibitive, require infeasible changes to a running production network, or do not provide timely data. The IPv6 Performance and Diagnostic Metrics destination option (PDM) provides a solution to these problems. This document will detail the background and need for the packet sequence number.

### 1.1 Why Packet Sequence Number

In many networks, during network diagnostics of an end-to-end connection, it becomes necessary to find the device along the network path creating problems. Diagnostic data may be collected at multiple places along the path (if possible), or at the source and destination. Then, the diagnostic data must be matched. Packet sequence number is critical in this matching process. The timestamp or even the IP addresses may be different at different devices. In IPv4 networks, the IPID field was used as a de facto sequence number. This will be discussed at greater length in section 1.2.

This method of data collection along the path is of special use on large multi-tier networks to determine where packet loss or packet corruption is happening. Multi-tier networks are those which have multiple routers or switches on the path between the sender and the receiver.

## 1.2 IPv4 IPID : DeFacto Sequence Number

With IPv4 networks, on many stack implementations, but not all, the IPID field has the property of sequentiality.

### 1.2.1 Description of IPID in IPv4

In IPv4, the 16 bit IP Identification (IPID) field is located at an offset of 4 bytes into the IPv4 header and is described in RFC0791 [RFC0791]. In IPv6, the IPID field is a 32-bit field contained in the Fragment Header defined by section 4.5 of RFC2460 [RFC2460]. Unfortunately, unless fragmentation is being done by the source node, the IPv6 packet will not contain this Fragment Header, and therefore will have no Identification field.

The intended purpose of the IPID field, in both IPv4 and IPv6, is to enable fragmentation and reassembly, and as currently specified is required to be unique within the maximum segment lifetime (MSL) on all datagrams. The MSL is often 2 minutes.

### 1.2.2 DeFacto Use of IPID

In many networks, the IPID field is used for more than fragmentation. During network diagnostics, packet traces may be taken at multiple places along the path, or at the source and destination. Then, packets can be matched by looking at the IPID.

The inclusion of the IPID makes it easier for a device(s) in the middle of the network, or on the receiving end of the network, to identify flows belonging to a single node, even if that node might have a different IP address. For example, in the case of sessions going through a NAT or proxy server.

For its de-facto diagnostic mode usage, the IPID field needs to be available whether or not fragmentation occurs. It also needs to be unique in the context of the session, and across all the connections controlled by the stack. In IPv4, the IPID is in the main header, so it is available for all packets. As it is a 16-bit field, it wrapped during the course of the session and thus had some limitations.

Even with these limitations, the IPID has been valuable and useful in IPv4 for diagnostics and problem resolution. It is a practical solution that is 'good enough' in many instances. Not having it available in IPv6, may be a major detriment to new IPv6 deployments and contribute to protracted downtimes in existing IPv6 operations.

### 1.2.3 Merits of DeFacto Usage

As network technology evolves, the uses to which fields are put can change as well. De-facto use is powerful, and should not be lightly ignored. In fact, it is a testament to the power and pervasiveness of the protocol that users create new uses for the original technology.

For example, the use of the IPID goes beyond the vision of the original authors. This sort of thing has happened with numerous other technologies and protocols.

The implementation of the traceroute command sends ICMP echo packets with a varying TTL. This is a very useful for diagnostics yet departs from the original purpose of TTL.

Similarly, cell phones have evolved to be more than just a means of vocal communication, including Internet communications, photo-sharing, stock exchange transactions, etc. Indeed, the Internet itself has evolved, from a small network for researchers and the military to share files into the pervasive global information superhighway that it is today.

### 1.2.4 Use Cases of IPv4 IPID in Diagnostics

Use Case # 1 --- Large Insurance Company

- (estimated time saved by use of IPID: 7 hours)

Performance Tool produces extraneous packets

- Issue was whether a performance tool was accurately replicating session flow during performance testing.
- Trace IPIDs showed more unique packets within same flow from performance tool compared to IE Browser.
- Having the clear IPID sequence numbers also showed where and why the extra packets were being generated.
- Solution: Problem rectified in subsequent version of performance tool.
- Without IPID, it was not clear if there was an issue at all.

Use Case #2 --- Large Bank

- (estimated time saved by use of IPID: 4 hours)

Batch transfer duration increases 12x

- A data transfer which formerly took 30 minutes to complete started taking 6-8 hours to complete.
- Was there packet loss? All the vendors said no.
- The other applications on the network did not report any

problems.

- 4 trace points were used, and the IPIDs in the packets were compared.
- The comparison showed 7% packet loss.
- Solution: WAN hardware was replaced and problem fixed.
- Without IPID, no one would agree a problem existed

Use Case #3 --- Large Bank

- (estimated time saved by use of IPID: 6 hours)

Very slow interactive performance

- All network links looked good.
- Traces showed duplicated small packets (which can be OK).
- We saw that the IPID was the same in both packets but the TTL was always + 1.
- A network device was "splitting" only small packets over two interfaces.
- The small packets were control info, telling other side to slow down.
- It erroneously looked like network congestion.
- Solution: Network device replaced and good interactive performance restored.
- Without IPID, flows would have appeared OK.

Use Case #4 --- Large Government Agency

- (estimated time saved by use of IPID: 9 hours)

VPN drops

- Cell phone connections to law enforcement were being dropped. The connections were going through a VPN.
- All parties (both sides of VPN connection, application, etc.) said it was not their problem. The problem went on for weeks.
- Finally, we took a trace which showed packets with IPID and TTL that did not match others in the flow AT ALL coming from the router nearest the application server end of VPN.
- Solution: Provider for VPN for application server changed. Problem resolved.
- Without IPID, much harder to diagnose problem.
- (Same case also happened with large corporation. Again, all parties saying not their fault until proven via packet trace.)

### 1.3 TCP sequence number is not enough

TCP Sequence number is defined in RFC0793 [RFC0793]. Indeed, the TCP Sequence Number along with the TCP Acknowledgment number can be used to calculate dropped packets, duplicate packets, out-of-order packets



etc. That is, IF the packet flow itself reflects accurately what happened on the wire!

See Scenario 1 (Section 1.5.2) and Scenario 2 (Section 1.5.3) for what happens with packet trace capture in real networks.

The TCP Sequence Number is, obviously, available only for TCP and not other transport protocols.

#### 1.4 Inadequacy of current measurement techniques

The question arises of whether current methods of instrumentation cannot be used without a change to the protocol. Current methods of measuring network data, other than packet traces, are inadequate because they assume unreasonable access to intermediate devices, are cost prohibitive, require infeasible changes to a running production network, or do not provide timely data. This section will discuss each of these in detail.

Current methods include both instrumentation and third party products. These include SNMP, CMIP, router logs, and firewall logs.

##### 1.4.1 SNMP / CMIP Counters

The traditional network performance counters measured by SNMP or CMIP do not provide information at the granularity desired on the behavior of application flows across the network. The problem is that such counters do not contain enough data to be able to provide a detailed and realistic view of the end-to-end behavior of a connection.

##### 1.4.2 Router / Firewall Logs

Router and firewall logs may provide some information for diagnostics. But as discussed in section 1.4.5, routers and firewalls in a production network are generally set to do minimal logging and diagnostics to allow maximum efficiency and throughput. Such devices cannot be asked to collect detailed data for an operational problem, as this requires a change to a production network.

##### 1.4.3 Netflow

Netflow is instrumentation which is available from some middle devices. As discussed in detail in section 1.4.5, such devices are generally set to do minimal logging and diagnostics to allow maximum efficiency and throughput.

Correlations to produce some level of response time data may be

possible from Netflow. But, it is not an adequate picture of end-to-end response time as Netflow is in an intermediate device and is not in a position to know what has happened at a client.

#### 1.4.4 Access to Intermediate Devices

The above current methods require access to the transport infrastructure - that is, the routers, switches or other intermediate devices. In some cases, this is possible; in others, the connections in question may cross a number of administrative entities (both in the transport and in the endpoints). When it is the enterprise at the endpoint which is interested in the diagnostics, the administrative entities who own the devices in the middle of the path have no stake in operational measurement at the enterprise or application level. They have no reason to provide the necessary data or to impact the basic transport with the instrumentation necessary to capture flow-oriented data as a continuous stream suitable for general consumption.

In other words, if you don't own the path end-to-end, you will not be able to get the data you need if you are required to get it from the devices in the middle. Not only that, the devices in the middle do not have the instrumentation necessary to make it easy to do end-to-end diagnostics because they are not responsible for that and so do not want to burden their devices with doing those kind of functions.

Many EDCO networks may not own the path end-to-end. They may be working with a business partner's network or crossing the Internet.

#### 1.4.5 Modifications to an Operational Production Network

Even when the enterprise does own all the devices along the entire path, to get enough data to adequately resolve a problem means changing the device configuration to do detailed diagnostics. In a production network, devices are generally set to do minimal logging and diagnostics. This is to allow maximum efficiency and throughput. The more logging and diagnostics such devices do, the fewer resources they have for actually transmitting traffic across the network.

So, if devices are to be asked to collect more data for an operational problem, this requires a change to a production network. This is generally not possible as it destabilizes a critical network during business hours, thus potentially disrupting many customers. Making changes is usually a lengthy process requiring change control, testing on a test network, etc. On networks which are critical to the business function, such as the networks we are discussing, it is hardly likely that changing configuration "in flight" is an option.

## 2 Solution Parameters

What is needed is:

- 1) A method to identify and/or track the behavior of a connection without assuming access to the transport devices.
- 2) A method to observe a connection in flight without introducing agents at endpoints.
- 3) a method to observe arbitrary flows at multiple points within a network and correlate the results of those observations in a consistent manner.
- 4) A method to signal and correlate transport issues to application end-to-end behavior.
- 5) A method which does not require changes to a production network in real time.
- 6) Adequate granularity in the measurement technique to provide the needed metrics.

### 2.1 Packet Trace Meets Criteria

The only instrumentation which provides enough detail to diagnose end-to-end problems is a packet trace. Packet traces do not require changes to devices in production mode because in many large EDCO networks, products are available to capture packets in passive mode. Such products continuously monitor network traffic. Often, they are used not for diagnostic reasons but for regulatory reasons. For example, there may be legal requirements to log all stock exchange transactions.

Products for packet tracing are available freely and can be used at a client host without disrupting major portions of the network.

#### 2.1.1 Limitations of Packet Capture

Even though packets are the only reliable way to provide data at the needed granularity, there are limitations with collecting packet traces in some situations. They are as follows:

#### 2.1.2 Problem Scenario 1

1. Packets are captured for analysis at places like large core switches. All packets are kept. Again, not necessarily for diagnostic reasons but for regulatory ones. For example, records of

all stock trades may need to be kept for a certain number of years.

2. When there is a problem, an analyst extracts the needed information.

3. If the extract is done incorrectly, as often happens, or the packet capture itself is incorrect, then there may be false duplicate packets which can be quite misleading and can lead to wrong conclusions. Are these real TCP duplicates? Is there congestion on the subnet? Are these retransmissions? Situations have been seen where routers incorrectly send two packets instead of one - is this such a situation?

### 2.1.3 Problem Scenario 2

1. In this scenario, packets are captured for analysis at places like a middleware box. It may be because problems are suspected with the box itself or it is a central point of the suspected failure.
2. The box may not offer any way to tailor the packet capture. "You will get what we give you, how we give it to you!" is their philosophy.
3. The packet capture incorrectly duplicates only packets going to certain nodes.
4. Again, there are false duplicate packets which can be misleading and can lead to wrong conclusions. Are these real TCP duplicates? Is there congestion on the subnet? Situations have been seen where routers incorrectly send two packets instead of one - is this such a situation?

### 3 Rationale for Proposed Solution (PDM)

The current IPv6 specification does not provide a packet sequence number or similar field in the IPv6 main header. One option might be to force all IPv6 packets to contain a Fragment Header. In packets which are entire in and of themselves, the fragment ID would be zero - that is, an atomic fragment. Why was a new destination option header defined rather than recommending that Fragment Header be used?

Our reasoning was that the PDM destination option header would provide multiple benefits : the packet sequence number and the timestamp to calculate response time. See draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01 [ELKRSP].

### 4 Backward Compatibility

The scheme proposed in this document is backward compatible with all the currently defined IPv6 extension headers. According to RFC2460 [RFC2460], if the destination node does not recognize this option, it should skip over this option and continue processing the header.

## 5 Security Considerations

No security considerations are seen.

## 6 IANA Considerations

There are no IANA considerations.

## 7 References

### 7.1 Normative References

- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.
- [RFC0793] Postel, J., "Transmission Control Protocol", STD 7, RFC 793, September 1981.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
- [ACKPDM] Ackermann, M., "draft-ackermann-tictoc-pdm-ntp-usage-00", Internet Draft, September 2013.
- [ELKPDM] Elkins, N., "draft-elkins-6man-ipv6-pdm-dest-option-02", Internet Draft, September 2013.
- [ELKRSP] Elkins, N., "draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01", Internet Draft, September 2013.
- [ELKUSE] Elkins, N., "draft-elkins-v6ops-ipv6-pdm-recommended-usage-01", Internet Draft, September 2013
- [ELKIPPM] Elkins, N., "draft-elkins-ippm-pdm-metrics-00", Internet Draft, September 2013.

## 8 Acknowledgments

The authors would like to thank Rick Troth and Fred Baker for their comments.

Authors' Addresses

Nalini Elkins  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States  
Phone: +1 831 659 8360  
Email: [nalini.elkins@insidethestack.com](mailto:nalini.elkins@insidethestack.com)  
<http://www.insidethestack.com>

Michael S. Ackermann  
Blue Cross Blue Shield of Michigan  
P.O. Box 2888  
Detroit, Michigan 48231  
United States  
Phone: +1 310 460 4080  
Email: [mackermann@bcbsmi.com](mailto:mackermann@bcbsmi.com)  
<http://www.bcbsmi.com>

Keven Haining  
US Bank  
16900 W Capitol Drive  
Brookfield, WI 53005  
United States  
Phone: +1 262 790 3551  
Email: [keven.haining@usbank.com](mailto:keven.haining@usbank.com)  
<http://www.usbank.com>

Sigfrido Perdomo  
Depository Trust and Clearing Corporation  
55 Water Street  
New York, NY 10055  
United States  
Phone: +1 917 842 7375  
Email: [s.perdomo@dtcc.com](mailto:s.perdomo@dtcc.com)  
<http://www.dtcc.com>

William Jouris  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States  
Phone: +1 925 855 9512  
Email: [bill.jouris@insidethestack.com](mailto:bill.jouris@insidethestack.com)

INTERNET-DRAFT  
Intended Status: Informational

N. Elkins  
Inside Products  
M. Ackermann  
BCBS Michigan  
W. Jouris  
Inside Products  
K. Haining  
US Bank  
S. Perdomo  
DTCC  
October 3, 2013

Expires: April 2014

Recommended Usage of IPv6 PDM Option  
draft-elkins-v6ops-ipv6-pdm-recommended-usage-01

Abstract

To diagnose performance and connectivity problems, metrics on real (non-synthetic) transmission are critical for timely end-to-end problem resolution. Such diagnostics may be real-time or after the fact, but must not impact an operational production network. The base metrics are: packet sequence number and packet timestamp. Metrics derived from these will be described separately. This document details the recommended usage for the IPv6 Performance and Diagnostic Metrics Destination Option (PDM).

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at  
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at  
<http://www.ietf.org/shadow.html>



## Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1	Introduction . . . . .	3
2	How to use Packet Sequence Number . . . . .	3
2.1	Limitations of Packet Capture . . . . .	4
2.1.1	Problem Scenario 1 . . . . .	4
2.1.2	Problem Scenario 2 . . . . .	4
2.1.3	Packet Sequence Number Provides Solution . . . . .	5
2.2	Packet Sequence Number Replaces IPv4 IPID DeFacto Diagnostic Function . . . . .	5
3	How to use the Timestamp . . . . .	5
3.1	Time Synchronization . . . . .	5
3.2	Response Time for Service Level Agreements . . . . .	5
3.3	Trending . . . . .	5
4	Security Considerations . . . . .	6
5	IANA Considerations . . . . .	6
6	References . . . . .	6
6.1	Normative References . . . . .	6
	Authors' Addresses . . . . .	7

## 1 Introduction

To diagnose performance and connectivity problems, metrics on real (non-synthetic) transmission are critical for timely end-to-end problem resolution. Such diagnostics may be real-time or after the fact, but must not impact an operational production network. The base metrics are: packet sequence number and packet timestamp. Metrics derived from these will be described separately.

For background, please see draft-ackermann-tictoc-pdm-ntp-usage-00 [ACKPDM], draft-elkins-6man-ipv6-pdm-dest-option-02 [ELKPDM], draft-elkins-v6ops-ipv6-packet-sequence-needed-01 [ELKPSN], draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01 [ELKRSP], and draft-elkins-ippm-pdm-metrics-00 [ELKIPPM]. These drafts are companions to this document.

As discussed in the above Internet Drafts, current methods are inadequate for these purposes because they assume unreasonable access to intermediate devices, are cost prohibitive, require infeasible changes to a running production network, or do not provide timely data. The IPv6 Performance and Diagnostic Metrics destination option (PDM) provides a solution to these problems. This document will discuss how best to use the PDM.

## 2 How to use Packet Sequence Number

In many large Enterprise Networks, during network diagnostics of an end-to-end connection, it becomes necessary to find the device along the path creating problems. Diagnostic data may be collected at multiple places along the path (if possible), or at the source and destination. Then, the diagnostic data must be matched. Packet sequence number is critical in this matching process. In IPv4 networks, the IPID field was used as a DeFacto sequence number. This was discussed at length in draft-elkins-v6ops-ipv6-packet-sequence-needed-01 [ELKPSN].

The packet sequence number provided in the PDM may be used in large multi-tier networks to see where the packet loss or packet corruption is happening. Multi-tier networks are those which have multiple routers or switches on the path between the sender and the receiver.

## 2.1 Limitations of Packet Capture

As discussed in draft-elkins-v6ops-ipv6-packet-sequence-needed-01 [ELKPSN], the only instrumentation which provides enough detail to diagnose end-to-end problems is a packet trace. Even though packets are the only reliable way to provide data at the needed granularity, there are limitations in operations on live production networks. How packet sequence number can alleviate the limitations are detailed below the problem description.

### 2.1.1 Problem Scenario 1

1. Packets are captured for analysis at places like large core switches. All packets are kept. Again, not necessarily for diagnostic reasons but for regulatory. (Ex. Records of all stock trades may need to be kept for a certain number of years.)
2. When there is a problem, an analyst extracts the needed information.
3. If the extract is done incorrectly, as often happens, or the packet capture itself is incorrect, then there are false duplicate packets which can be quite misleading and can lead to wrong conclusions. Are these real TCP duplicates? Is there congestion on the subnet? Are these retransmissions? Situations have been seen where routers incorrectly send two packets instead of one - is this such a situation?

### 2.1.2 Problem Scenario 2

Packet captures can be misleading for another.

1. In this scenario, packets are captured for analysis at places like a middleware box. The reason this is done is because problems are suspected with the box itself.
2. The box may not offer any way to tailor the packet capture. "You will get what we give you, how we give it to you!" is their philosophy,
3. The packet capture incorrectly duplicates only packets going to certain nodes. So, it is not possible to devise an algorithm or pattern whereby certain packets can be ignored
4. Again, there are false duplicate packets which can be misleading and can lead to wrong conclusions. Are these real TCP duplicates? Is there congestion on the subnet? Situations have been seen where routers incorrectly send two packets instead of one - is this such a

situation?

#### 2.1.3 Packet Sequence Number Provides Solution

If a packet is a duplicate sent by a stack at a source host, the packet sequence number will not be the same. If a duplicate packet is seen with the same packet sequence number, it can be safely assumed that this is a 'false' duplicate and can be ignored.

#### 2.2 Packet Sequence Number Replaces IPv4 IPID DeFacto Diagnostic Function

draft-elkins-v6ops-ipv6-packet-sequence-needed-01 [ELKPSN] discussed a number of use cases where the IPv4 IPID reduced the time to diagnose problems on networks. The packet sequence number in the PDM will serve the same function for IPv6. The recommendation is to have the PDM used for all packets for all protocols so that timely diagnosis can occur.

### 3 How to use the Timestamp

The timestamp contained in the PDM traveling along with the packet will be used to calculate end-to-end response time without requiring agents in devices along the path. The need for end-to-end response time, the background and current methods are discussed in draft-draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01 [ELKRSP].

#### 3.1 Time Synchronization

The timestamp used in the PDM is compatible with the Network Time Protocol (NTP) [RFC5905]. Many networks use NTP pervasively today. We recommend use of NTP so that the matching of timestamps and calculations of deltas can be easily done.

#### 3.2 Response Time for Service Level Agreements

In Networks, end-to-end response times are a critical component of Service Levels Agreements (SLAs). So, the recommended use of the PDM is to have it turned on for all applications which require SLAs and / or have a requirement for timely transmission.

#### 3.3 Trending

In addition to the need for tracking current service, end-to-end response time is valuable for capacity planning. By tracking response times, and identifying trends, it becomes possible to determine when network capacity is being approached. To allow tracking of trends in response time, we recommend having the PDM used

for applications which may need additional capacity so that summary data on response times and their distributions can be maintained.

#### 4 Security Considerations

There are no security considerations.

#### 5 IANA Considerations

There are no IANA considerations.

#### 6 References

##### 6.1 Normative References

- [RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.
- [ACKPDM] Ackermann, M., "draft-ackermann-tictoc-pdm-ntp-usage-00", Internet Draft, September 2013.
- [ELKPDM] Elkins, N., "draft-elkins-6man-ipv6-pdm-dest-option-02", Internet Draft, September 2013.
- [ELKRSP] Elkins, N., "draft-elkins-v6ops-ipv6-end-to-end-rt-needed-01", Internet Draft, September 2013.
- [ELKPSN] Elkins, N., "draft-elkins-v6ops-ipv6-packet-sequence-needed-01", Internet Draft, September 2013.
- [ELKIPPM] Elkins, N., "draft-elkins-ippm-pdm-metrics-00", Internet Draft, September 2013.

#### 7 Acknowledgments

The authors would like to thank David Boyes and Rick Troth for their comments.

Authors' Addresses

Nalini Elkins  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States  
Phone: +1 831 659 8360  
Email: [nalini.elkins@insidethestack.com](mailto:nalini.elkins@insidethestack.com)  
<http://www.insidethestack.com>

Michael S. Ackermann  
Blue Cross Blue Shield of Michigan  
P.O. Box 2888  
Detroit, Michigan 48231  
United States  
Phone: +1 310 460 4080  
Email: [mackermann@bcbsmi.com](mailto:mackermann@bcbsmi.com)  
<http://www.bcbsmi.com>

Keven Haining  
US Bank  
16900 W Capitol Drive  
Brookfield, WI 53005  
United States  
Phone: +1 262 790 3551  
Email: [keven.haining@usbank.com](mailto:keven.haining@usbank.com)  
<http://www.usbank.com>

Sigfrido Perdomo  
Depository Trust and Clearing Corporation  
55 Water Street  
New York, NY 10055  
United States  
Phone: +1 917 842 7375  
Email: [s.perdomo@dtcc.com](mailto:s.perdomo@dtcc.com)  
<http://www.dtcc.com>

William Jouris  
Inside Products, Inc.  
36A Upper Circle  
Carmel Valley, CA 93924  
United States  
Phone: +1 925 855 9512  
Email: [bill.jouris@insidethestack.com](mailto:bill.jouris@insidethestack.com)  
<http://www.insidethestack.com>

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 13, 2014

C. Grundemann  
C. Donley  
CableLabs  
J. Brzozowski  
Comcast Cable Communications  
L. Howard  
Time Warner Cable  
V. Kuarsingh  
Rogers Communications  
July 12, 2013

A Near Term Solution for Home IP Networking (HIPnet)  
draft-grundemann-hipnet-00

## Abstract

Home networks are becoming more complex. With the launch of new services such as home security, IP video, Smart Grid, etc., many Service Providers are placing additional IPv4/IPv6 routers on the subscriber network. This document describes a self-configuring home router that is capable of operating in such an environment, and that requires no user interaction to configure it. Compliant with draft-ietf-homenet-arch, it uses existing protocols in new ways without the need for a routing protocol.

## Requirements Language

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].

## 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 January 13, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	3
2. Terminology . . . . .	3
3. Architecture . . . . .	4
3.1. Current End-User Network Architecture . . . . .	5
3.2. HIPNet End-User Network Architecture . . . . .	5
4. Network Detection . . . . .	6
4.1. Edge Detection . . . . .	6
4.2. Directionless Home Routers . . . . .	7
5. Routing and Addressing . . . . .	9
5.1. Recursive Prefix Delegation . . . . .	9
5.2. Prefix Sub-Delegation Requirements . . . . .	11
5.3. Multiple Address Family Support . . . . .	11
5.4. Hierarchical Routing . . . . .	12
6. Multiple ISPs . . . . .	12
6.1. Backup Connection . . . . .	12
6.2. Multi-homing . . . . .	13
6.2.1. Multihoming Requirements . . . . .	15
7. Multicast Support . . . . .	15
7.1. Service Discovery . . . . .	15
7.2. Multicast Proxy Support . . . . .	15
7.3. Multicast Requirements . . . . .	15
8. Firewall Support . . . . .	16
8.1. Requirements . . . . .	17
9. Running Code . . . . .	18
10. IANA Considerations . . . . .	18
11. Security Considerations . . . . .	18
12. Acknowledgements . . . . .	18
13. References . . . . .	18
13.1. Normative References . . . . .	18
13.2. Informative References . . . . .	19



Authors' Addresses . . . . . 20

## 1. Introduction

This document expands upon [I-D.ietf-v6ops-6204bis] to describe IPv6/IPv4 features for a residential or small-office router, referred to as a HIPnet router. Consistent with [I-D.ietf-homenet-arch], it focuses on network technology evolution to support increasingly large residential/SoHo networks. While the primary focus is on IPv6 support, this document also describes how to leverage IPv6 to configure IPv4 in a manner better than nested NATs in operation on many networks today.

This document specifies how a HIPnet router automatically detects both the edge of the customer network and its upstream interface, how it subdivides an IPv6 prefix to distribute to downstream routers, and how it leverages IPv6 address assignment to distribute IPv4 addresses. It also discusses how such a router can operate with a backup ISP or limited multihoming across two ISPs.

This document is an update to and replacement of [I-D.grundemann-homenet-hipnet].

### 1.1. Requirements Language

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].

## 2. Terminology

End-User Network	one or more links attached to the HIPnet router that connect IPv6 and IPv4 hosts.
Home IP Network (HIPnet) Router	a node intended for home or small-office use that forwards packets not explicitly addressed to itself.
Customer Edge Router (CER)	a HIPnet router that connects the end-user network to a service provider network.
Internal Router	an additional HIPnet router deployed in the home or small-office network that is not attached to a service provider network. Note that this is a functional role; it is expected that there will not be a difference in hardware or software between a CER and IR, except in such cases when a

	CER has a dedicated non-Ethernet WAN interface (e.g. DSL/cable/ LTE modem) that would preclude it from operating as an IR.
Down Interface	a HIPnet router's attachment to a link in the end-user network on which it distributes addresses and/or prefixes. Examples are Ethernet (simple or bridged), 802.11 wireless, or other LAN technologies. A HIPnet router may have one or more network-layer down interfaces.
downstream router	a router directly connected to a HIPnet router's Down Interface.
Service Provider	an entity that provides access to the Internet. In this document, a service provider specifically offers Internet access using IPv6, and may also offer IPv4 Internet access. The service provider can provide such access over a variety of different transport methods such as DSL, cable, wireless, and others.
Up Interface	a HIPnet router's attachment to a link where it receives one or more IP addresses and/or prefixes. This is also the link to which the HIPnet router points its default route.
depth	the number of layers of routers in a network. A single router network would have a depth of 1, while a router behind a router behind a router would have a depth of 3.
width	The number of routers that can be directly subtended to an upstream router. A network with three directly attached routers behind the CER would have a width of 3.

### 3. Architecture

### 3.1. Current End-User Network Architecture

An end-user network will likely support both IPv4 and IPv6. A typical end-user network consists of a "plug and play" router with IPv4 NAT functionality and a single link behind it, connected to the service provider network.

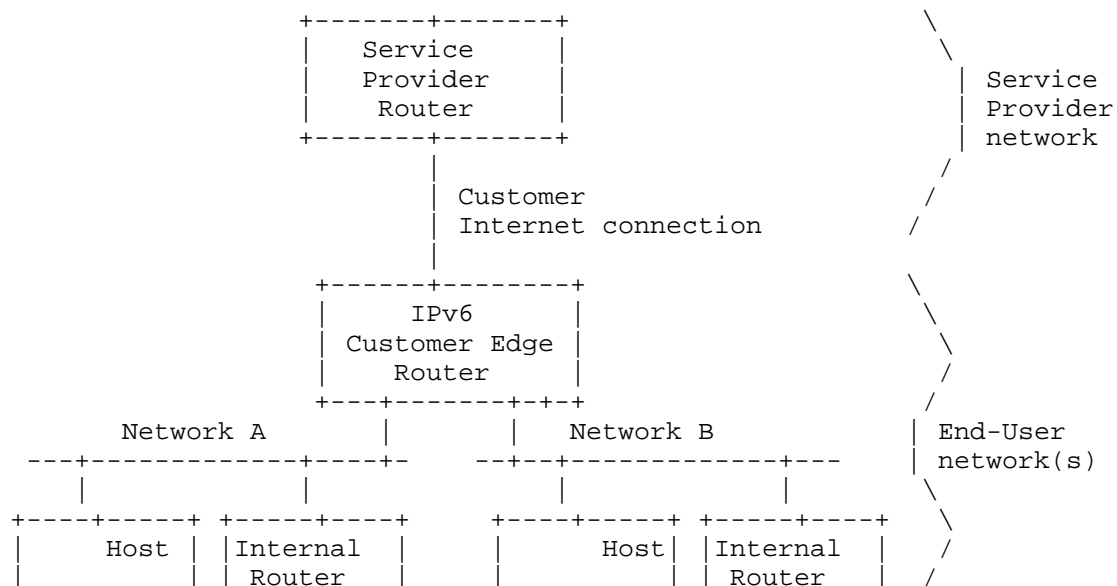
A typical IPv4 NAT deployment by default blocks all incoming connections. Opening of ports is typically allowed using a Universal Plug and Play Internet Gateway Device (UPnP IGD) [UPnP-IGD] or some other firewall control protocol.

Rewriting addresses on the edge of the network allows for some rudimentary multihoming, even though using NATs for multihoming does not preserve connections during a fail-over event [RFC4864].

Many existing routers support dynamic routing, and advanced end-users can build arbitrary, complex networks using manual configuration of address prefixes combined with a dynamic routing protocol.

### 3.2. HIPNet End-User Network Architecture

The end-user network architecture should provide equivalent or better capabilities and functionality than the current architecture. However, as end-user networks become more complex, the HIPnet architecture needs to support more complicated networks. Figure 1 illustrates the model topology for the end-user network.



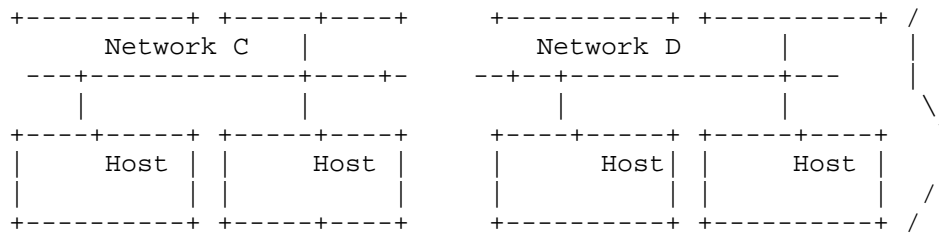


Figure 1: Example End-User Network

This architecture describes the following:

- o Prefix subdelegation supporting multiple subnets and routers
- o Border Detection
- o Router directionality supporting a hierarchical network
- o Multicast forwarding rules to support common service discovery protocols

While routers described in this document may be manually configured in an arbitrary topology with or without a dynamic routing protocol, this document only addresses automatic provisioning and configuration.

#### 4. Network Detection

In multirouter home networks, routers have to determine where they fit in the topology - whether they are at the edge or internal, and which interface is up (that is, which interface points out of the network).

##### 4.1. Edge Detection

Customer Edge Routers (CER) will often be required to behave differently from Internal Routers (IR) in several capacities. Some examples include: Firewall settings, IPv4 NAT, ULA generation (if supported), name services, multicast forwarding differences, and others. This is a functional role, and will not typically be differentiated by hardware/software (i.e. end users will not purchase a specific CER model of router distinct from IR models).

There are three methods that a router can use to determine if it is a CER for its given network:

"/48 Check" Service providers will provide IPv6 WAN addresses (DHCPv6 IA\_NA) and IPv6 prefixes (DHCPv6 IA\_PD) from different pools of addresses. The largest IPv6 prefix that we can expect to be delegated to a home router is a /48. Combining these two observations, a home router can compare the WAN address assigned to it with the prefix delegated to it to determine if it is attached directly to a service provider network. If the router is a CER, the WAN address will be from a different /48 than the prefix. If the router is an IR, the WAN address will be from the same /48 as the prefix. In this way, the router can determine if it is receiving an "external" prefix from a service provider or an "internal" prefix from another home router.

**CER\_ID** A home router can use the CER\_ID DHCPv6 option defined in [I-D.donley-dhc-cer-id-option] to determine if it is a CER or an IR. ISPs will not set the CER\_ID option, but the first CPE router sets its address in the option and other routers forward the completed CER\_ID to subdelegated routers.

**Physical** Some routers will have a physical differentiator built into them by design that will indicate that they are a CER. Examples include mobile routers, DSL routers, and cable eRouters. In the case of a mobile router, the presence of an active cellular connection indicates that the router is at the customer edge. Likewise, for an eRouter, the presence of an active DOCSIS(R) link tells the router that it is at the customer edge.

HIPnet routers can (and likely will) use more than one of the above techniques in combination to determine the edge. For example, an internal router will check for the CER\_ID option, but will also use the 48 check in case its upstream router does not support CER\_ID.

#### 4.2. Directionless Home Routers

As home networks grow in complexity and scale, it will become more common for end users to make mistakes with the physical connections between multiple routers in their home or small office. This is likely to produce loops and improper uplink connections. While we can safely assume that home networks will become more complex over time, we cannot make the same assumption of the users of home networks. Therefore, home routers will need to mitigate these physical topology problems and create a working multi-router home network dynamically, without any end user intervention.

Legacy home routers with a physically differentiated uplink port are "directional;" they are pre-set to route from the 'LAN' or Internal ports to a single, pre-defined uplink port labeled "WAN" or "Internet". This means that an end-user can make a cabling mistake

which renders the router unusable (e.g. connecting two router's uplink ports together). On the other hand, in enterprise and service provider networks, routers are "directionless;" that is to say they do not have a pre-defined 'uplink' port. While directional routers have a pre-set routing path, directionless routers are required to determine routing paths dynamically. Dynamic routing is often achieved through the implementation of a dynamic routing protocol, which all routers in a given network or network segment must support equally. This section introduces an alternative to dynamic routing protocols (such as OSPF) for creating routing paths on the fly in directionless home routers.

Note that some routers (e.g. those with a dedicated wireless/DSL/DOCSIS WAN interface) may continue to operate as directional routers. The HIPnet mechanism described below is intended for general-purpose routers.

The HIPnet mechanism uses address acquisition as described in [I-D.ietf-v6ops-6204bis] and various tiebreakers to determine directionality (up vs. down) and by so doing, creates a logical hierarchy (cf. [I-D.chakrabarti-homenet-prefix-alloc]) from any arbitrary physical topology:

1. After powering on, the HIPnet router sends Router Solicitations (RS) ([RFC4861] on all interfaces (except Wi-Fi\*))
2. Other routers respond with Router Advertisements (RA)
3. Router adds any interface on which it receives an RA to the candidate 'up' list
4. The router initiates DHCPv6 PD on all candidate 'up' interfaces. If no RAs are received, the router generates a /48 ULA prefix.
5. The router evaluates the offers received (in order of preference):
  - a) Valid GUA preferred (preferred/valid lifetimes >0)
  - b) Internal prefix preferred over external (for failover - see Section [6.1])
  - c) Largest prefix (e.g. /56 preferred to /60)
  - d) Link type/bandwidth (e.g. Ethernet vs. MoCA)
  - e) First response (wait 1 s after first response for additional offers)

f) Lowest numerical prefix

6. The router chooses the winning offer as its Up Interface.

Once directionality is established, the router continues to listen for RAs on all interfaces but doesn't acquire addresses on Down Interfaces. If the router initially receives only a ULA address on its Up Interface and GUA addressing becomes available on one of its Down Interfaces, it restarts the process. If the router stops receiving RAs on its Up Interface, it restarts the process.

In all cases, the router's Up Interface becomes its uplink interface; the router acts as a DHCP client on this interface. The router's remaining interfaces are Down Interfaces; it acts as a DHCP server on these interfaces. Also, per [I-D.ietf-v6ops-6204bis], the router only sends RAs on Down Interfaces.

\*Note: By default, Wi-Fi interfaces are considered to point "down." This requires manual configuration to enable a wireless uplink, which is preferred to avoid accidental or unwanted linking with nearby wireless networks.

## 5. Routing and Addressing

HIPnet routers use DHCPv6 prefix sub-delegation ([RFC3633]) to recursively build a hierarchical network ([I-D.chakrabarti-homenet-prefix-alloc]). This approach requires no new protocols to be supported on any home routers.

Default router settings: Only CER instantiates guest network. Wifi defaults to 'down' direction, default route uses wired interface. Firewall considers Wifi an inside port. Wi-Fi bridged with first wired Down Interface.

### 5.1. Recursive Prefix Delegation

Once directionality is established, the home router will acquire a WAN IPv6 address and an IPv6 prefix per [I-D.ietf-v6ops-6204bis]. As HIPnet routers (other than the CER) do not know their specific location in the hierarchical network, all HIPnet routers use the same generic rules for recursive prefix delegation to facilitate route aggregation, multihoming, and IPv4 support (described below). This methodology expounds upon that previously described in [I-D.chakrabarti-homenet-prefix-alloc].

The process can be illustrated in the following way:

1. Per [I-D.ietf-v6ops-6204bis], the HIPnet router assigns a separate /64 from its delegated prefix(es) for each of its Down Interfaces in numerical order, starting from the numerically lowest.
2. If the received prefix is too small to number all Down Interfaces, the router collapses them into a single interface, assigns a single /64 to that interface, and logs an error message.
3. The HIPnet router subdivides the IPv6 prefix received via DHCPv6 ([RFC3315]) into sub-prefixes. To support a suggested depth of three routers, with as large a width as possible, it is recommended to divide the prefix on 2-, 3-, or 4-bit boundaries. If the received prefix is not large enough, it is broken into as many /64 sub-prefixes as possible and logs an error message. By default, this document suggests that the router divide the delegated prefix based on the aggregate prefix size and the HIPnet router's number of physical Down Interfaces. This is to allow for enough prefixes to support a downstream router on each down port.
  - \* If the received prefix is smaller than a /56 (e.g. a /60),
    - + 8 or more port routers divide on 3-bit boundaries (e.g. /63).
    - + 7 or fewer port routers divide on 2-bit boundaries (e.g. /62).
  - \* If the received prefix is a /56 or larger,
    - + 8 or more port routers divide on 4-bit boundaries (e.g. /60).
    - + 7 or fewer port routers divide on 3-bit boundaries (e.g. /59).
4. The HIPnet router delegates remaining prefixes to downstream routers per [RFC3633] in reverse numerical order, starting with the numerically highest. This is to minimize the renumbering impact of enabling an inactive interface.

For example, a four port router with two LANs (two Down Interfaces) that receives 2001:db8:0:b0::/60 would start by numbering its two Down Interfaces with 2001:db8:0:b0::/64 and 2001:db8:0:b1::/64 respectively, and then begin prefix delegation by giving 2001:db8:0:bc::/62 to the first directly attached downstream router.



## 5.2. Prefix Sub-Delegation Requirements

- PSD-1: The HIPnet router MUST support [I-D.ietf-v6ops-6204bis] address acquisition and LAN addressing.
- PSD-2: The HIPnet router MUST support Delegating Router behavior for the IA-PD Option [RFC3633] on all Down Interfaces.
- PSD-3: HIPnet routers MUST NOT act as both a DHCP client and server on the same link.
- PSD-4: The HIPnet router MAY support other methods of dividing the received prefix.
- PSD-5: The HIPnet router MUST delegate prefixes of the same size to downstream routers.
- PSD-6: Per [I-D.ietf-v6ops-6204bis] L-2, the HIPnet router allocates a /64 to each Down Interface. The HIPnet router SHOULD allocate these /64 interface-prefixes in numerical order, starting with the lowest.
- PSD-7: If there are insufficient /64s for each Down Interface, the HIPnet router SHOULD assign the lowest numbered /64 for all Down Interfaces and log an error message.
- PSD-8: The HIPnet router MAY reserve additional /64 interface-prefixes for interfaces that will be enabled in the future.
- PSD-9: The HIPnet router SHOULD delegate sub-prefixes to downstream routers starting from the numerically highest sub-prefix and working down in reverse numerical order.
- PSD-10: If there are not enough sub-prefixes remaining to delegate to all downstream routers, the HIPnet router SHOULD log an error message.

## 5.3. Multiple Address Family Support

The recursive prefix delegation method described above can be extended to support additional address types such as IPv4, additional GUAs, or ULAs. When the HIPnet router receives its prefix via DHCPv6 ([RFC3633]), it computes its 8 or 16-bit Link ID (bits 56-64 or 48-64) from the received IA\_PD. It then prepends additional prefixes received in one or more IPv6 Router Advertisements ([RFC4861]) or from the DHCPv4-assigned ([RFC2131]) IPv4 network address received on the Up Interface.

As the network is hierarchical, upstream routers know the Link ID for each downstream router, and know the prefix(es) on each LAN segment. Accordingly, HIPnet routers automatically calculate downstream routes to all downstream routers.

In networks using this mechanism for IPv4 provisioning, it is suggested that the CER use addresses in the 10.0.0.0/8 ([RFC1918]) range for downstream interface provisioning. When used with a 16-bit Link ID, this results in an IPv4 /24 created for each LAN segment (8 network bits plus 16 Link ID bits equals a 24 bit subnet mask).

#### 5.4. Hierarchical Routing

The recursive prefix delegation method described above, coupled with "up detection", enables very simple hierarchical routing. By this we mean that each router installs a single default 'up' route and a more specific 'down' route for each prefix delegated to a downstream IR. Each of these 'down' routes simply points all packets destined to a given prefix to the WAN IP address of the router to which that prefix was delegated. This combination of a default 'up' route and more specific 'down' routes provides complete reachability within the home network with no need for any additional message exchange or routing protocol support.

#### 6. Multiple ISPs

HIPnet routers can support either active/standby multihoming with multiple ISPs or limited active/active multihoming without a routing protocol.

##### 6.1. Backup Connection

Using the procedure described above, multi-router home networks with multiple ISP connections can easily operate in an active/standby manner, switching from one Internet connection to the other when the active connection fails. Lacking a default priority, HIPnet routers will have to default to a "first online" method of primary CER selection. In other words, by default, the first CER to come online becomes the primary CER and the second CER to turn on becomes the backup. In this text, the primary ISP is the ISP connected to the primary CER and the backup ISP is simply the ISP attached to the backup CER.

In an active/standby multi-ISP scenario, a backup CER sets its Up Interface to point to the primary CER, not the backup ISP. Hence, it does not acquire or advertise the backup ISP prefix. Instead, it discovers the internally advertised GUA prefix being distributed by the currently active primary CER.

In the case of a primary ISP failure, per [I-D.ietf-v6ops-6204bis], the CER sends an RA advertising the preferred lifetime as 0 for the ISP-provided prefix, and its router lifetime as 0. The backup CER becomes active when it sees the primary ISP GUA prefix advertised with a preferred lifetime of 0. In the case of CER failure, if the backup CER sees the Primary CER stop sending RAs altogether, the Backup CER becomes active.

When the backup CER becomes active, it obtains and advertises its own external GUA. When advertising the GUA delegated by its ISP, the backup CER sets the valid, preferred, and router lifetimes to a value greater than 0. Other routers see this and re-determine the network topology via "up" detection, placing the new CER at the root of the new hierarchical tree.

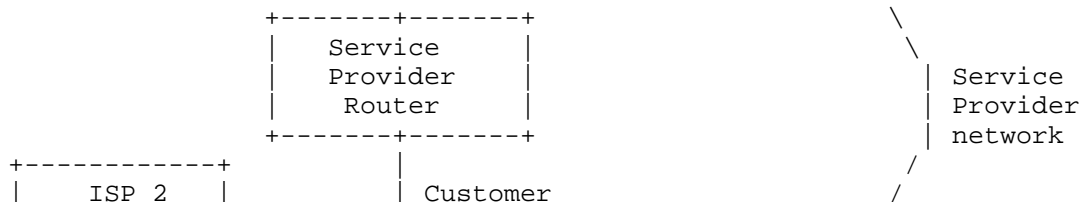
Using this approach, manual intervention may be required to transition back to the primary ISP. This prevents flapping in the event of intermittent network failures. Another alternative is to have a user-defined priority, which would facilitate pre-emption.

## 6.2. Multi-homing

The HIPnet algorithm also allows for limited active/active multihoming in two cases:

1. When one ISP router is used as the primary connection and the second ISP router is used for limited connectivity e.g. for a home office
2. When both ISP routers are connected to the same LAN segment at the top of the tree.

In case 1, the subscriber has a primary ISP connection and a secondary connection used for a limited special purpose. (e.g. for work VPN, video network, etc.). Devices connected under the secondary network router access the Internet through the secondary ISP. All devices still have access to all network resources in the home. Devices under the secondary connection can use the primary ISP if the secondary fails, but other devices do not use the secondary ISP.



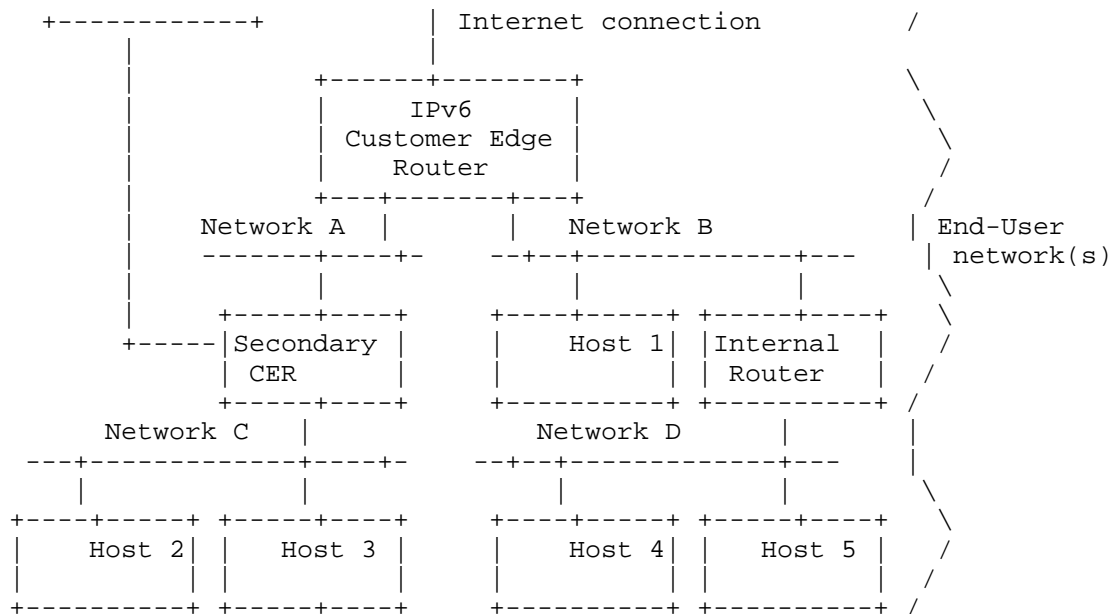


Figure 2: An Example of a multihomed End-User Network

As described above, the primary CER performs prefix sub-delegation to create the hierarchical tree network. The secondary edge router then obtains a second prefix from ISP2 and advertises the ISP2 prefix as part of its RA. The Secondary CER thus includes sub-prefixes from both ISPs in all IA\_PD messages to downstream routers with the same "router id.". In a change from the single-homing (or backup router) case, the secondary CER points its default route to ISP2, and adds an internal /48 route to its upstream internal router (e.g. R1). Devices below the the secondary CER (e.g. Host 2, Host 3) use ISP2, but have full access to all internal devices using the ISP1 prefix (and/or ULAs). If the ISP2 link fails, the secondary CER points its default route 'up' and traffic flows to ISP1. Devices not below the secondary CER (e.g. Hosts 1, 4, 5) use ISP1, but have full access to all internal devices using the ISP1 prefix (or ULAs).

In case 2, the secondary CER is installed on the same LAN segment as the primary CER. As above, it acquires a prefix from both the CER and secondary ISP. Since it is on the same LAN segment as the CER, the secondary CER does not delegate prefixes to that interface via DHCP. However, it does generate an RA for the ISP2 prefix on the LAN.

As described above, downstream routers receiving the secondary CER RA acquire an address using SLAAC and generate a prefix for sub-

delegation by prepending the secondary CER prefix with the router ID generated during the receipt of the prefix from the CER. Such routers then generate their own RAs on downstream interfaces and include the secondary prefix as an IA\_PD option in future prefix delegations.

#### 6.2.1. Multihoming Requirements

MH-1: HIPnet routers configured for active multi-ISP support MUST support DHCP address/prefix acquisition (per [I-D.ietf-v6ops-6204bis] on two interfaces (their WAN and upstream LAN interfaces).

MH-2: HIPnet routers configured for active multi-ISP support MAY route packets based on the source IP address of incoming packets using [RFC6724] logic. This allows traffic sourced from the first ISP prefix to be directed to the first ISP, and traffic sourced from the second ISP prefix to be directed to the second ISP.

MH-3: HIPnet routers configured for active multi-ISP support MUST advertise RAs for prefixes on all interfaces except the one from which the prefix was acquired or one directly attached to a Service Provider network.

### 7. Multicast Support

#### 7.1. Service Discovery

There are several common service discovery protocols such as mDNS [RFC6762]/DNS-SD [RFC6763] and SSDP [SSDP]. In a multirouter network, service discovery needs to work across the entire home network (e.g. site-scoped rather than link-scoped). This requires that HIPnet routers forward relevant multicast traffic appropriately, to enable service discovery across the home network.

#### 7.2. Multicast Proxy Support

In addition to multicast support for service discovery, it is recommended that HIPnet routers support external multicast traffic.

#### 7.3. Multicast Requirements

MULTI-1: A HIPnet router MUST discard IP multicast packets that fail a Reverse Path Forwarding Check (RPFC).

MULTI-2: A HIPnet router that determines itself to be at the edge of a home network (e.g. via CER\_ID option, /48 verification, or other mechanism) MUST NOT forward IPv4 administratively scoped (239.0.0.0/8) packets onto the WAN interface.

MULTI-3: HIPnet Routers MUST forward IPv4 Local Scope multicast packets (239.255.0.0/16) to all LAN interfaces except the one from which they were received.

MULTI-4: A HIPnet router that determines itself to be at the edge of a home network (e.g. via CER\_ID option, /48 verification, or other mechanism) MUST NOT forward site-scope (FF05::) IPv6 multicast packets onto the WAN interface.

MULTI-5: HIPnet routers MUST forward site-scoped (FF05::/16) IPv6 multicast packets to all LAN interfaces except the one from which they were received.

MULTI-6: A home router MAY discard IP multicast packets sent between Down Interfaces (different VLANs).

MULTI-7: HIPnet routers SHOULD support an IGMP/MLD proxy, as described in [RFC4605].

## 8. Firewall Support

In a home network, routers need to be equipped with stateful firewall capabilities. Home routers will need to provide "on by default" security where incoming traffic is limited to return traffic resulting from outgoing packets. They also need to allow users to create inbound 'pinholes' for specific purposes, such as online gaming, manually similar to those described in Simple Security ([RFC6092]). "Advanced Security" [I-D.vyncke-advanced-ipv6-security] features optionally could be added to provide intrusion detection (IDS/IPS) support.

Local Network Protection for IPv6 [RFC4864] recommends firewall functions that replace NAT security and calls for simple security. Simple Security [RFC6092] defines firewall filtering rules for IPv6 traffic. Advanced Security [I-D.vyncke-advanced-ipv6-security] supports the concept of end-to-end IPv6 reachability and uses adaptive filtering based on Intrusion Prevention System (IPS) functions.

It is recommended that the CER enable a stateful [RFC6092] firewall by default. IRs have three options described below:

IR Firewall Option 1 - Filtering Disabled: Once a home router determines that it is not the CER, it disables its firewall and allows all traffic to pass. The advantages of this approach are that it is simple and easy to troubleshoot and it facilitates whole-home service discovery and media sharing. The disadvantages are that it does not protect home devices from each other (e.g. infected machines could affect entire home network).

IR Firewall Option 2 - Simple Security + PCP: Home routers have a [RFC6092] firewall on by default, regardless of CER/IR status but IRs allow "pin-holing" using PCP [RFC6887]. CERs can restrict opening PCP pinholes on the up interface. The advantages of this approach are that it protects the home network from internal threats in other LAN segments and it mirrors legacy IPv4 router behavior. The disadvantages to this approach are that it is less predictable; it relies on application "pin-holing", a "default deny" rule that may interfere with service discovery and/or content sharing, and requires PCP clients (e.g. on PCs and CPE devices).

IR Firewall Option 3 - Advanced Security: Once a home router determines that it is not the CER, it disables its [RFC6092] firewall but activates an [I-D.vyncke-advanced-ipv6-security] firewall (IPS). The advantages to this approach are that it protects the home network from internal threats in other segments and is more predictable than Option 2, since internal traffic is allowed by default. The disadvantages are that adaptive filtering is more complex than static filtering and typically requires a "fingerprint" subscription to work well.

It is recommended that dual-stack routers configure IPv4 support to mirror IPv6, as described above.

While this section describes default router behavior, device manufacturers are encouraged to make router security options user-configurable.

#### 8.1. Requirements

SEC-1: The CER MUST enable a stateful [RFC6092] firewall by default.

SEC-2: HIPnet routers MUST only perform IPv4 NAT when serving as the CER.

SEC-3: By default, HIPnet routers SHOULD configure IPv4 firewalling rules to mirror IPv6.

SEC-4: HIPnet routers serving as CER SHOULD NOT enable UPnP IGD ([UPnP-IGD]) control by default.

## 9. Running Code

The HIPnet architecture described in this document was successfully demonstrated to work at Bits-N-Bytes in Orlando during IETF 86. The proof-of-concept software was simply a version of [OpenWRT] modified for HIPnet compliance by a small team of undergrads from the University of Colorado, Boulder. You can download the prototype/proof-of-concept software from [HIPnetPoC].

## 10. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

## 11. Security Considerations

Security considerations are discussed in the Firewall Support section above.

## 12. Acknowledgements

TBD

## 13. References

### 13.1. Normative References

- [I-D.ietf-v6ops-6204bis]  
Singh, H., Beebe, W., Donley, C., and B. Stark, "Basic Requirements for IPv6 Customer Edge Routers", draft-ietf-v6ops-6204bis-12 (work in progress), October 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.



- [RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick,  
"Internet Group Management Protocol (IGMP) / Multicast  
Listener Discovery (MLD)-Based Multicast Forwarding ("IGMP  
/MLD Proxying")", RFC 4605, August 2006.
- [RFC4864] Van de Velde, G., Hain, T., Droms, R., Carpenter, B., and  
E. Klein, "Local Network Protection for IPv6", RFC 4864,  
May 2007.
- [RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in  
Customer Premises Equipment (CPE) for Providing  
Residential IPv6 Internet Service", RFC 6092, January  
2011.

### 13.2. Informative References

- [HIPnetPoC]  
CableLabs, "HIPnetPoC", July 2013, <[http://  
www.cablelabs.com/cablemodem/ri/hipnet\\_prototype.html](http://www.cablelabs.com/cablemodem/ri/hipnet_prototype.html)>.
- [I-D.chakrabarti-homenet-prefix-alloc]  
Nordmark, E., Chakrabarti, S., Krishnan, S., and W.  
Haddad, "Simple Approach to Prefix Distribution in Basic  
Home Networks", draft-chakrabarti-homenet-prefix-alloc-01  
(work in progress), October 2011.
- [I-D.donley-dhc-cer-id-option]  
Donley, C. and C. Grundemann, "Customer Edge Router  
Identification Option", draft-donley-dhc-cer-id-option-01  
(work in progress), September 2012.
- [I-D.grundemann-homenet-hipnet]  
Grundemann, C., Donley, C., Brzozowski, J., Howard, L.,  
and V. Kuarsingh, "A Near Term Solution for Home IP  
Networking (HIPnet)", draft-grundemann-homenet-hipnet-01  
(work in progress), February 2013.
- [I-D.ietf-homenet-arch]  
Chown, T., Arkko, J., Brandt, A., Troan, O., and J. Weil,  
"Home Networking Architecture for IPv6", draft-ietf-  
homenet-arch-08 (work in progress), May 2013.
- [I-D.vyncke-advanced-ipv6-security]  
Vyncke, E., Yourtchenko, A., and M. Townsley, "Advanced  
Security for IPv6 CPE", draft-vyncke-advanced-  
ipv6-security-03 (work in progress), October 2011.
- [OpenWRT] OpenWRT, "OpenWRT", July 2013, <<http://openwrt.org/>>.

- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC2131] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.
- [RFC6724] Thaler, D., Draves, R., Matsumoto, A., and T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, September 2012.
- [RFC6762] Cheshire, S. and M. Krochmal, "Multicast DNS", RFC 6762, February 2013.
- [RFC6763] Cheshire, S. and M. Krochmal, "DNS-Based Service Discovery", RFC 6763, February 2013.
- [RFC6887] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", RFC 6887, April 2013.
- [SSDP] UPnP Forum, "SSDP", October 2008, <<http://www.upnp.org/>>.
- [UPnP-IGD] UPnP Forum, "UPnP-IGD", November 2001, <<http://www.upnp.org/>>.

#### Authors' Addresses

Chris Grundemann  
CableLabs  
858 Coal Creek Circle  
Louisville, CO 80027  
USA

Phone: +1-303-351-1539  
Email: [c.grundemann@cablelabs.com](mailto:c.grundemann@cablelabs.com)

Chris Donley  
CableLabs  
858 Coal Creek Circle  
Louisville, CO 80027  
USA

Email: c.donley@cablelabs.com

John Jason Brzozowski  
Comcast Cable Communications  
1306 Goshen Parkway  
Chester, PA 19380  
USA

Email: john\_brzozowski@cable.comcast.com

Lee Howard  
Time Warner Cable  
13241 Woodland Park Rd  
Herndon, VA 20171  
USA

Email: william.howard@twcable.com

Victor Kuarsingh  
Rogers Communications  
8200 Dixie Road  
Brampton, ON L6T 0C1  
Canada

Email: victor.kuarsingh@rci.rogers.com

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: February 1, 2015

K. Chittimaneni  
Dropbox Inc.  
T. Chown  
University of Southampton  
L. Howard  
Time Warner Cable  
V. Kuarsingh  
Dyn Inc  
Y. Pouffary  
Hewlett Packard  
E. Vyncke  
Cisco Systems  
July 31, 2014

Enterprise IPv6 Deployment Guidelines  
draft-ietf-v6ops-enterprise-incremental-ipv6-06

Abstract

Enterprise network administrators worldwide are in various stages of preparing for or deploying IPv6 into their networks. The administrators face different challenges than operators of Internet access providers, and have reasons for different priorities. The overall problem for many administrators will be to offer Internet-facing services over IPv6, while continuing to support IPv4, and while introducing IPv6 access within the enterprise IT network. The overall transition will take most networks from an IPv4-only environment to a dual stack network environment and eventually an IPv6-only operating mode. This document helps provide a framework for enterprise network architects or administrators who may be faced with many of these challenges as they consider their IPv6 support strategies.

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 February 1, 2015.

#### Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction . . . . .	3
1.1. Enterprise Assumptions . . . . .	4
1.2. IPv4-only Considerations . . . . .	4
1.3. Reasons for a Phased Approach . . . . .	5
2. Preparation and Assessment Phase . . . . .	6
2.1. Program Planning . . . . .	6
2.2. Inventory Phase . . . . .	7
2.2.1. Network infrastructure readiness assessment . . . . .	7
2.2.2. Applications readiness assessment . . . . .	8
2.2.3. Importance of readiness validation and testing . . . . .	8
2.3. Training . . . . .	9
2.4. Security Policy . . . . .	9
2.4.1. IPv6 is no more secure than IPv4 . . . . .	9
2.4.2. Similarities between IPv6 and IPv4 security . . . . .	10
2.4.3. Specific Security Issues for IPv6 . . . . .	10
2.5. Routing . . . . .	12
2.6. Address Plan . . . . .	13
2.7. Tools Assessment . . . . .	15
3. External Phase . . . . .	16
3.1. Connectivity . . . . .	16
3.2. Security . . . . .	17
3.3. Monitoring . . . . .	19
3.4. Servers and Applications . . . . .	19
3.5. Network Prefix Translation for IPv6 . . . . .	20
4. Internal Phase . . . . .	20
4.1. Security . . . . .	21
4.2. Network Infrastructure . . . . .	21
4.3. End user devices . . . . .	22
4.4. Corporate Systems . . . . .	23

5. IPv6-only . . . . .	23
6. Considerations For Specific Enterprises . . . . .	25
6.1. Content Delivery Networks . . . . .	25
6.2. Data Center Virtualization . . . . .	25
6.3. University Campus Networks . . . . .	25
7. Security Considerations . . . . .	27
8. Acknowledgements . . . . .	27
9. IANA Considerations . . . . .	27
10. Informative References . . . . .	27
Authors' Addresses . . . . .	32

## 1. Introduction

An Enterprise Network is defined in [RFC4057] as a network that has multiple internal links, one or more router connections to one or more Providers, and is actively managed by a network operations entity (the "administrator", whether a single person or department of administrators). Administrators generally support an internal network, consisting of users' workstations, personal computers, mobile devices, other computing devices and related peripherals, a server network, consisting of accounting and business application servers, and an external network, consisting of Internet-accessible services such as web servers, email servers, VPN systems, and customer applications. This document is intended as guidance for enterprise network architects and administrators in planning their IPv6 deployments.

The business reasons for spending time, effort, and money on IPv6 will be unique to each enterprise. The most common drivers are due to the fact that when Internet service providers, including mobile wireless carriers, run out of IPv4 addresses, they will provide native IPv6 and non-native IPv4. The non-native IPv4 service may be NAT64, NAT444, Dual-stack Lite, MAP-T, MAP-E, or other transition technologies. Compared to tunneled or translated service, native traffic typically performs better and more reliably than non-native. For example, for client networks trying to reach enterprise networks, the IPv6 experience will be better than the transitional IPv4 if the enterprise deploys IPv6 in its public-facing services. The native IPv6 network path should also be simpler to manage and, if necessary, troubleshoot. Further, enterprises doing business in growing parts of the world may find IPv6 growing faster there, where again potential new customers, employees and partners are using IPv6. It is thus in the enterprise's interests to deploy native IPv6, at the very least in its public-facing services, but ultimately across the majority or all of its scope.

The text in this document provides specific guidance for enterprise networks, and complements other related work in the IETF, including [I-D.ietf-v6ops-design-choices] and [RFC5375].

### 1.1. Enterprise Assumptions

For the purpose of this document, we assume:

- o The administrator is considering deploying IPv6 (but see Section 1.2 below).
- o The administrator has existing IPv4 networks and devices which will continue to operate and be supported.
- o The administrator will want to minimize the level of disruption to the users and services by minimizing number of technologies and functions that are needed to mediate any given application. In other words: provide native IP wherever possible.

Based on these assumptions, an administrator will want to use technologies which minimize the number of flows being tunnelled, translated or intercepted at any given time. The administrator will choose transition technologies or strategies which allow most traffic to be native, and will manage non-native traffic. This will allow the administrator to minimize the cost of IPv6 transition technologies, by containing the number and scale of transition systems.

Tunnels used for IPv6/IPv4 transition are expected as near/mid- term mechanisms, while IPv6 tunneling will be used for many long-term operational purposes such as security, routing control, mobility, multi-homing, traffic engineering, etc. We refer to the former class of tunnels as "transition tunnels"

### 1.2. IPv4-only Considerations

As described in [RFC6302] administrators should take certain steps even if they are not considering IPv6. Specifically, Internet-facing servers should log the source port number, timestamp (from a reliable source), and the transport protocol. This will allow investigation of malefactors behind address-sharing technologies such as NAT444, MAP, or Dual-stack Lite. Such logs should be protected for integrity, safeguarded for privacy and periodically purged within applicable regulations for log retention.

Other IPv6 considerations may impact ostensibly IPv4-only networks, e.g. [RFC6104] describes the rogue IPv6 RA problem, which may cause problems in IPv4-only networks where IPv6 is enabled in end systems

on that network. Further discussion of the security implications of IPv6 in IPv4-only networks can be found in [RFC7123]).

### 1.3. Reasons for a Phased Approach

Given the challenges of transitioning user workstations, corporate systems, and Internet-facing servers, a phased approach allows incremental deployment of IPv6, based on the administrator's own determination of priorities. This document outlines suggested phases: a Preparation and Assessment Phase, an Internal Phase, and an External Phase. The Preparation Phase is highly recommended to all administrators, as it will save errors and complexity in later phases. Each administrator must decide whether to begin with an External Phase (enabling IPv6 for Internet-facing systems, as recommended in [RFC5211]) or an Internal Phase (enabling IPv6 for internal interconnections first).

Each scenario is likely to be different to some extent, but we can highlight some considerations:

- o In many cases, customers outside the network will have IPv6 before the internal enterprise network. For these customers, IPv6 may well perform better, especially for certain applications, than translated or tunneled IPv4, so the administrator may want to prioritize the External Phase such that those customers have the simplest and most robust connectivity to the enterprise, or at least its external-facing elements.
- o Employees who access internal systems by VPN may find that their ISPs provide translated IPv4, which does not support the required VPN protocols. In these cases, the administrator may want to prioritize the External Phase, and any other remotely-accessible internal systems. It is worth noting that a number of emerging VPN solutions provide dual-stack connectivity; thus a VPN service may be useful for employees in IPv4-only access networks to access IPv6 resources in the enterprise network (much like many public tunnel broker services, but specifically for the enterprise). Some security considerations are described in [I-D.ietf-opsec-vpn-leakages].
- o Internet-facing servers cannot be managed over IPv6 unless the management systems are IPv6-capable. These might be Network Management Systems (NMS), monitoring systems, or just remote management desktops. Thus in some cases, the Internet-facing systems are dependent on IPv6-capable internal networks. However, dual-stack Internet-facing systems can still be managed over IPv4.



- o Virtual machines may enable a faster rollout once initial system deployment is complete. Management of VMs over IPv6 is still dependent on the management software supporting IPv6.
- o IPv6 is enabled by default on all modern operating systems, so it may be more urgent to manage and have visibility on the internal traffic. It is important to manage IPv6 for security purposes, even in an ostensibly IPv4-only network, as described in [RFC7123].
- o In many cases, the corporate accounting, payroll, human resource, and other internal systems may only need to be reachable from the internal network, so they may be a lower priority. As enterprises require their vendors to support IPv6, more internal applications will support IPv6 by default and it can be expected that eventually new applications will only support IPv6. The inventory, as described in Section 2.2, will help determine the systems' readiness, as well as the readiness of the supporting network elements and security, which will be a consideration in prioritization of these corporate systems.
- o Some large organizations (even when using private IPv4 addresses[RFC1918]) are facing IPv4 address exhaustion because of the internal network growth (for example the vast number of virtual machines) or because of the acquisition of other companies that often raise private IPv4 address overlapping issues.
- o IPv6 restores end to end transparency even for internal applications (of course security policies must still be enforced). When two organizations or networks merge [RFC6879], the unique addressing of IPv6 can make the merger much easier and faster. A merger may, therefore, prioritize IPv6 for the affected systems.

These considerations are in conflict; each administrator must prioritize according to their company's conditions. It is worth noting that the reasons given in one "Large Corporate User's View of IPng", described in [RFC1687], for reluctance to deploy have largely been satisfied or overcome in the intervening years.

## 2. Preparation and Assessment Phase

### 2.1. Program Planning

Since enabling IPv6 is a change to the most fundamental Internet Protocol, and since there are so many interdependencies, having a professional project manager organize the work is highly recommended. In addition, an executive sponsor should be involved in determining

the goals of enabling IPv6 (which will establish the order of the phases), and should receive regular updates.

It may be necessary to complete the Preparation Phase before determining whether to prioritize the Internal or External Phase, since needs and readiness assessments are part of that phase. For a large enterprise, it may take several iterations to really understand the level of effort required. Depending on the required schedule, it may be useful to roll IPv6 projects into other architectural upgrades--this can be an excellent way to improve the network and reduce costs. However, by increasing the scope of projects, the schedule is often affected. For instance, a major systems upgrade may take a year to complete, where just patching existing systems may take only a few months.

The deployment of IPv6 will not generally stop all other technology work. Once IPv6 has been identified as an important initiative, all projects, both new and in-progress, will need to be reviewed to ensure IPv6 support.

It is normal for assessments to continue in some areas while execution of the project begins in other areas. This is fine, as long as recommendations in other parts of this document are considered, especially regarding security (for instance, one should not deploy IPv6 on a system before security has been evaluated).

## 2.2. Inventory Phase

To comprehend the scope of the inventory phase we recommend dividing the problem space in two: network infrastructure readiness and applications readiness.

### 2.2.1. Network infrastructure readiness assessment

The goal of this assessment is to identify the level of IPv6 readiness of network equipment. This will identify the effort required to move to an infrastructure that supports IPv6 with the same functional service capabilities as the existing IPv4 network. This may also require a feature comparison and gap analysis between IPv4 and IPv6 functionality on the network equipment and software. IPv6 support will require testing; features often work differently in vendors' labs than production networks. Some devices and software will require IPv4 support for IPv6 to work.

The inventory will show which network devices are already capable, which devices can be made IPv6 ready with a code/firmware upgrade, and which devices will need to be replaced. The data collection consists of a network discovery to gain an understanding of the

topology and inventory network infrastructure equipment and code versions with information gathered from static files and IP address management, DNS and DHCP tools.

Since IPv6 might already be present in the environment, through default configurations or VPNs, an infrastructure assessment (at minimum) is essential to evaluate potential security risks.

#### 2.2.2. Applications readiness assessment

Just like network equipment, application software needs to support IPv6. This includes OS, firmware, middleware and applications (including internally developed applications). Vendors will typically handle IPv6 enablement of off-the-shelf products, but often enterprises need to request this support from vendors. For internally developed applications it is the responsibility of the enterprise to enable them for IPv6. Analyzing how a given application communicates over the network will dictate the steps required to support IPv6. Applications should avoid instructions specific to a given IP address family. Any applications that use APIs, such as the C language, that expose the IP version specifically, need to be modified to also work with IPv6.

There are two ways to IPv6-enable applications. The first approach is to have separate logic for IPv4 and IPv6, thus leaving the IPv4 code path mainly untouched. This approach causes the least disruption to the existing IPv4 logic flow, but introduces more complexity, since the application now has to deal with two logic loops with complex race conditions and error recovery mechanisms between these two logic loops. The second approach is to create a combined IPv4/IPv6 logic, which ensures operation regardless of the IP version used on the network. Knowing whether a given implementation will use IPv4 or IPv6 in a given deployment is a matter of some art; see Source Address Selection [RFC6724] and Happy Eyeballs [RFC6555]. It is generally recommended that the application developer use industry IPv6-porting tools to locate the code that needs to be updated. Some discussion of IPv6 application porting issues can be found in [RFC4038].

#### 2.2.3. Importance of readiness validation and testing

Lastly IPv6 introduces a completely new way of addressing endpoints, which can have ramifications at the network layer all the way up to the applications. So to minimize disruption during the transition phase we recommend complete functionality, scalability and security testing to understand how IPv6 impacts the services and networking infrastructure.

### 2.3. Training

Many organizations falter in IPv6 deployment because of a perceived training gap. Training is important for those who work with addresses regularly, as with anyone whose work is changing. Better knowledge of the reasons IPv6 is being deployed will help inform the assessment of who needs training, and what training they need.

### 2.4. Security Policy

It is obvious that IPv6 networks should be deployed in a secure way. The industry has learnt a lot about network security with IPv4, so, network operators should leverage this knowledge and expertise when deploying IPv6. IPv6 is not so different than IPv4: it is a connectionless network protocol using the same lower layer service and delivering the same service to the upper layer. Therefore, the security issues and mitigation techniques are mostly identical with same exceptions that are described further.

#### 2.4.1. IPv6 is no more secure than IPv4

Some people believe that IPv6 is inherently more secure than IPv4 because it is new. Nothing can be more wrong. Indeed, being a new protocol means that bugs in the implementations have yet to be discovered and fixed and that few people have the operational security expertise needed to operate securely an IPv6 network. This lack of operational expertise is the biggest threat when deploying IPv6: the importance of training is to be stressed again.

One security myth is that thanks to its huge address space, a network cannot be scanned by enumerating all IPv6 address in a /64 LAN hence a malevolent person cannot find a victim. [RFC5157] describes some alternate techniques to find potential targets on a network, for example enumerating all DNS names in a zone. Additional advice in this area is also given in [I-D.ietf-opsec-ipv6-host-scanning].

Another security myth is that IPv6 is more secure because it mandates the use of IPsec everywhere. While the original IPv6 specifications may have implied this, [RFC6434] clearly states that IPsec support is not mandatory. Moreover, if all the intra-enterprise traffic is encrypted, both malefactors and security tools that rely on payload inspection (IPS, firewall, ACL, IPFIX ([RFC7011] and [RFC7012]), etc) will be thwarted. Therefore, IPsec is as useful in IPv6 as in IPv4 (for example to establish a VPN overlay over a non-trusted network or reserved for some specific applications).

The last security myth is that amplification attacks (such as [SMURF]) do not exist in IPv6 because there is no more broadcast.

Alas, this is not true as ICMP error (in some cases) or information messages can be generated by routers and hosts when forwarding or receiving a multicast message (see Section 2.4 of [RFC4443]). Therefore, the generation and the forwarding rate of ICMPv6 messages must be limited as in IPv4.

It should be noted that in a dual-stack network the security implementation for both IPv4 and IPv6 needs to be considered, in addition to security considerations related to the interaction of (and transition between) the two, while they coexist.

#### 2.4.2. Similarities between IPv6 and IPv4 security

As mentioned earlier, IPv6 is quite similar to IPv4, therefore several attacks apply for both protocol families, including:

- o Application layer attacks: such as cross-site scripting or SQL injection
- o Rogue device: such as a rogue Wi-Fi Access Point
- o Flooding and all traffic-based denial of services (including the use of control plane policing for IPv6 traffic see [RFC6192])

A specific case of congruence is IPv6 Unique Local Addresses (ULAs) [RFC4193] and IPv4 private addressing [RFC1918], which do not provide any security by 'magic'. In both cases, the edge router must apply strict filters to block those private addresses from entering and, just as importantly, leaving the network. This filtering can be done by the enterprise or by the ISP, but the cautious administrator will prefer to do it in the enterprise.

IPv6 addresses can be spoofed as easily as IPv4 addresses and there are packets with bogon IPv6 addresses (see [CYMRU]). Anti-bogon filtering must be done in the data and routing planes. It can be done by the enterprise or by the ISP, or both, but again the cautious administrator will prefer to do it in the enterprise.

#### 2.4.3. Specific Security Issues for IPv6

Even if IPv6 is similar to IPv4, there are some differences that create some IPv6-only vulnerabilities or issues. We give examples of such differences in this section.

Privacy extension addresses [RFC4941] are usually used to protect individual privacy by periodically changing the interface identifier part of the IPv6 address to avoid tracking a host by its otherwise always identical and unique MAC-based EUI-64. While this presents a

real advantage on the Internet, moderated by the fact that the prefix part remains the same, it complicates the task of following an audit trail when a security officer or network operator wants to trace back a log entry to a host in their network, because when the tracing is done the searched IPv6 address could have disappeared from the network. Therefore, the use of privacy extension addresses usually requires additional monitoring and logging of the binding of the IPv6 address to a data-link layer address (see also the monitoring section of [I-D.ietf-opsec-v6]). Some early enterprise deployments have taken the approach of using tools that harvest IP/MAC address mappings from switch and router devices to provide address accountability; this approach has been shown to work, though it can involve gathering significantly more address data than in equivalent IPv4 networks. An alternative is to try to prevent the use of privacy extension addresses by enforcing the use of DHCPv6, such that hosts only get addresses assigned by a DHCPv6 server. This can be done by configuring routers to set the M-bit in Router Advertisements, combined with all advertised prefixes being included without the A-bit set (to prevent the use of stateless auto-configuration). This technique of course requires that all hosts support stateful DHCPv6. It is important to note that not all operating systems exhibit the same behavior when processing RAs with the M-Bit set. The varying OS behavior is related to the lack of prescriptive definition around the A, M and O-bits within the ND protocol. [I-D.liu-bonica-dhcpv6-slaac-problem] provides a much more detailed analysis on the interaction of the M-Bit and DHCPv6.

Extension headers complicate the task of stateless packet filters such as ACLs. If ACLs are used to enforce a security policy, then the enterprise must verify whether its ACL (but also stateful firewalls) are able to process extension headers (this means understand them enough to parse them to find the upper layers payloads) and to block unwanted extension headers (e.g., to implement [RFC5095]). This topic is discussed further in [RFC7045].

Fragmentation is different in IPv6 because it is done only by source host and never during a forwarding operation. This means that ICMPv6 packet-too-big messages must be allowed to pass through the network and not be filtered [RFC4890]. Fragments can also be used to evade some security mechanisms such as RA-guard [RFC6105]. See also [RFC5722], and [RFC7113].

One of the biggest differences between IPv4 and IPv6 is the introduction of the Neighbor Discovery Protocol [RFC4861], which includes a variety of important IPv6 protocol functions, including those provided in IPv4 by ARP [RFC0826]. NDP runs over ICMPv6 (which as stated above means that security policies must allow some ICMPv6 messages to pass, as described in RFC 4890), but has the same lack of

security as, for example, ARP, in that there is no inherent message authentication. While Secure Neighbour Discovery (SeND) [RFC3971] and CGA [RFC3972] have been defined, they are not widely implemented). The threat model for Router Advertisements within the NDP suite is similar to that of DHCPv4 (and DHCPv6), in that a rogue host could be either a rogue router or a rogue DHCP server. An IPv4 network can be made more secure with the help of DHCPv4 snooping in edge switches, and likewise RA snooping can improve IPv6 network security (in IPv4-only networks as well). Thus enterprises using such techniques for IPv4 should use the equivalent techniques for IPv6, including RA-guard [RFC6105] and all work in progress from the SAVI WG, e.g. [RFC6959], which is similar to the protection given by dynamic ARP monitoring in IPv4. Other DoS vulnerabilities are related to NDP cache exhaustion, and mitigation techniques can be found in ([RFC6583]).

As stated previously, running a dual-stack network doubles the attack exposure as a malevolent person has now two attack vectors: IPv4 and IPv6. This simply means that all routers and hosts operating in a dual-stack environment with both protocol families enabled (even if by default) must have a congruent security policy for both protocol versions. For example, permit TCP ports 80 and 443 to all web servers and deny all other ports to the same servers must be implemented both for IPv4 and IPv6. It is thus important that the tools available to administrators readily support such behaviour.

## 2.5. Routing

An important design choice to be made is what IGP to use inside the network. A variety of IGPs (IS-IS, OSPFv3 and RIPng) support IPv6 today and picking one over the other is a design choice that will be dictated mostly by existing operational policies in an enterprise network. As mentioned earlier, it would be beneficial to maintain operational parity between IPv4 and IPv6 and therefore it might make sense to continue using the same protocol family that is being used for IPv4. For example, in a network using OSPFv2 for IPv4, it might make sense to use OSPFv3 for IPv6. It is important to note that although OSPFv3 is similar to OSPFv2, they are not the same. On the other hand, some organizations may chose to run different routing protocols for different IP versions. For example, one may chose to run OSPFv2 for IPv4 and IS-IS for IPv6. An important design question to consider here is whether to support one IGP or two different IGPs in the longer term. [I-D.ietf-v6ops-design-choices] presents advice on the design choices that arise when considering IGPs and discusses the advantages and disadvantages to different approaches in detail.

## 2.6. Address Plan

The most common problem encountered in IPv6 networking is in applying the same principles of conservation that are so important in IPv4. IPv6 addresses do not need to be assigned conservatively. In fact, a single larger allocation is considered more conservative than multiple non-contiguous small blocks, because a single block occupies only a single entry in a routing table. The advice in [RFC5375] is still sound, and is recommended to the reader. If considering ULAs, give careful thought to how well it is supported, especially in multiple address and multicast scenarios, and assess the strength of the requirement for ULA. [I-D.ietf-v6ops-ula-usage-recommendations] provides much more detailed analysis and recommendations on the usage of ULAs.

The enterprise administrator will want to evaluate whether the enterprise will request address space from a LIR (Local Internet Registry, such as an ISP), a RIR (Regional Internet Registry, such as AfriNIC, APNIC, ARIN, LACNIC, or RIPE-NCC) or a NIR (National Internet Registry, operated in some countries). The normal allocation is Provider Aggregatable (PA) address space from the enterprise's ISP, but use of PA space implies renumbering when changing provider. Instead, an enterprise may request Provider Independent (PI) space; this may involve an additional fee, but the enterprise may then be better able to be multihomed using that prefix, and will avoid a renumbering process when changing ISPs (though it should be noted that renumbering caused by outgrowing the space, merger, or other internal reason would still not be avoided with PI space).

The type of address selected (PI vs. PA) should be congruent with the routing needs of the enterprise. The selection of address type will determine if an operator will need to apply new routing techniques and may limit future flexibility. There is no right answer, but the needs of the external phase may affect what address type is selected.

Each network location or site will need a prefix assignment. Depending on the type of site/location, various prefix sizes may be used. In general, historical guidance suggests that each site should get at least a /48, as documented in RFC 5375 and [RFC6177]. In addition to allowing for simple planning, this can allow a site to use its prefix for local connectivity, should the need arise, and if the local ISP supports it.

When assigning addresses to end systems, the enterprise may use manually-configured addresses (common on servers) or SLAAC or DHCPv6 for client systems. Early IPv6 enterprise deployments have used SLAAC, both for its simplicity but also due to the time DHCPv6 has



taken to mature. However, DHCPv6 is now very mature, and thus workstations managed by an enterprise may use stateful DHCPv6 for addressing on corporate LAN segments. DHCPv6 allows for the additional configuration options often employed by enterprise administrators, and by using stateful DHCPv6, administrators correlating system logs know which system had which address at any given time. Such an accountability model is familiar from IPv4 management, though for DHCPv6 hosts are identified by DUID rather than MAC address. For equivalent accountability with SLAAC (and potentially privacy addresses), a monitoring system that harvests IP/MAC mappings from switch and router equipment could be used.

A common deployment consideration for any enterprise network is how to get host DNS records updated. Commonly, either the host will send DNS updates or the DHCP server will update records. If there is sufficient trust between the hosts and the DNS server, the hosts may update (and the enterprise may use SLAAC for addressing). Otherwise, the DHCPv6 server can be configured to update the DNS server. Note that an enterprise network with this more controlled environment will need to disable SLAAC on network segments and force end hosts to use DHCPv6 only.

In the data center or server room, assume a /64 per VLAN. This applies even if each individual system is on a separate VLAN. In a /48 assignment, typical for a site, there are then still 65,535 /64 blocks. Some administrators reserve a /64 but configure a small subnet, such as /112, /126, or /127, to prevent rogue devices from attaching and getting numbers; an alternative is to monitor traffic for surprising addresses or ND tables for new entries. Addresses are either configured manually on the server, or reserved on a DHCPv6 server, which may also synchronize forward and reverse DNS (though see [RFC6866] for considerations on static addressing). SLAAC is not recommended for servers, because of the need to synchronize RA timers with DNS TTLs so that the DNS entry expires at the same time as the address.

All user access networks should be a /64. Point-to-point links where Neighbor Discovery Protocol is not used may also utilize a /127 (see [RFC6164]).

Plan to aggregate at every layer of network hierarchy. There is no need for VLSM [RFC1817] in IPv6, and addressing plans based on conservation of addresses are short-sighted. Use of prefixes longer than /64 on network segments will break common IPv6 functions such as SLAAC[RFC4862]. Where multiple VLANs or other layer two domains converge, allow some room for expansion. Renumbering due to outgrowing the network plan is a nuisance, so allow room within it. Generally, plan to grow to about twice the current size that can be

accommodated; where rapid growth is planned, allow for twice that growth. Also, if DNS (or reverse DNS) authority may be delegated to others in the enterprise, assignments need to be on nibble boundaries (that is, on a multiple of 4 bits, such as /64, /60, /56, ..., /48, /44), to ensure that delegated zones align with assigned prefixes.

If using ULAs, it is important to note that AAAA and PTR records for ULA are not recommended to be installed in the global DNS. Similarly, reverse (address-to-name) queries for ULA must not be sent to name servers outside of the organization, due to the load that such queries would create for the authoritative name servers for the ip6.arpa zone. For more details please refer to section 4.4 of [RFC4193].

Enterprise networks more and more include virtual networks where a single physical node may host many virtualized addressable devices. It is imperative that the addressing plans assigned to these virtual networks and devices be consistent and non-overlapping with the addresses assigned to real networks and nodes. For example, a virtual network established within an isolated lab environment may at a later time become attached to the production enterprise network.

## 2.7. Tools Assessment

Enterprises will often have a number of operational tools and support systems which are used to provision, monitor, manage and diagnose the network and systems within their environment. These tools and systems will need to be assessed for compatibility with IPv6. The compatibility may be related to the addressing and connectivity of various devices as well as IPv6 awareness of the tools and processing logic.

The tools within the organization fall into two general categories, those which focus on managing the network, and those which are focused on managing systems and applications on the network. In either instance, the tools will run on platforms which may or may not be capable of operating in an IPv6 network. This lack in functionality may be related to Operating System version, or based on some hardware constraint. Those systems which are found to be incapable of utilizing an IPv6 connection, or which are dependent on an IPv4 stack, may need to be replaced or upgraded.

In addition to devices working on an IPv6 network natively, or via a transition tunnel, many tools and support systems may require additional software updates to be IPv6 aware, or even a hardware upgrade (usually for additional memory: IPv6 addresses are larger and for a while, IPv4 and IPv6 addresses will coexist in the tool). This awareness may include the ability to manage IPv6 elements and/or

applications in addition to the ability to store and utilize IPv6 addresses.

Considerations when assessing the tools and support systems may include the fact that IPv6 addresses are significantly larger than IPv4, requiring data stores to support the increased size. Such issues are among those discussed in [RFC5952]. Many organizations may also run dual-stack networks, therefore the tools need to not only support IPv6 operation, but may also need to support the monitoring, management and intersection with both IPv6 and IPv4 simultaneously. It is important to note that managing IPv6 is not just constrained to using large IPv6 addresses, but also that IPv6 interfaces and nodes are likely to use two or more addresses as part of normal operation. Updating management systems to deal with these additional nuances will likely consume time and considerable effort.

For networking systems, like node management systems, it is not always necessary to support local IPv6 addressing and connectivity. Operations such as SNMP MIB polling can occur over IPv4 transport while seeking responses related to IPv6 information. Where this may seem advantageous to some, it should be noted that without local IPv6 connectivity, the management system may not be able to perform all expected functions - such as reachability and service checks.

Organizations should be aware that changes to older IPv4-only SNMP MIB specifications have been made by the IETF related to legacy operation in [RFC2096] and [RFC2011]. Updated specifications are now available in [RFC4292] and [RFC4293] which modified the older MIB framework to be IP protocol agnostic, supporting both IPv4 and IPv6. Polling systems will need to be upgraded to support these updates as well as the end stations which are polled.

### 3. External Phase

The external phase for enterprise IPv6 adoption covers topics which deal with how an organization connects its infrastructure to the external world. These external connections may be toward the Internet at large, or to other networks. The external phase covers connectivity, security and monitoring of various elements and outward facing or accessible services.

#### 3.1. Connectivity

The enterprise will need to work with one or more Service Providers to gain connectivity to the Internet or transport service infrastructure such as a BGP/MPLS IP VPN as described in [RFC4364] and [RFC4659]. One significant factor that will guide how an organization may need to communicate with the outside world will

involve the use of PI (Provider Independent) and/or PA (Provider Aggregatable) IPv6 space.

Enterprises should be aware that depending on which address type they selected (PI vs. PA) in their planning phase, they may need to implement new routing functions and/or behaviours to support their connectivity to the ISP. In the case of PI, the upstream ISP may offer options to route the prefix (typically a /48) on the enterprise's behalf and update the relevant routing databases. Otherwise, the enterprise may need to perform this task on their own and use BGP to inject the prefix into the global BGP system.

Note that the rules set by the RIRs for an enterprise acquiring PI address space have changed over time. For example, in the European region the RIPE-NCC no longer requires an enterprise to be multihomed to be eligible for an IPv6 PI allocation. Requests can be made directly or via a LIR. It is possible that the rules may change again, and may vary between RIRs.

When seeking IPv6 connectivity to a Service Provider, Native IPv6 connectivity is preferred since it provides the most robust and efficient form of connectivity. If native IPv6 connectivity is not possible due to technical or business limitations, the enterprise may utilize readily available transition tunnel IPv6 connectivity. There are IPv6 transit providers which provide robust tunnelled IPv6 connectivity which can operate over IPv4 networks. It is important to understand the transition tunnel mechanism used, and to consider that it will have higher latency than native IPv4 or IPv6, and may have other problems, e.g. related to MTUs.

It is important to evaluate MTU considerations when adding IPv6 to an existing IPv4 network. It is generally desirable to have the IPv6 and IPv4 MTU congruent to simplify operations (so the two address families behave similarly, that is, as expected). If the enterprise uses transition tunnels inside or externally for IPv6 connectivity, then modification of the MTU on hosts/routers may be needed as mid-stream fragmentation is no longer supported in IPv6. It is preferred that pMTUD is used to optimize the MTU, so erroneous filtering of the related ICMPv6 message types should be monitored. Adjusting the MTU may be the only option if undesirable upstream ICMPv6 filtering cannot be removed.

### 3.2. Security

The most important part of security for external IPv6 deployment is filtering and monitoring. Filtering can be done by stateless ACLs or a stateful firewall. The security policies must be consistent for IPv4 and IPv6 (else the attacker will use the less protected protocol

stack), except that certain ICMPv6 messages must be allowed through and to the filtering device (see [RFC4890]):

- o Packet Too Big: essential to allow Path MTU discovery to work
- o Parameter Problem
- o Time Exceeded

In addition, Neighbor Discovery Protocol messages (including Neighbor Solicitation, Router Advertisements, etc.) are required for local hosts.

It could also be safer to block all fragments where the transport layer header is not in the first fragment to avoid attacks as described in [RFC5722]. Some filtering devices allow this filtering. Ingress filters and firewalls should follow [RFC5095] in handling routing extension header type 0, dropping the packet and sending ICMPv6 Parameter Problem, unless Segments Left = 0 (in which case, ignore the header).

If an Intrusion Prevention System (IPS) is used for IPv4 traffic, then an IPS should also be used for IPv6 traffic. In general, make sure IPv6 security is at least as good as IPv4. This also includes all email content protection (anti-spam, content filtering, data leakage prevention, etc.).

The edge router must also implement anti-spoofing techniques based on [RFC2827] (also known as BCP 38).

In order to protect the networking devices, it is advised to implement control plane policing as per [RFC6192].

The potential NDP cache exhaustion attack (see [RFC6583]) can be mitigated by two techniques:

- o Good NDP implementation with memory utilization limits as well as rate-limiters and prioritization of requests.
- o Or, as the external deployment usually involves just a couple of exposed statically configured IPv6 addresses (virtual addresses of web, email, and DNS servers), then it is straightforward to build an ingress ACL allowing traffic for those addresses and denying traffic to any other addresses. This actually prevents the attack as a packet for a random destination will be dropped and will never trigger a neighbor resolution.

### 3.3. Monitoring

Monitoring the use of the Internet connectivity should be done for IPv6 as it is done for IPv4. This includes the use of IP Flow Information eXport (IPFIX) [RFC7012] to report abnormal traffic patterns (such as port scanning, SYN-flooding, related IP source addresses) from monitoring tools and evaluating data read from SNMP MIBs [RFC4293] (some of which also enable the detection of abnormal bandwidth utilization) and syslogs (finding server and system errors). Where Netflow is used, version 9 is required for IPv6 support. Monitoring systems should be able to examine IPv6 traffic, use IPv6 for connectivity, record IPv6 address, and any log parsing tools and reporting need to support IPv6. Some of this data can be sensitive (including personally identifiable information) and care in securing it should be taken, with periodic purges. Integrity protection on logs and sources of log data is also important to detect unusual behavior (misconfigurations or attacks). Logs may be used in investigations, which depend on trustworthy data sources (tamper resistant).

In addition, monitoring of external services (such as web sites) should be made address-specific, so that people are notified when either the IPv4 or IPv6 version of a site fails.

### 3.4. Servers and Applications

The path to the servers accessed from the Internet usually involves security devices (firewall, IPS), server load balancing (SLB) and real physical servers. The latter stage is also multi-tiered for scalability and security between presentation and data storage. The ideal transition is to enable native dual-stack on all devices; but as part of the phased approach, operators have used the following techniques with success:

- o Use a network device to apply NAT64 and basically translate an inbound TCP connection (or any other transport protocol) over IPv6 into a TCP connection over IPv4. This is the easiest to deploy as the path is mostly unchanged but it hides all IPv6 remote users behind a single IPv4 address which leads to several audit trail and security issues (see [RFC6302]).
- o Use the server load balancer which acts as an application proxy to do this translation. Compared to the NAT64, it has the potential benefit of going through the security devices as native IPv6 (so more audit and trace abilities) and is also able to insert a HTTP X-Forward-For header which contains the remote IPv6 address. The latter feature allows for logging, and rate-limiting on the real

servers based on the IPV6 address even if those servers run only IPv4.

In either of these cases, care should be taken to secure logs for privacy reasons, and to periodically purge them.

### 3.5. Network Prefix Translation for IPv6

Network Prefix Translation for IPv6, or NPTv6 as described in [RFC6296] provides a framework to utilize prefix ranges within the internal network which are separate (address-independent) from the assigned prefix from the upstream provider or registry. As mentioned above, while NPTv6 has potential use-cases in IPv6 networks, the implications of its deployment need to be fully understood, particularly where any applications might embed IPv6 addresses in their payloads.

Use of NPTv6 can be chosen independently from how addresses are assigned and routed within the internal network, how prefixes are routed towards the Internet, or whether PA or PI addresses are used.

## 4. Internal Phase

This phase deals with the delivery of IPv6 to the internal user-facing side of the IT infrastructure, which comprises various components such as network devices (routers, switches, etc.), end user devices and peripherals (workstations, printers, etc.), and internal corporate systems.

An important design paradigm to consider during this phase is "dual-stack when you can, tunnel when you must". Dual-stacking allows a more robust, production-quality IPv6 network than is typically facilitated by internal use of transition tunnels that are harder to troubleshoot and support, and that may introduce scalability and performance issues. Tunnels may of course still be used in production networks, but their use needs to be carefully considered, e.g. where the transition tunnel may be run through a security or filtering device. Tunnels do also provide a means to experiment with IPv6 and gain some operational experience with the protocol. [RFC4213] describes various transition mechanisms in more detail. [RFC6964] suggests operational guidance when using ISATAP tunnels [RFC5214], though we would recommend use of dual-stack wherever possible.

#### 4.1. Security

IPv6 must be deployed in a secure way. This means that all existing IPv4 security policies must be extended to support IPv6; IPv6 security policies will be the IPv6 equivalent of the existing IPv4 ones (taking into account the difference for ICMPv6 [RFC4890]). As in IPv4, security policies for IPv6 will be enforced by firewalls, ACL, IPS, VPN, and so on.

Privacy extension addresses [RFC4941] raise a challenge for an audit trail as explained in section Section 2.4.3. The enterprise may choose to attempt to enforce use of DHCPv6, or deploy monitoring tools that harvest accountability data from switches and routers (thus making the assumption that devices may use any addresses inside the network).

One major issue is threats against Neighbor Discovery. This means, for example, that the internal network at the access layer (where hosts connect to the network over wired or wireless) should implement RA-guard [RFC6105] and the techniques being specified by SAVI WG [RFC6959]; see also Section 2.4.3 for more information.

#### 4.2. Network Infrastructure

The typical enterprise network infrastructure comprises a combination of the following network elements - wired access switches, wireless access points, and routers (although it is fairly common to find hardware that collapses switching and routing functionality into a single device). Basic wired access switches and access points operate only at the physical and link layers, and don't really have any special IPv6 considerations other than being able to support IPv6 addresses themselves for management purposes. In many instances, these devices possess a lot more intelligence than simply switching packets. For example, some of these devices help assist with link layer security by incorporating features such as ARP inspection and DHCP Snooping, or they may help limit where multicast floods by using IGMP (or, in the case of IPv6, MLD) snooping.

Another important consideration in enterprise networks is first hop router redundancy. This directly ties into network reachability from an end host's point of view. IPv6 Neighbor Discovery (ND), [RFC4861], provides a node with the capability to maintain a list of available routers on the link, in order to be able to switch to a backup path should the primary be unreachable. By default, ND will detect a router failure in 38 seconds and cycle onto the next default router listed in its cache. While this feature provides a basic level of first hop router redundancy, most enterprise IPv4 networks are designed to fail over much faster. Although this delay can be



improved by adjusting the default timers, care must be taken to protect against transient failures and to account for increased traffic on the link. Another option to provide robust first hop redundancy is to use the Virtual Router Redundancy Protocol for IPv6 (VRRPv3), [RFC5798]. This protocol provides a much faster switchover to an alternate default router than default ND parameters. Using VRRPv3, a backup router can take over for a failed default router in around three seconds (using VRRPv3 default parameters). This is done without any interaction with the hosts and a minimum amount of VRRP traffic.

Last but not the least, one of the most important design choices to make while deploying IPv6 on the internal network is whether to use Stateless Automatic Address Configuration (SLAAC), [RFC4862], or Dynamic Host Configuration Protocol for IPv6 (DHCPv6), [RFC3315], or a combination thereof. Each option has advantages and disadvantages, and the choice will ultimately depend on the operational policies that guide each enterprise's network design. For example, if an enterprise is looking for ease of use, rapid deployment, and less administrative overhead, then SLAAC makes more sense for workstations. Manual or DHCPv6 assignments are still needed for servers, as described in the External Phase and Address Plan sections of this document. However, if the operational policies call for precise control over IP address assignment for auditing then DHCPv6 may be preferred. DHCPv6 also allows you to tie into DNS systems for host entry updates and gives you the ability to send other options and information to clients. It is worth noting that in general operation RAs are still needed in DHCPv6 networks, as there is no DHCPv6 Default Gateway option. Similarly, DHCPv6 is needed in RA networks for other configuration information, e.g. NTP servers or, in the absence of support for DNS resolvers in RAs [RFC6106], DNS resolver information.

#### 4.3. End user devices

Most operating systems (OSes) that are loaded on workstations and laptops in a typical enterprise support IPv6 today. However, there are various out-of-the-box nuances that one should be mindful about. For example, the default behavior of OSes vary; some may have IPv6 turned off by default, some may only have certain features such as privacy extensions to IPv6 addresses (RFC 4941) turned off while others have IPv6 fully enabled. Further, even when IPv6 is enabled, the choice of which address is used may be subject to Source Address Selection (RFC 6724) and Happy Eyeballs (RFC 6555). Therefore, it is advised that enterprises investigate the default behavior of their installed OS base and account for it during the Inventory phases of their IPv6 preparations. Furthermore, some OSes may have some transition tunneling mechanisms turned on by default and in such

cases it is recommended to administratively shut down such interfaces unless required.

It is important to note that it is recommended that IPv6 be deployed at the network and system infrastructure level before it is rolled out to end user devices; ensure IPv6 is running and routed on the wire, and secure and correctly monitored, before exposing IPv6 to end users.

Smartphones and tablets are significant IPv6-capable platforms, depending on the support of the carrier's data network.

IPv6 support for peripherals varies. Much like servers, printers are generally configured with a static address (or DHCP reservation) so clients can discover them reliably.

#### 4.4. Corporate Systems

No IPv6 deployment will be successful without ensuring that all the corporate systems that an enterprise uses as part of its IT infrastructure support IPv6. Examples of such systems include, but are not limited to, email, video conferencing, telephony (VoIP), DNS, RADIUS, etc. All these systems must have their own detailed IPv6 rollout plan in conjunction with the network IPv6 rollout. It is important to note that DNS is one of the main anchors in an enterprise deployment, since most end hosts decide whether or not to use IPv6 depending on the presence of IPv6 AAAA records in a reply to a DNS query. It is recommended that system administrators selectively turn on AAAA records for various systems as and when they are IPv6 enabled; care must be taken though to ensure all services running on that host name are IPv6-enabled before adding the AAAA record. Care with web proxies is advised; a mismatch in the level of IPv6 support between the client, proxy, and server can cause communication problems. All monitoring and reporting tools across the enterprise will need to be modified to support IPv6.

#### 5. IPv6-only

Early IPv6 enterprise deployments have generally taken a dual-stack approach to enabling IPv6, i.e. the existing IPv4 services have not been turned off. Although IPv4 and IPv6 networks will coexist for a long time, the long term enterprise network roadmap should include steps to simplify engineering and operations by deprecating IPv4 from the dual-stack network. In some extreme cases, deploying dual-stack networks may not even be a viable option for very large enterprises due to the RFC 1918 address space not being large enough to support the network's growth. In such cases, deploying IPv6-only networks might be the only choice available to sustain network growth. In

other cases, there may be elements of an otherwise dual-stack network that may be run IPv6-only.

If nodes in the network don't need to talk to an IPv4-only node, then deploying IPv6-only networks should be straightforward. However, most nodes will need to communicate with some IPv4-only nodes; an IPv6-only node may therefore require a translation mechanism. As [RFC6144] points out, it is important to look at address translation as a transition strategy towards running an IPv6-only network.

There are various stateless and stateful IPv4/IPv6 translation methods available today that help IPv6 to IPv4 communication. RFC 6144 provides a framework for IPv4/IPv6 translation and describes in detail various scenarios in which such translation mechanisms could be used. [RFC6145] describes stateless address translation. In this mode, a specific IPv6 address range will represent IPv4 systems (IPv4-converted addresses), and the IPv6 systems have addresses (IPv4-translatable addresses) that can be algorithmically mapped to a subset of the service provider's IPv4 addresses. [RFC6146], NAT64, describes stateful address translation. As the name suggests, the translation state is maintained between IPv4 address/port pairs and IPv6 address/port pairs, enabling IPv6 systems to open sessions with IPv4 systems. [RFC6147], DNS64, describes a mechanism for synthesizing AAAA resource records (RRs) from A RRs. Together, RFCs 6146 and RFC 6147 provide a viable method for an IPv6-only client to initiate communications to an IPv4-only server. As described in the assumptions section, the administrator will usually want most traffic or flows to be native, and only translate as needed.

The address translation mechanisms for the stateless and stateful translations are defined in [RFC6052]. It is important to note that both of these mechanisms have limitations as to which protocols they support. For example, RFC 6146 only defines how stateful NAT64 translates unicast packets carrying TCP, UDP, and ICMP traffic only. The classic problems of IPv4 NAT also apply, e.g. handling IP literals in application payloads. The ultimate choice of which translation mechanism to choose will be dictated mostly by existing operational policies pertaining to application support, logging requirements, etc.

There is additional work being done in the area of address translation to enhance and/or optimize current mechanisms. For example, [I-D.xli-behave-divi] describes limitations with the current stateless translation, such as IPv4 address sharing and application layer gateway (ALG) problems, and presents the concept and implementation of dual-stateless IPv4/IPv6 translation (dIVI) to address those issues.

It is worth noting that for IPv6-only access networks that use technologies such as NAT64, the more content providers (and enterprises) that make their content available over IPv6, the less the requirement to apply NAT64 to traffic leaving the access network. This particular point is important for enterprises which may start their IPv6 deployment well into the global IPv6 transition. As time progresses, and given the current growth in availability of IPv6 content, IPv6-only operation using NAT64 to manage some flows will become less expensive to run versus the traditional NAT44 deployments since only IPv6 to IPv4 flows need translation. [RFC6883] provides guidance and suggestions for Internet Content Providers and Application Service Providers in this context.

Enterprises should also be aware that networks may be subject to future convergence with other networks (i.e. mergers, acquisitions, etc). An enterprise considering IPv6-only operation may need to be aware that additional transition technologies and/or connectivity strategies may be required depending on the level of IPv6 readiness and deployment in the merging networking.

## 6. Considerations For Specific Enterprises

### 6.1. Content Delivery Networks

Some guidance for Internet Content and Application Service Providers can be found in [RFC6883], which includes a dedicated section on Content Delivery Networks (CDNs). An enterprise that relies on a CDN to deliver a 'better' e-commerce experience needs to ensure that their CDN provider also supports IPv4/IPv6 traffic selection so that they can ensure 'best' access to the content. A CDN could enable external IPv6 content delivery even if the enterprise provides that content over IPv4.

### 6.2. Data Center Virtualization

IPv6 Data Center considerations are described in [I-D.ietf-v6ops-dc-ipv6].

### 6.3. University Campus Networks

A number of campus networks around the world have made some initial IPv6 deployment. This has been encouraged by their National Research and Education Network (NREN) backbones having made IPv6 available natively since the early 2000's. Universities are a natural place for IPv6 deployment to be considered at an early stage, perhaps compared to other enterprises, as they are involved by their very nature in research and education.

Campus networks can deploy IPv6 at their own pace; there is no need to deploy IPv6 across the entire enterprise from day one, rather specific projects can be identified for an initial deployment, that are both deep enough to give the university experience, but small enough to be a realistic first step. There are generally three areas in which such deployments are currently made.

In particular those initial areas commonly approached are:

- o External-facing services. Typically the campus web presence and commonly also external-facing DNS and MX services. This ensures early IPv6-only adopters elsewhere can access the campus services as simply and as robustly as possible.
- o Computer science department. This is where IPv6-related research and/or teaching is most likely to occur, and where many of the next generation of network engineers are studying, so enabling some or all of the campus computer science department network is a sensible first step.
- o The eduroam wireless network. Eduroam [I-D.wierenga-ietf-eduroam] is the de facto wireless roaming system for academic networks, and uses 802.1X-based authentication, which is agnostic to the IP version used (unlike web-redirection gateway systems). Making a campus' eduroam network dual-stack is a very viable early step.

The general IPv6 deployment model in a campus enterprise will still follow the general principles described in this document. While the above early stage projects are commonly followed, these still require the campus to acquire IPv6 connectivity and address space from their NREN (or other provider in some parts of the world), and to enable IPv6 on the wire on at least part of the core of the campus network. This implies a requirement to have an initial address plan, and to ensure appropriate monitoring and security measures are in place, as described elsewhere in this document.

Campuses which have deployed to date do not use ULAs, nor do they use NPTv6. In general, campuses have very stable PA-based address allocations from their NRENS (or their equivalent). However, campus enterprises may consider applying for IPv6 PI; some have already done so. The discussions earlier in this text about PA vs. PI still apply.

Finally, campuses may be more likely than many other enterprises to run multicast applications, such as IP TV or live lecture or seminar streaming, so may wish to consider support for specific IPv6 multicast functionality, e.g. Embedded-RP [RFC3956] in routers and MLDv1 and MLDv2 snooping in switches.

## 7. Security Considerations

This document has multiple security sections detailing how to securely deploy an IPv6 network within an enterprise network.

## 8. Acknowledgements

The authors would like to thank Robert Sparks, Steve Hanna, Tom Taylor, Brian Haberman, Stephen Farrell, Chris Grundemann, Ray Hunter, Kathleen Moriarty, Benoit Claise, Brian Carpenter, Tina Tsou, Christian Jaquet, and Fred Templin for their substantial comments and contributions.

## 9. IANA Considerations

There are no IANA considerations or implications that arise from this document.

## 10. Informative References

- [RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.
- [RFC1687] Fleischman, E., "A Large Corporate User's View of IPng", RFC 1687, August 1994.
- [RFC1817] Rekhter, Y., "CIDR and Classful Routing", RFC 1817, August 1995.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC2011] McCloghrie, K., "SNMPv2 Management Information Base for the Internet Protocol using SMIV2", RFC 2011, November 1996.
- [RFC2096] Baker, F., "IP Forwarding Table MIB", RFC 2096, January 1997.
- [RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

- [RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.
- [RFC3956] Savola, P. and B. Haberman, "Embedding the Rendezvous Point (RP) Address in an IPv6 Multicast Address", RFC 3956, November 2004.
- [RFC3971] Arkko, J., Kempf, J., Zill, B., and P. Nikander, "SEcure Neighbor Discovery (SEND)", RFC 3971, March 2005.
- [RFC3972] Aura, T., "Cryptographically Generated Addresses (CGA)", RFC 3972, March 2005.
- [RFC4038] Shin, M-K., Hong, Y-G., Hagino, J., Savola, P., and E. Castro, "Application Aspects of IPv6 Transition", RFC 4038, March 2005.
- [RFC4057] Bound, J., "IPv6 Enterprise Network Scenarios", RFC 4057, June 2005.
- [RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.
- [RFC4293] Routhier, S., "Management Information Base for the Internet Protocol (IP)", RFC 4293, April 2006.
- [RFC4292] Haberman, B., "IP Forwarding Table MIB", RFC 4292, April 2006.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, September 2006.
- [RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC4890] Davies, E. and J. Mohacsi, "Recommendations for Filtering ICMPv6 Messages in Firewalls", RFC 4890, May 2007.
- [RFC4941] Narten, T., Draves, R., and S. Krishnan, "Privacy Extensions for Stateless Address Autoconfiguration in IPv6", RFC 4941, September 2007.
- [RFC5095] Abley, J., Savola, P., and G. Neville-Neil, "Deprecation of Type 0 Routing Headers in IPv6", RFC 5095, December 2007.
- [RFC7012] Claise, B. and B. Trammell, "Information Model for IP Flow Information Export (IPFIX)", RFC 7012, September 2013.
- [RFC5157] Chown, T., "IPv6 Implications for Network Scanning", RFC 5157, March 2008.
- [RFC5211] Curran, J., "An Internet Transition Plan", RFC 5211, July 2008.
- [RFC5214] Templin, F., Gleeson, T., and D. Thaler, "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)", RFC 5214, March 2008.
- [RFC5375] Van de Velde, G., Popoviciu, C., Chown, T., Bonness, O., and C. Hahn, "IPv6 Unicast Address Assignment Considerations", RFC 5375, December 2008.
- [RFC5722] Krishnan, S., "Handling of Overlapping IPv6 Fragments", RFC 5722, December 2009.
- [RFC5798] Nadas, S., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, March 2010.
- [RFC5952] Kawamura, S. and M. Kawashima, "A Recommendation for IPv6 Address Text Representation", RFC 5952, August 2010.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6104] Chown, T. and S. Venaas, "Rogue IPv6 Router Advertisement Problem Statement", RFC 6104, February 2011.



- [RFC6105] Levy-Abegnoli, E., Van de Velde, G., Popoviciu, C., and J. Mohacsi, "IPv6 Router Advertisement Guard", RFC 6105, February 2011.
- [RFC6106] Jeong, J., Park, S., Beloeil, L., and S. Madanapalli, "IPv6 Router Advertisement Options for DNS Configuration", RFC 6106, November 2010.
- [RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, April 2011.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.
- [RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", RFC 6146, April 2011.
- [RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 2011.
- [RFC6177] Narten, T., Huston, G., and L. Roberts, "IPv6 Address Assignment to End Sites", BCP 157, RFC 6177, March 2011.
- [RFC6164] Kohno, M., Nitzan, B., Bush, R., Matsuzaki, Y., Colitti, L., and T. Narten, "Using 127-Bit IPv6 Prefixes on Inter-Router Links", RFC 6164, April 2011.
- [RFC6192] Dugal, D., Pignataro, C., and R. Dunn, "Protecting the Router Control Plane", RFC 6192, March 2011.
- [RFC6296] Wasserman, M. and F. Baker, "IPv6-to-IPv6 Network Prefix Translation", RFC 6296, June 2011.
- [RFC6302] Durand, A., Gashinsky, I., Lee, D., and S. Sheppard, "Logging Recommendations for Internet-Facing Servers", BCP 162, RFC 6302, June 2011.
- [RFC6434] Jankiewicz, E., Loughney, J., and T. Narten, "IPv6 Node Requirements", RFC 6434, December 2011.
- [RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.
- [RFC6583] Gashinsky, I., Jaeggli, J., and W. Kumari, "Operational Neighbor Discovery Problems", RFC 6583, March 2012.

- [RFC6724] Thaler, D., Draves, R., Matsumoto, A., and T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, September 2012.
- [RFC6866] Carpenter, B. and S. Jiang, "Problem Statement for Renumbering IPv6 Hosts with Static Addresses in Enterprise Networks", RFC 6866, February 2013.
- [RFC6879] Jiang, S., Liu, B., and B. Carpenter, "IPv6 Enterprise Network Renumbering Scenarios, Considerations, and Methods", RFC 6879, February 2013.
- [RFC6883] Carpenter, B. and S. Jiang, "IPv6 Guidance for Internet Content Providers and Application Service Providers", RFC 6883, March 2013.
- [RFC6959] McPherson, D., Baker, F., and J. Halpern, "Source Address Validation Improvement (SAVI) Threat Scope", RFC 6959, May 2013.
- [RFC6964] Templin, F., "Operational Guidance for IPv6 Deployment in IPv4 Sites Using the Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)", RFC 6964, May 2013.
- [RFC7011] Claise, B., Trammell, B., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, September 2013.
- [RFC7113] Gont, F., "Implementation Advice for IPv6 Router Advertisement Guard (RA-Guard)", RFC 7113, February 2014.
- [RFC7123] Gont, F. and W. Liu, "Security Implications of IPv6 on IPv4 Networks", RFC 7123, February 2014.
- [RFC7045] Carpenter, B. and S. Jiang, "Transmission and Processing of IPv6 Extension Headers", RFC7045, December 2013, <<http://tools.ietf.org/html/rfc7045>>.
- [I-D.xli-behave-divi]  
    Bao, C., Li, X., Zhai, Y., and W. Shang, "dIVI: Dual-Stateless IPv4/IPv6 Translation", draft-xli-behave-divi-06 (work in progress), January 2014.
- [I-D.wierenga-ietf-eduroam]  
    Wierenga, K., Winter, S., and T. Wolniewicz, "The eduroam architecture for network roaming", draft-wierenga-ietf-eduroam-03 (work in progress), February 2014.

- [I-D.ietf-v6ops-dc-ipv6]  
Lopez, D., Chen, Z., Tsou, T., Zhou, C., and A. Servin,  
"IPv6 Operational Guidelines for Datacenters", draft-ietf-  
v6ops-dc-ipv6-01 (work in progress), February 2014.
- [I-D.ietf-v6ops-design-choices]  
Matthews, P. and V. Kuarsingh, "Design Choices for IPv6  
Networks", draft-ietf-v6ops-design-choices-01 (work in  
progress), March 2014.
- [I-D.ietf-opsec-v6]  
Chittimaneni, K., Kaeo, M., and E. Vyncke, "Operational  
Security Considerations for IPv6 Networks", draft-ietf-  
opsec-v6-04 (work in progress), October 2013.
- [I-D.ietf-opsec-ipv6-host-scanning]  
Gont, F. and T. Chown, "Network Reconnaissance in IPv6  
Networks", draft-ietf-opsec-ipv6-host-scanning-04 (work in  
progress), June 2014.
- [I-D.liu-bonica-dhcpv6-slaac-problem]  
Liu, B. and R. Bonica, "DHCPv6/SLAAC Address Configuration  
Interaction Problem Statement", draft-liu-bonica-dhcpv6-  
slaac-problem-02 (work in progress), September 2013.
- [I-D.ietf-v6ops-ula-usage-recommendations]  
Liu, B. and S. Jiang, "Considerations of Using Unique  
Local Addresses", draft-ietf-v6ops-ula-usage-  
recommendations-03 (work in progress), July 2014.
- [I-D.ietf-opsec-vpn-leakages]  
Gont, F., "Layer-3 Virtual Private Network (VPN) tunnel  
traffic leakages in dual- stack hosts/networks", draft-  
ietf-opsec-vpn-leakages-06 (work in progress), April 2014.
- [SMURF] "CERT Advisory CA-1998-01 Smurf IP Denial-of-Service  
Attacks",  
<<http://www.cert.org/advisories/CA-1998-01.html>>.
- [CYMRU] "THE BOGON REFERENCE",  
<<http://www.team-cymru.org/Services/Bogons/>>.

Authors' Addresses

Kiran K. Chittimaneni  
Dropbox Inc.  
1600 Amphitheater Pkwy  
Mountain View, California CA 94043  
USA

Email: [kk@dropbox.com](mailto:kk@dropbox.com)

Tim Chown  
University of Southampton  
Highfield  
Southampton, Hampshire SO17 1BJ  
United Kingdom

Email: [tjc@ecs.soton.ac.uk](mailto:tjc@ecs.soton.ac.uk)

Lee Howard  
Time Warner Cable  
13820 Sunrise Valley Drive  
Herndon, VA 20171  
US

Phone: +1 703 345 3513  
Email: [lee.howard@twcable.com](mailto:lee.howard@twcable.com)

Victor Kuarsingh  
Dyn Inc  
150 Dow Street  
Manchester, NH  
US

Email: [victor@jvknet.com](mailto:victor@jvknet.com)

Yanick Pouffary  
Hewlett Packard  
950 Route Des Colles  
Sophia-Antipolis 06901  
France

Email: [Yanick.Pouffary@hp.com](mailto:Yanick.Pouffary@hp.com)

Eric Vyncke  
Cisco Systems  
De Kleetlaan 6a  
Diegem 1831  
Belgium

Phone: +32 2 778 4677  
Email: [evyncke@cisco.com](mailto:evyncke@cisco.com)

Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: September 11, 2014

G. Chen  
Z. Cao  
China Mobile  
C. Xie  
China Telecom  
D. Binet  
France Telecom-Orange  
March 10, 2014

NAT64 Deployment Options and Experience  
draft-ietf-v6ops-nat64-experience-10

Abstract

This document summarizes NAT64 function deployment scenarios and operational experience. Both NAT64 Carrier Grade NAT (NAT64-CGN) and NAT64 server Front End (NAT64-FE) are considered in this document.

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 11, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Terminology . . . . .	3
3. NAT64 Networking Experience . . . . .	4
3.1. NAT64-CGN Consideration . . . . .	4
3.1.1. NAT64-CGN Usages . . . . .	4
3.1.2. DNS64 Deployment . . . . .	4
3.1.3. NAT64 Placement . . . . .	5
3.1.4. Co-existence of NAT64 and NAT44 . . . . .	5
3.2. NAT64-FE Consideration . . . . .	6
4. High Availability . . . . .	7
4.1. Redundancy Design . . . . .	7
4.2. Load Balancing . . . . .	9
5. Source Address Transparency . . . . .	9
5.1. Traceability . . . . .	9
5.2. Geo-location . . . . .	10
6. Quality of Experience . . . . .	11
6.1. Service Reachability . . . . .	11
6.2. Resource Reservation . . . . .	12
7. MTU Considerations . . . . .	13
8. ULA Usages . . . . .	14
9. Security Considerations . . . . .	15
10. IANA Considerations . . . . .	15
11. Acknowledgements . . . . .	15
12. Additional Author List . . . . .	16
13. References . . . . .	16
13.1. Normative References . . . . .	16
13.2. Informative References . . . . .	18
Appendix A. Testing Results of Application Behavior . . . . .	20
Authors' Addresses . . . . .	21

## 1. Introduction

IPv6 is the only sustainable solution for numbering nodes on Internet due to the IPv4 depletion. Network operators have to deploy IPv6-only networks in order to meet the needs of the expanding internet without available IPv4 addresses.

Single-stack IPv6 network deployment can simplify networks provisioning, some justification was provided in 464xlat [RFC6877]. IPv6-only connectivity confers some benefits to mobile operators as an example. In the mobile context, IPv6-only usage enables the use of a single IPv6 Packet Data Protocol(PDP) context or Evolved Packet System (EPS) bearer on Long Term Evolution (LTE) networks. This

eliminates significant network costs caused by employing two PDP contexts in some cases, and the need for IPv4 addresses to be assigned to customers. In broadband networks overall, it can allow for the scaling of edge-network growth to be decoupled from IPv4 numbering limitations.

In transition scenarios, some existing networks are likely to be IPv4-only for quite a long time. IPv6 networks and hosts IPv6-only hosts will need to coexist with IPv4 numbered resources. Widespread dual-stack deployments have not materialized at the anticipated rate over the last 10 years, one possible conclusion being that legacy networks will not make the jump quickly. The Internet will include nodes that are dual-stack, nodes that remain IPv4-only, and nodes that can be deployed as IPv6-only nodes. A translation mechanism based on a NAT64[RFC6146] [RFC6145]function is likely to be a key element of Internet connectivity for IPv6-IPv4 interoperability.

[RFC6036] reports at least 30% of operators plan to run some kind of translator (presumably NAT64/DNS64). Advice on NAT64 deployment and operations are therefore of some importance. [RFC6586] documents the implications for IPv6 only networks. This document intends to be specific to NAT64 network planning.

## 2. Terminology

Regarding IPv4/IPv6 translation, [RFC6144] has described a framework for enabling networks to make interworking possible between IPv4 and IPv6 networks. This document has further categorized different NAT64 functions, locations and use-cases. The principle distinction of location is whether the NAT64 is located in a Carrier Grade NAT or server Front End. The terms of NAT-CGN/FE are understood to be a topological distinction indicating different features employed in a NAT64 deployment.

**NAT64 Carrier Grade NAT (NAT64-CGN):** A NAT64-CGN is placed in an ISP network. IPv6 enabled subscribers leverage the NAT64-CGN to access existing IPv4 internet services. The ISP as an administrative entity takes full control of the IPv6 side, but has limited or no control on the IPv4 internet side. NAT64-CGN deployments may have to consider the IPv4 Internet environment and services, and make appropriate configuration choices accordingly.

**NAT64 server Front End (NAT64-FE):** A NAT64-FE is generally a device with NAT64 functionality in a content provider or data center network. It could be for example a traffic load balancer or a firewall. The operator of the NAT64-FE has full control over the IPv4 network within the data center, but only limited influence or control over the external Internet IPv6 network.



### 3. NAT64 Networking Experience

#### 3.1. NAT64-CGN Consideration

##### 3.1.1. NAT64-CGN Usages

Fixed network operators and mobile operators may locate NAT64 translators in access networks or in mobile core networks. It can be built into various devices, including routers, gateways or firewalls in order to connect IPv6 users to the IPv4 Internet. With regard to the numbers of users and the shortage of public IPv4 addresses, stateful NAT64[RFC6146] is more suited to maximize sharing of public IPv4 addresses. The usage of stateless NAT64 can provide better transparency features [I-D.ietf-softwire-stateless-4v6-motivation], but has to be coordinated with A+P[RFC6346] processes as specified in [I-D.ietf-softwire-map-t] in order to address an IPv4 address shortage.

##### 3.1.2. DNS64 Deployment

DNS64[RFC6147] is recommended for use in combination with stateful NAT64, and will likely be an essential part of an IPv6 single-stack network that couples to the IPv4 Internet. 464xlat[RFC6877] can enable access of IPv4 only applications or applications that call IPv4 literal addresses. Using DNS64 will help 464xlat to automatically discover NAT64 prefix through [RFC7050]. Berkeley Internet Name Daemon (BIND) software supports the function. It's important to note that DNS64 generates the synthetic AAAA reply when services only provide A records. Operators should not expect to access IPv4 parts of a dual-stack server using NAT64/DNS64. The traffic is forwarded on IPv6 paths if dual-stack servers are targeted. IPv6 traffic may be routed around rather than going through NAT64. Only the traffic going to IPv4-only service would traverse the NAT64 translator. In some sense, it encourages IPv6 usage and limits NAT translation compared to employing NAT44, where all traffic flows have to be translated. In some cases, NAT64-CGNs may serve double roles, i.e. as a translator and IPv6 forwarder. In mobile networks, NAT64 may be deployed as the default gateway serving all the IPv6 traffic. The traffic heading to a dual-stack server is only forwarded on the NAT64. Therefore, both IPv6 and IPv4 are suggested to be configured on the Internet faced interfaces of NAT64. We tested on Top100 websites (referring to [Alexa] statistics). 43% of websites are connected and forwarded on the NAT64 since those websites have both AAAA and A records. With expansion of IPv6 support, the translation process on NAT64 will likely become less-important over time. It should be noted the DNS64-DNSSEC Interaction[RFC6147] may impact validation of Resource Records retrieved from the the DNS64 process. In particular, DNSSEC

validation will fail when DNS64 synthesizes AAAA records where there is a DNS query with the "DNSSEC OK" (DO) bit set and the "Checking Disabled" (CD) bit set received.

### 3.1.3. NAT64 Placement

All connections to IPv4 services from IPv6-only clients must traverse the NAT64-CGN. It can be advantageous from the vantage-point of troubleshooting and traffic engineering to carry the IPv6 traffic natively for as long as possible within an access network and translate packets only at or near the network egress. NAT64 may be a feature of the Autonomous System (AS) border in fixed networks. It may be deployed in an IP node beyond the Gateway GPRS Support Node (GGSN) or Public Data Network- Gateway (PDN-GW) in mobile networks or directly as part of the gateway itself in some situations. This allows consistent attribution and traceability within the service provider network. It has been observed that the process of correlating log information is problematic from multiple-vendor's equipment due to inconsistent formats of log records. Placing NAT64 in a centralized location may reduce diversity of log format and simplify the network provisioning. Moreover, since NAT64 is only targeted at serving traffic flows from IPv6 to IPv4-only services, the user traffic volume should not be as high as in a NAT44 scenario, and therefore, the gateway's capacity in such location may be less of a concern or a hurdle to deployment. On the other-hand, placement in a centralized fashion would require more strict high availability (HA) design. It would also make geo-location based on IPv4 addresses rather inaccurate as is currently the case for NAT44 CGN already deployed in ISP networks. More considerations or workarounds on HA and traceability could be found at Section 4 and Section 5.

### 3.1.4. Co-existence of NAT64 and NAT44

NAT64 will likely co-exist with NAT44 in a dual-stack network where IPv4 private addresses are allocated to customers. The coexistence has already been observed in mobile networks, in which dual stack mobile phones normally initiate some dual-stack PDN/PDP Type[RFC6459] to query both IPv4/IPv6 address and IPv4 allocated addresses are very often private ones. [RFC6724] always prioritizes IPv6 connections regardless of whether the end-to-end path is native IPv6 or IPv6 translated to IPv4 via NAT64/DNS64. Conversely, Happy Eyeballs[RFC6555] will direct some IP flows across IPv4 paths. The selection of IPv4/IPv6 paths may depend on particular implementation choices or settings on a host-by-host basis, and may differ from an operator's deterministic scheme. Our tests verified that hosts may find themselves switching between IPv4 and IPv6 paths as they access identical service, but at different times [I-D.kaliwoda-sunset4-dual-ipv6-coexist]. Since the topology on each

path is potentially different, it may cause unstable user experience and some degradation of Quality of Experience (QoE) when falling back to the other protocol. It's also difficult for operators to find a solution to make a stable network with optimal resource utilization. In general, it's desirable to figure out the solution that will introduce IPv6/IPv4 translation service to IPv6-only hosts connecting to IPv4 servers while making sure dual-stack hosts to have at least one address family accessible via native service if possible. With the end-to-end native IPv6 environment available, hosts should be upgraded aggressively to migrate in favor of IPv6-only. There are ongoing efforts to detect host connectivity and propose a new DHCPv6 option[I-D.wing-dhc-dns-reconfigure] to convey appropriate configuration information to the hosts.

### 3.2. NAT64-FE Consideration

Some Internet Content Providers (ICPs) may locate NAT64 in front of an Internet Data Center (IDC), for example co-located with a load-balancing function. Load-balancers are employed to connect different IP family domains, and distribute workloads across multiple domains or internal servers. In some cases, IPv4 addresses exhaustion may not be a problem in some IDC's internal networks. IPv6 support for some applications may require some investments and workloads so IPv6 support may not be a priority. The use of NAT64 may be served to support widespread IPv6 adoption on the Internet while maintaining IPv4-only applications access.

Different strategy has been described in [RFC6883] referred to as "inside out" and "outside in". An IDC operator may implement the following practices in the NAT64-FE networking scenario.

- o Some ICPs who already have satisfactory operational experience might adopt single stack IPv6 operation in building data-center networks, servers and applications, as it allows new services delivery without having to integrate consideration of IPv4 NAT and address limitations of IPv4 networks. Stateless NAT64[RFC6145] can be used to provide services for IPv4-only enabled customers. [I-D.anderson-siit-dc] has provided further descriptions and guidelines.
- o ICPs who attempt to offer customers IPv6 support in their application farms at an early stage may likely run proxies load-balancers or translators, which are configured to handle incoming IPv6 flows and proxy them to IPv4 back-end systems. Many load balancers integrate proxy functionality. IPv4 addresses configured in the proxy may be multiplexed like a stateful NAT64 translator. A similar challenge exists once increasingly numerous users in IPv6 Internet access an IPv4 network. High loads on

load-balancers may be apt to cause additional latency, IPv4 pool exhaustion, etc. Therefore, this approach is only reasonable at an early stage. ICPs may employ dual-stack or IPv6 single stack in a further stage, since the native IPv6 is frequently more desirable than any of the transition solutions.

[RFC6144] recommends that AAAA records of load-balancers or application servers can be directly registered in the authoritative DNS servers. In this case, there is no need to deploy DNS64 name-servers. Those AAAA records can point to natively assigned IPv6 addresses or IPv4-converted IPv6 addresses[RFC6052]. Hosts are not aware of the NAT64 translator on communication path. For the testing purpose, operators could employ an independent sub domain e.g. ipv6exp.example.com to identify experimental ipv6 services to users. How to design the FQDN for the IPv6 service is out-of-scope of this document.

#### 4. High Availability

##### 4.1. Redundancy Design

High Availability (HA) is a major requirement for every service and network services. The deployment of redundancy mechanisms is an essential approach to avoid failure and significantly increase the network reliability. It's not only useful to stateful NAT64 cases, but also to stateless NAT64 gateways.

Three redundancy modes are mainly used: cold standby, warm standby and hot standby.

- o Cold standby HA devices do not replicate the NAT64 states from the primary equipment to the backup. Administrators switch on the backup NAT64 only if the primary NAT64 fails. As a result, all existing established sessions through a failed translator will be disconnected. The translated flows will need to be recreated by end-systems. Since the backup NAT64 is manually configured to switch over to active NAT64, it may have unpredictable impacts to the ongoing services.
- o Warm standby is a flavor of the cold standby mode. Backup NAT64 would keep running once the primary NAT64 is working. This makes warm standby less time consuming during the traffic failover. Virtual Router Redundancy Protocol (VRRP)[RFC5798] can be a solution to enable automatic handover in the warm standby. It was tested that the handover takes as maximum as 1 minute if the backup NAT64 needs to take over routing and re-construct the Binding Information Bases (BIBs) for 30 million sessions. In

deployment phase, operators could balance loads on distinct NAT64s devices. Those NAT64s make a warm backup of each other.

- o Hot standby must synchronize the BIBs between the primary NAT64 and backup. When the primary NAT64 fails, backup NAT64 would take over and maintain the state of all existing sessions. The internal hosts don't have to re-connect the external hosts. The handover time has been extremely reduced. Employing Bidirectional Forwarding Detection (BFD) [RFC5880] combined with VRRP, a delay of only 35ms for 30 million sessions handover was observed during testing. Under ideal conditions hotstandby deployments could guarantee the session continuity for every service. In order to timely transmit session states, operators may have to deploy extra transport links between primary NAT64 and distant backup. The scale of synchronization data instance is depending on the particular deployment. For example, If a NAT64-CGN is served for 200,000 users, the average amount of 800, 000 sessions per second is roughly estimated for new created and expired sessions. A physical 10Gbps transport link may have to be deployed for the sync data transmission considering the amount of sync sessions at the peak and capacity redundancy

In general, cold-standby and warm-standby is simpler and less resource intensive, but it requires clients to re-establish sessions when a fail-over occurs. Hot standby increases resource consumption in order to synchronize state, but potentially achieves seamless handover. For stateless NAT64 considerations are simple, because state synchronization is unnecessary. Regarding stateful NAT64, it may be useful to investigate performance tolerance of applications and the traffic characteristics in a particular network. Some testing results are shown in the Appendix A.

Our statistics in a mobile network shown that almost 91.21% of of traffic is accounted by http/https services. These services generally don't require session continuity. Hot-standby does not offer much benefit for those sessions on this point. In fixed networks, HTTP streaming, p2p and online games would be the major traffic beneficiaries of hot-standby replication[Cisco-VNI]. Consideration should be given to the importance of maintaining bindings for those sessions across failover. Operators may also consider the Average Revenue Per User (ARPU) factors to deploy suitable redundancy mode. Warm standby may still be adopted to cover most services while hot standby could be used to upgrade Quality of Experience (QoE) using DNS64 to generate different synthetic responses for limited traffic or destinations. Further considerations are discussed at Section 6.

#### 4.2. Load Balancing

Load balancing is used to accompany redundancy design so that better scalability and resiliency could be achieved. Stateless NAT64s allow asymmetric routing while anycast-based solutions are recommended in [I-D.ietf-softwire-map-deployment]. The deployment of load balancing may make more sense to stateful NAT64s for the sake of single-point failure avoidance. Since the NAT64-CGN and NAT64-FE have distinct facilities, the following lists the considerations for each case.

- o NAT64-CGN equipment doesn't typically implement load-balancing functions onboard. Therefore, the gateways have to resort to DNS64 or internal host's behavior. Once DNS64 is deployed, the load balancing can be performed by synthesizing AAAA response with different IPv6 prefixes. For the applications not requiring DNS resolver, internal hosts could learn multiple IPv6 prefixes through the approaches defined in[RFC7050] and then select one based on a given prefix selection policy.
- o A dedicated Load Balancer could be deployed at front of a NAT64-FE farm. Load Balancer uses proxy mode to redirect the flows to the appropriate NAT64 instance. Stateful NAT64s require a deterministic pattern to arrange the traffic in order to ensure outbound/inbound flows traverse the identical NAT64. Therefore, static scheduling algorithms, for example source-address based policy, is preferred. A dynamic algorithm, for example Round-Robin, may have impacts on applications seeking session continuity, which described in the Table 1.

#### 5. Source Address Transparency

##### 5.1. Traceability

Traceability is required in many cases such as identifying malicious attacks sources and accounting requirements. Operators are asked to record the NAT64 log information for specific periods of time. In our lab testing, the log information from 200,000 subscribers have been collected from a stateful NAT64 gateway for 60 days. Syslog[RFC5424] has been adopted to transmit log message from NAT64 to a log station. Each log message contains transport protocol, source IPv6 address:port, translated IPv4 address: port and timestamp. It takes almost 125 bytes in ASCII format. It has been verified that the rate of traffic flow is around 72 thousand flows per second and the volume of recorded information reaches up to 42.5 terabytes in the raw format. The volume is 29.07 terabytes in a compact format. At scale, operators have to build up dedicated transport links, storage system and servers for the purpose of managing such logging.

There are also several improvements that can be made to mitigate the issue. For example, stateful NAT64 could configure with bulk port allocation method. Once a subscriber creates the first session, a number of ports are pre-allocated. A bulk allocation message is logged indicating this allocation. Subsequent session creations will use one of the pre-allocated port and hence does not require logging. The log volume in this case may be only one thousandth of dynamic port allocation. Some implementations may adopt static port-range allocations [I-D.donley-behave-deterministic-cgn] which eliminates the need for per-subscriber logging. As a side effect, the IPv4 multiplexing efficiency is decreased regarding to those methods. For example, the utilization ratio of public IPv4 address is dropped approximately to 75% when NAT64 gateway is configured with bulk port allocation (The lab testing allocates each subscriber with 400 ports). In addition, port-range based allocation should also consider port randomization described in [RFC6056]. A trade-off among address multiplexing efficiency, logging storage compression and port allocation complexity should be considered. More discussions could be found in [I-D.chen-sunset4-cgn-port-allocation]. The decision can balance usable IPv4 resources against investments in log systems.

## 5.2. Geo-location

IP addresses are usually used as inputs to geo-location services. The use of address sharing prevents these systems from resolving the location of a host based on IP address alone. Applications that assume such geographic information may not work as intended. The possible solutions listed in [RFC6967] are intended to bridge the gap. However, those solutions can only provide a sub-optimal substitution to solve the problem of host identification, in particular it may not today solve problems with source identification through translation. The following lists current practices to mitigate the issue.

- o Operators who adopt NAT64-FE may leverage the application layer proxies, e.g. X-Forwarded-For (XFF) [I-D.ietf-appsawg-http-forwarded], to convey the IPv6 source address in HTTP headers. Those messages would be passed on to web-servers. The log parsing tools are required to be able to support IPv6 and may lookup Radius servers for the target subscribers based on IPv6 addresses included in XFF HTTP headers. XFF is the de facto standard which has been integrated in most Load Balancers. Therefore, it may be superior to use in a NAT-FE environment. In the downsides, XFF is specific to HTTP. It restricts the usages so that the solution can't be applied to requests made over HTTPs. This makes geo-location problematic for HTTPs based services.

- o The NAT64-CGN equipment may not implement XFF. Geo-location based on shared IPv4 address is rather inaccurate in that case. Operators could subdivide the outside IPv4 address pool so an IPv6 address can be translated depending on their geographical locations. As consequence, location information can be identified from a certain IPv4 address range. [RFC6967] also enumerates several options to reveal the host identifier. Each solution likely has their-own specific usage. For the geo-location systems relying on a Radius database[RFC5580], we have investigated to deliver NAT64 BIBs and Session Table Entries (STEs) to a Radius server[I-D.chen-behave-nat64-radius-extension]. This method could provide geo-location system with an internal IPv6 address to identify each user. It can get along with [RFC5580] to convey original source address through same message bus.

## 6. Quality of Experience

### 6.1. Service Reachability

NAT64 is providing a translation capability between IPv6 and IPv4 end-nodes. In order to provide the reachability between two IP address families, NAT64-CGN has to implement appropriate application aware functions, i.e. Application Layer Gateway (ALG), where address translation is not itself sufficient and security mechanisms do not render it infeasible. Most NAT64-CGNs mainly provide FTP-ALG[RFC6384]. NAT64-FEs may have functional richness on Load Balancer, for example HTTP-ALG, HTTPS-ALG, RTSP-ALG and SMTP-ALG have been supported. Those application protocols exchange IP address and port parameters within control session, for example the "Via" field in a HTTP header, "Transport" field in a RTSP SETUP message and "Received: " header in a SMTP message. ALG functions will detect those fields and make IP address translations. It should be noted that ALGs may impact the performance on a NAT64 box to some extent. ISPs as well as content providers might choose to avoid situations where the imposition of an ALG might be required. At the same time, it is also important to remind customers and application developers that IPv6 end-to-end usage does not require ALG imposition and therefore results in a better overall user experience.

The service reachability is also subject to the IPv6 support in the client side. We tested several kinds of applications as shown in the below table to verify the IPv6 supports. The experiences of some applications are still align with [RFC6586]. For example, we have tested P2P file sharing and streaming applications including eMule v0.50a, Thunder v7.9 and PPS TV v3.2.0. It has been found there are some software issues to support IPv6 at this time. The application software would benefit from 464xlat[RFC6877] until the software adds IPv6 support.. A SIP based voice call has been tested in LTE mobile



environment as specified in [IR.92]. The voice call is failed due to the lack of NAT64 traversal when an IPv6 SIP user agent communicates with an IPv4 SIP user agent. In order to address the failure, Interactive Connectivity Establishment (ICE) described in [RFC5245] is recommended to be supported for the SIP IPv6 transition. [RFC6157] describes both signaling and media layer process, which should be followed. In addition, it may be worth to notice that ICE is not only useful for NAT traversal, but also firewall[RFC6092] traversal in native IPv6 deployment.

Different IPsec modes for VPN services have been tested, including IPsec-AH and IPsec-ESP. It has been testified IPsec-AH can't survive since the destination host detects the IP header changes and invalidate the packets. IPsec-ESP failed in our testing because the NAT64 does not translate IPsec ESP (i.e. protocol 50) packets. It has been suggested that IPsec ESP should succeed if the IPsec client supports NAT-Traversal in the IKE[RFC3947] and uses IPsec ESP over UDP[RFC3948].

Table 1: The tested applications

APPs	Results and Found Issues
Webservice	Mostly pass, some failure cases due to IPv4 Literals
Instant Message	Mostly fail, software can't support IPv6
Games	Mostly pass for web-based games; mostly fail for standalone games due to the lack of IPv6 support in software
SIP-VoIP	Fail, due to the lack of NAT64 traversal
IPsec-VPN	Fail, the translated IPsec packets are invalidated
P2P file sharing and streaming	Mostly fail, software can't support IPv6, e.g. eMule Thunder and PPS TV
FTP	Pass
Email	Pass

## 6.2. Resource Reservation

Session status normally is managed by a static timer. For example, the value of the "established connection idle-timeout" for TCP sessions must not be less than 2 hours 4 minutes[RFC5382] and 5

minutes for UDP sessions[RFC4787]. In some cases, NAT resource maybe significantly consumed by largely inactive users. The NAT translator and other customers would suffer from service degradation due to port consummation by other subscribers using the same NAT64 device. A flexible NAT session control is desirable to resolve the issues. PCP[RFC6887] could be a candidate to provide such capability. A NAT64-CGN should integrate with a PCP server, to allocate available IPv4 address/port resources. Resources could be assigned to PCP clients through PCP MAP/PEER mode. Such ability can be considered to upgrade user experiences, for example assigning different sizes of port ranges for different subscribers. Those mechanisms are also helpful to minimize terminal battery consumption and reduce the number of keep-alive messages to be sent by mobile terminal devices.

Subscribers can also benefit from network reliability. It has been discussed that hot-standby offers satisfactory experience once outage of primary NAT64 is occurred. Operators may rightly be concerned about the considerable investment required for NAT64 equipment relative to low ARPU income. For example, transport links may cost much, because primary NAT64 and backup are normally located at different locations, separated by a relatively large distance. Additional cost has to be assumed to ensure the connectivity quality. However, that may be necessary to some applications, which are delay-sensitive and seek session continuity, for example on-line games and live-streaming. Operators may be able to get added-values from those services by offering first-class services. It can be pre-configured on the gateway to hot-standby modes depending on subscriber's profile. The rest of other sessions can be covered by cold/warm standby.

## 7. MTU Considerations

IPv6 requires that every link in the internet have an Maximum Transmission Unit (MTU) of 1280 octets or greater[RFC2460]. However, in case of NAT64 translation deployment, some IPv4 MTU constrained link will be used in some communication path and originating IPv6 nodes may therefore receive an ICMP Packet Too Big (PTB) message, reporting a Next-Hop MTU less than 1280 bytes. The result would be that IPv6 allows packets to contain a fragmentation header, without the packet being fragmented into multiple pieces. A NAT64 would receive IPv6 packets with fragmentation header in which "M" flag equal to 0 and "Fragment Offset" equal to 0. Those packets likely impact other fragments already queued with the same set of {IPv6 Source Address, IPv6 Destination Address, Fragment Identification}. If the NAT64 box is compliant with [RFC5722], there is risk that all the fragments have to be dropped.

[RFC6946] discusses how this situation could be exploited by an attacker to perform fragmentation-based attacks, and also proposes an improved handling of such packets. It required enhancements on NAT64 gateway implementations to isolate packet's processing. NAT64 should follow the recommendation and take steps to prevent the risks of fragmentation.

Another approach that potentially avoids this issue is to configure IPv4 MTU more than 1260 bytes. It would forbid the occurrence of PTB smaller than 1280 bytes. Such an operational consideration is hard to universally apply to the legacy "IPv4 Internet" NAT64-CGN bridged. However, it's a feasible approach in NAT64-FE cases, since a IPv4 network NAT64-FE connected is rather well-organized and operated by a IDC operator or content provider. Therefore, the MTU of IPv4 network in NAT64-FE case are strongly recommended to set to more than 1260 bytes.

## 8. ULA Usages

Unique Local Addresses (ULAs) are defined in [RFC4193] to be renumbered within a network site for local communications. Operators may use ULAs as NAT64 prefixes to provide site-local IPv6 connectivity. Those ULA prefixes are stripped when the packets going to the IPv4 Internet, therefore ULAs are only valid in the IPv6 site. The use of ULAs could help in identifying the translation traffic. [I-D.ietf-v6ops-ula-usage-recommendations] provides further guidance for the ULAs usages.

We configure ULAs as NAT64 prefixes on a NAT64-CGN. If a host is only assigned with an IPv6 address and connected to NAT64-CGN, when connect to an IPv4 service, it would receive AAAA record generated by the DNS64 with the ULA prefix. A Global Unicast Address (GUA) will be selected as the source address to the ULA destination address. When the host has both IPv4 and IPv6 address, it would initiate both A and AAAA record lookup, then both original A record and DNS64-generated AAAA record would be received. A host, which is compliant with [RFC6724], will never prefer ULA over IPv4. An IPv4 path will be always selected. It may be undesirable because the NAT64-CGN will never be used. Operators may consider to add additional site-specific rows into the default policy table for host address selection in order to steer traffic flows going through NAT64-CGN. However, it involves significant costs to change terminal's behavior. Therefore, operators are not suggested to configure ULAs on a NAT64-CGN.

ULAs can't work when hosts transit the Internet to connect with NAT64. Therefore, ULAs are inapplicable to the case of NAT64-FE.

## 9. Security Considerations

This document presents the deployment experiences of NAT64 in CGN and FE scenarios. In general, RFC 6146[RFC6146] provides TCP-tracking, address-dependent filtering mechanisms to protect NAT64 from Distributed Denial of Service (DDoS). In NAT64-CGN cases, operators also could adopt unicast Reverse Path Forwarding (uRPF)[RFC3704] and black/white-list to enhance the security by specifying access policies. For example, NAT64-CGN should forbid establish NAT64 BIB for incoming IPv6 packets if uRPF in Strict or Loose mode check does not pass or whose source IPv6 address is associated to black-lists.

The stateful NAT64-FE creates state and maps that connection to an internally-facing IPv4 address and port. An attacker can consume the resources of the NAT64-FE device by sending an excessive number of connection attempts. Without a DDoS limitation mechanism, the NAT64-FE is exposed to attacks. Load Balancer is recommended to enable the capabilities of line rate DDOS defense, such as the employment of SYN PROXY-COOKIE. Security domain division is necessary as well in this case. Therefore, Load Balancers could not only serve for optimization of traffic distribution, but also prevent service from quality deterioration due to security attacks.

The DNS64 process will potentially interfere with the DNSSEC functions[RFC4035], since DNS response is modified and DNSSEC intends to prevent such changes. More detailed discussions can be found in [RFC6147].

## 10. IANA Considerations

This memo includes no request to IANA.

## 11. Acknowledgements

The authors would like to thank Jari Arkko, Dan Wing, Remi Despres, Fred Baker, Hui Deng, Iljitsch van Beijnum, Philip Matthews, Randy Bush, Mikael Abrahamsson, Lorenzo Colitti, Sheng Jiang, Nick Heatley, Tim Chown, Gert Doering and Simon Perreault for their helpful comments.

Many thanks to Wesley George, Lee Howard and Satoru Matsushima for their detailed reviews.

The authors especially thank Joel Jaeggli and Ray Hunter for his efforts and contributions on editing which substantially improves the legibility of the document.

Thanks to Cameron Byrne who was an active co-author of some earlier versions of this draft.

## 12. Additional Author List

The following are extended authors who contributed to the effort:

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

QiBo Niu  
ZTE  
50,RuanJian Road.  
YuHua District,  
Nan Jing 210012  
P.R.China  
Email: niu.qibo@zte.com.cn

## 13. References

### 13.1. Normative References

- [I-D.ietf-appsawg-http-forwarded]  
Petersson, A. and M. Nilsson, "Forwarded HTTP Extension",  
draft-ietf-appsawg-http-forwarded-10 (work in progress),  
October 2012.
- [RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6  
(IPv6) Specification", RFC 2460, December 1998.
- [RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed  
Networks", BCP 84, RFC 3704, March 2004.
- [RFC3947] Kivinen, T., Swander, B., Huttunen, A., and V. Volpe,  
"Negotiation of NAT-Traversal in the IKE", RFC 3947,  
January 2005.
- [RFC3948] Huttunen, A., Swander, B., Volpe, V., DiBurro, L., and M.  
Stenberg, "UDP Encapsulation of IPsec ESP Packets", RFC  
3948, January 2005.

- [RFC4035] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Protocol Modifications for the DNS Security Extensions", RFC 4035, March 2005.
- [RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.
- [RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.
- [RFC5245] Rosenberg, J., "Interactive Connectivity Establishment (ICE): A Protocol for Network Address Translator (NAT) Traversal for Offer/Answer Protocols", RFC 5245, April 2010.
- [RFC5382] Guha, S., Biswas, K., Ford, B., Sivakumar, S., and P. Srisuresh, "NAT Behavioral Requirements for TCP", BCP 142, RFC 5382, October 2008.
- [RFC5424] Gerhards, R., "The Syslog Protocol", RFC 5424, March 2009.
- [RFC5580] Tschofenig, H., Adrangi, F., Jones, M., Lior, A., and B. Aboba, "Carrying Location Objects in RADIUS and Diameter", RFC 5580, August 2009.
- [RFC5722] Krishnan, S., "Handling of Overlapping IPv6 Fragments", RFC 5722, December 2009.
- [RFC5798] Nadas, S., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, March 2010.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.
- [RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", RFC 6146, April 2011.

- [RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 2011.
- [RFC6157] Camarillo, G., El Malki, K., and V. Gurbani, "IPv6 Transition in the Session Initiation Protocol (SIP)", RFC 6157, April 2011.
- [RFC6384] van Beijnum, I., "An FTP Application Layer Gateway (ALG) for IPv6-to-IPv4 Translation", RFC 6384, October 2011.
- [RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.
- [RFC6724] Thaler, D., Draves, R., Matsumoto, A., and T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, September 2012.
- [RFC6887] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", RFC 6887, April 2013.
- [RFC6946] Gont, F., "Processing of IPv6 "Atomic" Fragments", RFC 6946, May 2013.
- [RFC7050] Savolainen, T., Korhonen, J., and D. Wing, "Discovery of the IPv6 Prefix Used for IPv6 Address Synthesis", RFC 7050, November 2013.

### 13.2. Informative References

- [Alexa] Alexa, "<http://www.alexa.com/topsites>", April 2013.
- [Cisco-VNI] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2012-2017, <http://ciscovni.com/forecast-widget/index.html>", May 2013.
- [I-D.anderson-siit-dc] Anderson, T., "Stateless IP/ICMP Translation in IPv6 Data Centre Environments", draft-anderson-siit-dc-00 (work in progress), November 2012.
- [I-D.chen-behave-nat64-radius-extension] Chen, G. and D. Binet, "Radius Attributes for Stateful NAT64", draft-chen-behave-nat64-radius-extension-00 (work in progress), July 2013.

- [I-D.chen-sunset4-cgn-port-allocation]  
Chen, G., Tsou, T., Donley, C., and T. Taylor, "Analysis of NAT64 Port Allocation Method", draft-chen-sunset4-cgn-port-allocation-03 (work in progress), February 2014.
- [I-D.donley-behave-deterministic-cgn]  
Donley, C., Grundemann, C., Sarawat, V., Sundaresan, K., and O. Vautrin, "Deterministic Address Mapping to Reduce Logging in Carrier Grade NAT Deployments", draft-donley-behave-deterministic-cgn-07 (work in progress), January 2014.
- [I-D.ietf-softwire-map-deployment]  
Qiong, Q., Chen, M., Chen, G., Tsou, T., and S. Perreault, "Mapping of Address and Port (MAP) - Deployment Considerations", draft-ietf-softwire-map-deployment-03 (work in progress), October 2013.
- [I-D.ietf-softwire-map-t]  
Li, X., Bao, C., Dec, W., Troan, O., Matsushima, S., and T. Murakami, "Mapping of Address and Port using Translation (MAP-T)", draft-ietf-softwire-map-t-05 (work in progress), February 2014.
- [I-D.ietf-softwire-stateless-4v6-motivation]  
Boucadair, M., Matsushima, S., Lee, Y., Bonness, O., Borges, I., and G. Chen, "Motivations for Carrier-side Stateless IPv4 over IPv6 Migration Solutions", draft-ietf-softwire-stateless-4v6-motivation-05 (work in progress), November 2012.
- [I-D.ietf-v6ops-ula-usage-recommendations]  
Liu, B. and S. Jiang, "Recommendations of Using Unique Local Addresses", draft-ietf-v6ops-ula-usage-recommendations-02 (work in progress), February 2014.
- [I-D.kaliwoda-sunset4-dual-ipv6-coexist]  
Kaliwoda, A. and D. Binet, "Co-existence of both dual-stack and IPv6-only hosts", draft-kaliwoda-sunset4-dual-ipv6-coexist-01 (work in progress), October 2012.
- [I-D.wing-dhc-dns-reconfigure]  
Patil, P., Boucadair, M., Wing, D., and T. Reddy, "DHCPv6 Dynamic Reconfiguration", draft-wing-dhc-dns-reconfigure-02 (work in progress), September 2013.



- [IR.92] Global System for Mobile Communications Association (GSMA), , "IMS Profile for Voice and SMS Version 7.0", March 2013.
- [RFC6036] Carpenter, B. and S. Jiang, "Emerging Service Provider Scenarios for IPv6 Deployment", RFC 6036, October 2010.
- [RFC6056] Larsen, M. and F. Gont, "Recommendations for Transport-Protocol Port Randomization", BCP 156, RFC 6056, January 2011.
- [RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service", RFC 6092, January 2011.
- [RFC6144] Baker, F., Li, X., Bao, C., and K. Yin, "Framework for IPv4/IPv6 Translation", RFC 6144, April 2011.
- [RFC6346] Bush, R., "The Address plus Port (A+P) Approach to the IPv4 Address Shortage", RFC 6346, August 2011.
- [RFC6459] Korhonen, J., Soininen, J., Patil, B., Savolainen, T., Bajko, G., and K. Iisakkila, "IPv6 in 3rd Generation Partnership Project (3GPP) Evolved Packet System (EPS)", RFC 6459, January 2012.
- [RFC6586] Arkko, J. and A. Keranen, "Experiences from an IPv6-Only Network", RFC 6586, April 2012.
- [RFC6877] Mawatari, M., Kawashima, M., and C. Byrne, "464XLAT: Combination of Stateful and Stateless Translation", RFC 6877, April 2013.
- [RFC6883] Carpenter, B. and S. Jiang, "IPv6 Guidance for Internet Content Providers and Application Service Providers", RFC 6883, March 2013.
- [RFC6967] Boucadair, M., Touch, J., Levis, P., and R. Penno, "Analysis of Potential Solutions for Revealing a Host Identifier (HOST\_ID) in Shared Address Deployments", RFC 6967, June 2013.

#### Appendix A. Testing Results of Application Behavior

We test several application behaviors in a lab environment to evaluate the impact when a primary NAT64 is out of service. In this testing, participants are asked to connect a IPv6-only WiFi network

using laptops, tablets or mobile phones. NAT64 is deployed as the gateway to connect Internet service. The tested applications are shown in the below table. Cold standby, warm standby and hot standby are taken turn to be tested. The participants may experience service interruption due to the NAT64 handover. Different interruption intervals are tested to gauge application behaviors. The results are illuminated as below.

Table 2: The acceptable delay of applications

APPs	Acceptable Interrupt Recovery	Session Continuity
Web Browse	As maximum as 6s	No
Http streaming	As maximum as 10s(cache)	Yes
Gaming	200ms~400ms	Yes
P2P streaming, file sharing	10~16s	Yes
Instant Message	1 minute	Yes
Mail	30 seconds	No
Downloading	1 minutes	No

## Authors' Addresses

Gang Chen  
 China Mobile  
 Xuanwumenxi Ave. No.32,  
 Xuanwu District,  
 Beijing 100053  
 China

Email: phdgang@gmail.com

Zhen Cao  
China Mobile  
Xuanwumenxi Ave. No.32,  
Xuanwu District,  
Beijing 100053  
China

Email: caozhen@chinamobile.com, zehn.cao@gmail.com

Chongfeng Xie  
China Telecom  
Room 708 No.118, Xizhimenneidajie  
Beijing 100035  
P.R.China

Email: xiechf@ctbri.com.cn

David Binet  
France Telecom-Orange  
Rennes  
35000  
France

Email: david.binet@orange.com

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

S. Jiang, Ed.  
Huawei Technologies Co., Ltd  
Q. Sun  
China Telecom  
I. Farrer  
Deutsche Telekom AG  
Y. Bo  
Huawei Technologies Co., Ltd  
T. Yang  
China Mobile  
July 15, 2013

Analysis of Semantic Embedded IPv6 Address Schemas  
draft-jiang-v6ops-semantic-prefix-04

Abstract

This informational document discusses the use of embedded semantics within IPv6 address schemas. Network operators who have large IPv6 address space may choose to embed some semantics into their IPv6 addressing by assigning additional significance to specific bits within the prefix. By embedding semantics into IPv6 prefixes, the semantics of packets can be easily inspected. This can simplify the packet differentiation process. However, semantic embedded IPv6 address schemas have their own operational cost and even potential pitfalls. Some complex semantic embedded IPv6 address schemas may also require new technologies in addition to existing Internet protocols.

The document aims to understand the usage of semantic embedded IPv6 address schemas, and neutrally analyze on the associated advantages, drawbacks and technical gaps for more complex address schemas.

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 January 16, 2014.

#### Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction . . . . .	3
2. Terminology . . . . .	4
3. Understanding of Semantic IPv6 Prefix Address Schema . . . . .	4
3.1. Overview of Semantic IPv6 Prefix Address Schema . . . . .	4
3.2. Existing Approaches to Traffic Differentiation . . . . .	5
3.3. Justification for Semantics with the IPv6 Prefix . . . . .	6
3.4. The Semantic Prefix Domain . . . . .	7
3.5. The Embedded Semantics . . . . .	8
3.6. Network Operations Based on Semantic Prefixes . . . . .	8
4. Potential Benefits . . . . .	9
5. Potential Drawbacks . . . . .	10
6. Gaps for complex semantic prefix scenarios . . . . .	11
6.1. Semantic Notification in the Network . . . . .	11
6.2. Semantic Relevant Interactions between Hosts and the Network . . . . .	12
6.3. Additional Technical Extensions . . . . .	12
7. IANA Considerations . . . . .	13
8. Change Log (removed by RFC editor) . . . . .	13
9. Security Considerations . . . . .	14
10. Acknowledgements . . . . .	14
11. References . . . . .	14
11.1. Normative References . . . . .	14
11.2. Informative References . . . . .	15
Appendix A. An ISP Semantic Prefix Example . . . . .	15
A.1. Function Type Semantic Bits . . . . .	16
A.2. Network Device Type Bits within Network Device Address Space . . . . .	17
A.3. Subscriber Type Bits within Subscriber Address Space . . . . .	17
A.4. Service Platform Type Bits within Service Platform	

Address Space . . . . .	18
Appendix B. An Enterprise Semantic Prefix example . . . . .	19
Appendix C. A Multi-Prefix Semantic example . . . . .	20
Authors' Addresses . . . . .	21

## 1. Introduction

As the global Internet expands, it is being used for an increasingly diverse range of services. These services place differentiated requirements upon packet delivery networks meaning that Internet Service Providers and enterprises need to be aware of more information about each packet in order to best meet a specific service's needs. Dividing a network into different subnets according to different semantics is already widely existing today, mostly motivated by either topological aspects, logical user/device groups, and/or trust/security domains.

In order to inspect the semantics of packets so that they can be treated differently, some network operators have chosen to embed semantics into IPv6 prefixes. Routers and other intermediary devices can easily apply relevant policies as required. User types, service types, applications, security requirements, traffic identity types, quality requirements and other criteria may be used according to how a network operator may want to differentiate its services. Packet-level differentiation can also enable flow-level and user-level differentiation. Consequently, the network operators can treat network packets differently and efficiently. It is believed this mechanism can simplify the management and maintenance of networks.

However, semantic embedded IPv6 address schemas come with their own operational cost and even pitfalls. Some complex semantic embedded IPv6 address schemas may also require technologies additional to existing Internet protocols.

While network operators, who already have large IPv6 address space allocations, are free to plan and deploy addressing in their preferred way (including semantic embedded IPv6 address schemas), it is useful to analyze the benefits and drawbacks of a semantic approach to addressing.

The document only discusses the usage of semantics within a single network, or group of interconnected networks which share a common addressing policy, referred to as a Semantic Prefix Domain.

This document does not intend to suggest the standardization of any common global semantics. It does not intend to draw any conclusions, either recommending this kind of address schemas or not. It aims to provide network operators with relevant information to use in the creation of their own addressing policy.

## 2. Terminology

The following terms are used throughout this document:

**Semantic Prefix:** A flexible-length IPv6 prefix which embeds certain semantics.

**Semantic Prefix Domain:** A portion of the Internet over which a consistent semantic-prefix based policy is in operation.

**Semantic Prefix Policy:** A policy based on the embedded semantics within IPv6 prefix.

## 3. Understanding of Semantic IPv6 Prefix Address Schema

Some network operators (either ISPs or enterprise network operators), who have large IPv6 address space, have chosen to embed certain pre-defined semantics into their IPv6 address schemas by assigning additional significance to specific bits within the prefix. The IPv6 addresses of each packet can then explicitly express semantics. Consequently, intermediate devices can easily apply relevant packet differentiating operations accordingly. This mechanism may divert much network complexity to the planning and management of IPv6 addressing and IP address based policies.

For illustrations of how semantic prefixes could be applied in real-world scenarios, Appendix A describes an ISP example semantic IPv6 prefix address schema; Appendix B introduces an enterprise semantic IPv6 prefix example; and Appendix C introduces an enterprise example in which a multiple-site enterprise network with several prefixes of different lengths is organized as a single, contiguous Semantic Prefix Domain.

### 3.1. Overview of Semantic IPv6 Prefix Address Schema

A network operator first plans their IPv6 address schema, in which useful semantics (see Section 3.5) are embedded into prefix. They then delegate prefixes with the corresponding semantics to users. The users generate their IPv6 addresses based on assigned prefixes. Then, when the IPv6 stack on the user devices forms packets, the source addresses comprise compliance semantics. For trust reasons, the filters on the edge router may drop packets which are not compliant with assigned prefixes.

The embedded semantics are only meaningful within a network domain which implements a single policy (see Section 3.4). Different service providers may make very different choices regarding the specific semantics which are relevant to their networks. Therefore, it is not possible or even desirable to attempt to standardize a general semantic prefix policy.

Forwarding policies, access control lists, policy-based routing, security isolation and other network operations (see Section 3.6) can be easily applied according to semantics, which are self-expressed by the source address of every packet. Also, the semantics of the destination address may be taken in account if the destination is in the same Semantic Prefix Domain or the peer Semantic Prefix Domain whose semantics has been notified.

### 3.2. Existing Approaches to Traffic Differentiation

There are several existing approaches which have been developed that can assist operators in identifying and marking traffic. These solutions were mainly developed in the IPv4 era, where the IP address is used as a host locator and little else. The limited capacity of a 32-bit IPv4 address provides very little room for encoding additional information. Correspondingly, these approaches are indirect, inefficient and expensive for operators.

#### 3.2.1. Differentiated Services

Quality of Service (QoS) based on and Differentiated Services [RFC2474] is a widely deployed framework specifying a simple and scalable coarse-grained mechanism for classifying and managing network traffic. But in a service provider's network, DiffServ codepoint (DSCP) values cannot be trusted when they are set by the customer as these are arbitrary values.

In real-world scenarios, ISPs deploy "remarking" points at the customer edge of their network, re-classifying received packets by rewriting the DSCP field according to local policy using information such as the source/destination address, IP protocol number and transport layer source/destination ports.



The traffic classification process leads to increased packet processing overhead and complexity at the edge of the service provider's network.

DSCP mechanism abstracts all the semantics into a single-dimension service classes. This abstract processing has lost a lot of semantic information, which providers want to inspect for every packet, then process the packet accordingly.

The DSCP in the IPv6 header traffic class field allows 6-bits for encoding service provider specific information related to the contents of the packet. Whilst this is a useful part of an overall packet differentiation architecture, the relative small number of available bits (when compared to the available number of bits within the service providers prefix) means that it cannot be used in isolation.

### 3.2.2. Deep Packet Inspection

Deep Packet Inspection (DPI) may also be used by ISPs to learn the characteristics of users packets. This involves looking into the packet well beyond the network-layer header to identify the specific application traffic type. Once identified, the traffic type can be used as an input for setting the packet's DSCP or other actions.

But DPI is expensive both in processing costs and latency. The processing costs means that dedicated infrastructure is necessary to carry out the function. The incurred latency may be too much for use with any delay/jitter sensitive applications. As a result, DPI is difficult for large-scale deployment and it's usage is usually limited to small and specific functions in the network. In short, it is not scalable, and cannot support realtime network operations.

### 3.3. Justification for Semantics with the IPv6 Prefix

Although the interface identifier portion of an IPv6 address has arbitrary bits and extension headers can carry significantly more information, these fields can not be trusted by network operators. Users may easily change the setting of interface identifier or extension headers in order to obtain undeserved priorities/privileges, while servers or enterprise users may be much more self-restricted since they are charged accordingly.

With proper access control filters deployed, the prefix can be trusted by the network operators and is simple to inspect in the IP header of a packet. The packets with the noncompliance source addresses should be filtered. The prefix is delegated by the network and therefore the network is able to detect any undesired

modifications and filter the packet accordingly. This also makes it possible for the service provider to increase the level of trust in a customer-generated packet. If the packet has an source or destination address which is outside of the network operator's policy then a session will simply fail to establish.

### 3.4. The Semantic Prefix Domain

A Semantic Prefix Domain is a portion of the Internet over which a consistent set of semantic-prefix-based policies are administered in a coordinated fashion. It is analogous to a Differentiated Services Domain [RFC2474]. Some of the characteristics that a single Semantic Prefix Domain could represent include:

- a. Administrative domains
- b. Autonomous systems
- c. Trust regions
- d. Network technologies
- e. Hosts
- f. Routers
- g. User groups
- h. Services
- i. Traffic groups
- j. Applications

A Semantic Prefix Domain has a set of pre-defined semantic definitions, which are only meaningful locally. Without an efficient semantics notification, exchanging mechanism or service agreement, the definitions of semantics are only meaningful within local Semantic Prefix Domain. Agreements on definitions between network operators could be made. However, this may involve trust models among network operators. Sharing semantic definition among Semantic Prefix Domains enables more semantic based network operations.

An enterprise Semantic Prefix Domain may span several physical networks and traverse ISP networks. However, when an interim network is traversed (such as when an intermediary ISP is used for interconnectivity), the relevance of the semantics is limited to network domains that share a common Semantic Prefix Policy.

If an ISP has several non-contiguous address blocks, they may be organized as a single Semantic Prefix Domain if the same Semantic Prefix Policy is shared across these non-contiguous address blocks.

### 3.5. The Embedded Semantics

The size of the operator assigned prefix means that there is potentially much more scope for embedding semantics than has previously been possible. The following list describes some suggested semantics which may be useful to network operators besides source/destination location:

- a. User types
- b. Applications
- c. Security domain
- d. Traffic identity types
- e. Quality requirements
- f. Geo-location

The selection of semantics varies among different network operators. They may choose one or more semantics to be embedded into their IPv6 address schemas, depending on what is important for them and what may trigger packet differentiation processes in their networks. The selection criterion and the impact of each choice are out of scope of this document.

### 3.6. Network Operations Based on Semantic Prefixes

From the explicit semantics contained within the addresses of each packet, many network operations can be applied. Compared with traditional operations, these operations are easier to realize and stable. Although detailed operation vary depending on various embedded semantics, the network operations based on semantic prefix can be abstracted into following categories:

- a. Statistic based on certain semantic. Any embedded semantic can be set as a statistic condition. In other words, any embedded semantic can be measured independently.
- b. Differentiate packet processing. Many packet processing operations can be applied based on the semantic differentiation, such as queueing, path selection, forwarding to certain process devices, etc.

- c. Security isolation. A set of packet filters that are based on semantic can fulfil network security isolation.
- d. Access control. Resource access, authentication, service access can be directly based on semantics.
- e. Resource allocation. Resources, such as bandwidth, fast queue, caching, etc., can be allocated or reserved for certain semantic users/packets.
- f. Virtualization. Within a Semantic Prefix Domain, organizing virtual networks is simplified by assigning all the nodes the same semantic identifier so that the packets from them can be distinguished from other virtual networks.

It should also be noticed that these operations do not have to be processed on the same single device. They may be separated among network devices. In other words, if there are multiple semantics in a Semantics Prefix Domain, various semantics may be understood and treated on different network devices. It is not necessary for all network devices in such domain to capable of understanding all semantics.

#### 4. Potential Benefits

Depending on various embedded semantics, different beneficial scenarios can be expected.

- a. Semantic prefix address schema provides a directly and explicitly mechanism for packet inspection. It improves the inspecting efficiency on IPv6 network devices.
- b. Simplified measurement and statistics gathering: the semantic prefix provides explicit identifiers which can be used for measurement and statistical information collection. This can be achieved by checking certain bits of the source and/or destination address in each packet.
- c. Simplified flow control: by applying policies according to certain bit values, packets carrying the same semantics in their source/destination addresses can.
- d. Service segregation: when service related information is encoded within the semantic prefix, this can be used to create simple access-control lists which can be applied uniformly across all network devices. Security zones are such typical services that need to be segregated.

- e. Policy aggregation: the semantic prefix allows many policies to be aggregated according to the same semantics within the policy based routing system [RFC1104].
- f. Easy dynamic reconfiguration of semantic oriented policy: network operators may want to dynamically change the policy actions that are operated on certain semantic packets. The semantic prefix allows such changes be operated easily, as only a small number of consistent policy rules need to be updated on all devices within the semantic prefix domain.
- g. Application-aware routing: embedding application information into IP addresses is the simplest way to realize application aware routing.
- h. Easy user behavior management: based on the user type reading from the addresses, any improper user behaviors can be easily detected and automatically handled by network policies.
- i. Easy network resources access rights management: the authentication of access right may already be embedded into the addresses. Simple matching policies can filter improper access requests.
- j. Easy virtualization: virtual network based on any semantics can be easily deployed using the semantic prefix mechanism.

## 5. Potential Drawbacks

- a. Address consumption caused by lower address utility rate.  
Embedding semantics into IPv6 addresses causes the network to use more of the address space than it normally would. The wastage comes from aligning. 1) A small addressing requirement for a separate type may get the same large address space as a large addressing requirement. 2) The number of types in each semantic has to align to  $2^n$ , for example, 5 types use to take 3 bits in the prefix.

Network operators should be aware they may not get more addresses because they have allocated their assigned address block(s) for semantic use without the addresses actually being in use - leading to a lower address utility rate. Although the current Regional Internet Registry (RIR) policies do not disallow such address usage, such usage has not been taken into account in calculating reasonable addressing quotients.

- b. Complexity that is created within the semantic prefix policy.  
Encoding too many semantics into prefixes can come at the expense

of future addressing flexibility. At the same time, embedding too many semantics may induce semantic overlap. Careful consideration should be taken with semantics definition.

- c. The risk of privacy/information leakage. The semantics in the address may be guessable, or leaked to outside the organisation. Therefore, some information of either subscribers or networks may be leaked, too.
- d. Burdening the host OS. In some complex semantic prefix scenarios, the semantics prefix mechanism puts extra burden on the originator. In such scenarios, host devices are given multiple IPv6 prefixes and required to choose correctly. When forming a packet, the originator of packets (normally the host OS) has to pick the right address/prefix according to the semantics to access a service.
- e. In order to perform policies based on trusted user/prefix, tight/strict access control filter linked with prefix assignment is requested. It is the filter who makes sure the prefix right. The filter should link back to other states of the user, like user authentication, etc, in order to match the packet to its properties and check whether it is mapped to right semantics or not.

## 6. Gaps for complex semantic prefix scenarios

The simplest semantic prefix model is to embed only abstracted user type semantics into the prefix. Current network architectures can support this semantic prefix model, in which each subscriber is still assigned a single prefix, while they are not notified the semantic embedded in the prefix.

In order to fulfill more benefits of the semantic prefix design, additional functions are needed to allow semantic relevant operations in networks and semantic relevant interactions with hosts.

IPv6 provides a facility for multiple addresses to be configured on a single interface. This creates a precondition for the approach that user chooses addresses differently for different purposes/usages.

### 6.1. Semantic Notification in the Network

In order to manage semantic prefixes and their relevant network actions, the network should be able to notify semantics along with prefix delegation.

When an prefix is delegated using a DHCPv6 IA\_PD [RFC3633], the associated semantics should also be propagated to the requesting router. This is particularly useful for autonomic process when a new device is connected.

## 6.2. Semantic Relevant Interactions between Hosts and the Network

The more that semantics are embedded into a prefix, the more complicated functions are needed for semantic relevant interactions between hosts and the network, such as prefix delegation, host notification, address selections, etc.

In practice, a single host may belong to multiple semantics. This means that several IPv6 addresses are configured on a single physical interface and should be selected for use depending on the service that a host wishes to access. A certain packet would only serve a certain semantic.

The host's IPv6 stack must have a mechanism for understanding these semantics in order to select the right source address when forming a packet. If the embedded semantic is application relevant, applications on the hosts should also be involved in the address choosing process: the host IPv6 stack reports multiple available addresses to the application through socket API (one example is "IPv6 Socket API for Source Address Selection" [RFC5014]). The application then needs to apply the semantic logic so that it can correctly select from the offered candidate addresses.

Although [RFC6724] provides an algorithm for source address selection, some semantic prefix policies may conflict with this algorithm. In this case, source address selection mechanisms may need further supporting functions to be developed.

## 6.3. Additional Technical Extensions

There are several areas in which the semantic prefix could be extended in order to increase the usefulness and applicability of the semantic prefix address schema. They are listed here for future study. Currently, their feasibility, usefulness and applicability are not carefully studied yet.

### - Dynamic Policy Configuration

Dynamic policy configuration would simplify the distribution of policy across devices in the semantic prefix domain. New functions or protocol extension are needed to enable dynamic changes to the policy actions in operation on certain semantic packets.

- Semantics Announcements to peer networks

A network may announce all, or some of its Semantic Prefix Policy to connected peer networks. This could be used to enable more dynamic configuration and enable traffic from different semantic prefix domains to traverse different networks whilst having the same semantic prefix policy applied. To achieve this automatically by message exchanging would require new functions or protocol extensions.

- Extension of Prefix Semantics beyond the left-most 64 bits

The prefix concept refers here to the left-most bits in the IP addresses delegated by the network management plane. The prefix could be longer than 64-bits if the network operators strictly manage the address assignment by using Dynamic Host Configuration Protocol for IPv6 (DHCPv6) [RFC3315] (but in this case standard Stateless Address AutoConfiguration - SLAAC [RFC4862] cannot be used).

- Organizing consumer/home networks according to semantics

Consumers or subscribers are currently assigned /48 or /56 prefixes. They have bits, which may also count the right-most 64 bits too, to organize their networks into subnets. These subnets may be organized according to some semantics that are meaningful for the user himself. In such scenario, the user acts as the network operator for his own network. Some additional technologies/functions may be needed to make such organizing and follow-up management efficient.

## 7. IANA Considerations

This document has no IANA considerations.

## 8. Change Log (removed by RFC editor)

draft-jiang-v6ops-semantic-prefix-04: add new pitfalls section; restructure to be a neatrul analysis document; 2013-07-15.

draft-jiang-v6ops-semantic-prefix-03: reword to emphasis this mechanism is a (not the) method that network operators use their addresses; add text to clarify the increased trust is actually from the deployment of source address filter, which is a compliance requirement by semantic prefix; restructure the document, move examples and gap analysis into appendixes, reorganize most content into a frame section; add summarized description for framework at the beginning of Section 3; add description for network operations based on semantic prefix; add a new coauthor who contributes an enterprise semantic prefix network example; combine most of draft-sun-v6ops-



semantic-usecase into the draft as ISP example in appendix;  
2013-5-28.

draft-jiang-v6ops-semantic-prefix-02: add new coauthor, re-organize  
the content, and refine the English, 2013-1-31.

draft-jiang-v6ops-semantic-prefix-01: add the concept of hierarchical  
Semantic Prefix Domain and more gap analysis, 2012-10-22.

draft-jiang-v6ops-semantic-prefix-00: resubmitted to v6ops WG.  
Removed detailed examples and recommendations for semantics bits,  
2012-10-15.

draft-jiang-semantic-prefix-01: added enterprise considerations and  
scenarios, emphasizing semantics only for local meaning and no intend  
to standardize any common global semantics, 2012-07-16.

draft-jiang-semantic-prefix-00: original version, 2012-07-09

## 9. Security Considerations

Embedding semantics in prefix is actually exposing more information  
of packets explicit. These informations may also provide convenient  
for malicious attackers to track or attack certain type of packets.  
If networks announce their local prefix semantics to their peer  
networks, it may also increase the vulnerable risk.

Prefix-based filters should be deployed, in order to protect against  
address spoofing attacks or denial of service for packets with forged  
source addresses.

## 10. Acknowledgements

Useful comments were made by Erik Nygren, Dan Wing, Nick Hilliard,  
Ray Hunter, David Farmer, Fred Baker, Joel Jaeggli, John Curran, Tim  
Chown, Ted Lemon, Owen DeLong, Lorenzo Colitti, George Michaelson,  
Joel Halpern, Vizdal Ales, Bless Roland, Manning Bill, Manfred Albert  
and other participants in the V6OPS working group.

## 11. References

### 11.1. Normative References

- [RFC1104] Braun, H., "Models of policy based routing", RFC 1104,  
June 1989.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black,  
"Definition of the Differentiated Services Field (DS

Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

- [RFC3315] Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., and M. Carney, "Dynamic Host Configuration Protocol for IPv6 (DHCPv6)", RFC 3315, July 2003.
- [RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, December 2003.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.
- [RFC4862] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, September 2007.
- [RFC6724] Thaler, D., Draves, R., Matsumoto, A., and T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, September 2012.

#### 11.2. Informative References

- [RFC5014] Nordmark, E., Chakrabarti, S., and J. Laganier, "IPv6 Socket API for Source Address Selection", RFC 5014, September 2007.

#### Appendix A. An ISP Semantic Prefix Example

This ISP semantic prefix example is abstracted from a real ISP address architecture design.

Note: for now, this example only covers unicast address within IP Version 6 Addressing Architecture [RFC4291].

For ISPs, several motivations to use semantic prefixes are as follows:

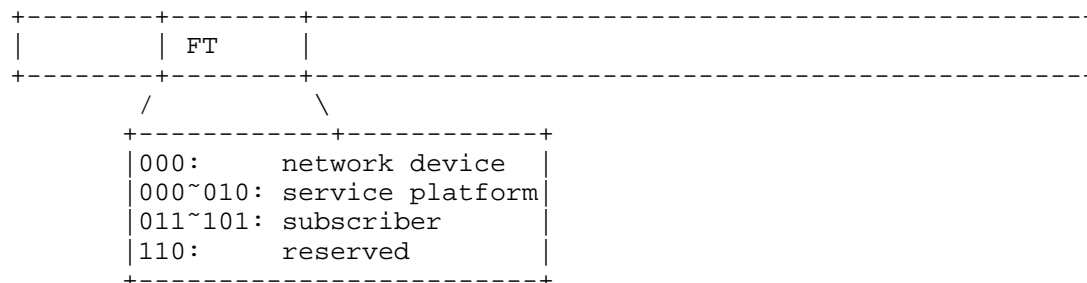
- a. Network Device management: Separated and specialized address space for network device will help to identify the network device among numerous addresses and apply policy accordingly.
- b. Differentiated user management: In ISPs' network, different kinds of customers may have different requirements for service provisioning.
- c. High-priority service guarantee: Different priorities may be divided into apply differentiated policy.

- d. Service-based Routing: ISPs may offer different routing policy for specific service platforms .e.g.video streaming, VOIP, etc.
- e. Security Control: For security requirement, operators need to take control and identify of certain devices/customers in a quick manner.
- f. Easy measurement and statistic: The semantic prefix provides explicit identifiers for measurement and statistic.

These requirements are largely falling into two categories: some is regarding to the network device features, and the others are related to services provision and subscriber identification. The functional usage of the semantics for the two categories are quite different. Therefore, an ISP semantic IPv6 prefix example is designed as a two-level hierarchical architecture, in which the first level is the function types of prefixes, and the second level is the further usage within an specific prefix type.

### A.1. Function Type Semantic Bits

Function Type (FT): the value of this field is to indicate the functional usage of this prefix. The typical types for operators include network device, subscriber and service platform.



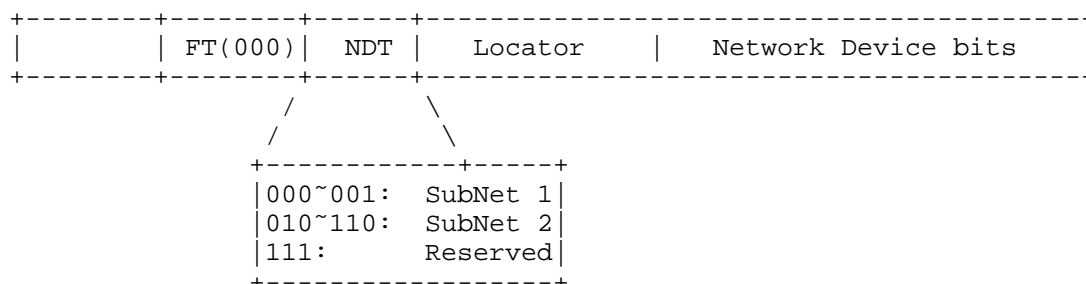
Function	Type	Bits	Example
add	int	32	add(1, 2)
sub	int	32	sub(1, 2)
mul	int	32	mul(1, 2)
div	int	32	div(1, 2)
mod	int	32	mod(1, 2)
and	int	32	and(1, 2)
or	int	32	or(1, 2)
xor	int	32	xor(1, 2)
lsh	int	32	lsh(1, 2)
rsh	int	32	rsh(1, 2)
neg	int	32	neg(1)
abs	int	32	abs(1)
iszero	bool	1	iszero(1)
isnotzero	bool	1	isnotzero(1)
ispos	bool	1	ispos(1)
isneg	bool	1	isneg(1)
isodd	bool	1	isodd(1)
iseven	bool	1	iseven(1)
isprime	bool	1	isprime(1)
isdivisible	bool	1	isdivisible(1, 2)
isdivisible3	bool	1	isdivisible3(1)
isdivisible5	bool	1	isdivisible5(1)
isdivisible7	bool	1	isdivisible7(1)
isdivisible11	bool	1	isdivisible11(1)
isdivisible13	bool	1	isdivisible13(1)
isdivisible17	bool	1	isdivisible17(1)
isdivisible19	bool	1	isdivisible19(1)
isdivisible23	bool	1	isdivisible23(1)
isdivisible29	bool	1	isdivisible29(1)
isdivisible31	bool	1	isdivisible31(1)
isdivisible37	bool	1	isdivisible37(1)
isdivisible41	bool	1	isdivisible41(1)
isdivisible43	bool	1	isdivisible43(1)
isdivisible47	bool	1	isdivisible47(1)
isdivisible53	bool	1	isdivisible53(1)
isdivisible59	bool	1	isdivisible59(1)
isdivisible61	bool	1	isdivisible61(1)
isdivisible67	bool	1	isdivisible67(1)
isdivisible71	bool	1	isdivisible71(1)
isdivisible73	bool	1	isdivisible73(1)
isdivisible79	bool	1	isdivisible79(1)
isdivisible83	bool	1	isdivisible83(1)
isdivisible89	bool	1	isdivisible89(1)
isdivisible97	bool	1	isdivisible97(1)
isdivisible101	bool	1	isdivisible101(1)
isdivisible103	bool	1	isdivisible103(1)
isdivisible107	bool	1	isdivisible107(1)
isdivisible113	bool	1	isdivisible113(1)
isdivisible127	bool	1	isdivisible127(1)
isdivisible131	bool	1	isdivisible131(1)
isdivisible137	bool	1	isdivisible137(1)
isdivisible149	bool	1	isdivisible149(1)
isdivisible151	bool	1	isdivisible151(1)
isdivisible157	bool	1	isdivisible157(1)
isdivisible163	bool	1	isdivisible163(1)
isdivisible167	bool	1	isdivisible167(1)
isdivisible173	bool	1	isdivisible173(1)
isdivisible179	bool	1	isdivisible179(1)
isdivisible181	bool	1	isdivisible181(1)
isdivisible187	bool	1	isdivisible187(1)
isdivisible191	bool	1	isdivisible191(1)
isdivisible193	bool	1	isdivisible193(1)
isdivisible197	bool	1	isdivisible197(1)
isdivisible199	bool	1	isdivisible199(1)
isdivisible211	bool	1	isdivisible211(1)
isdivisible223	bool	1	isdivisible223(1)
isdivisible227	bool	1	isdivisible227(1)
isdivisible229	bool	1	isdivisible229(1)
isdivisible233	bool	1	isdivisible233(1)
isdivisible239	bool	1	isdivisible239(1)
isdivisible241	bool	1	isdivisible241(1)
isdivisible251	bool	1	isdivisible251(1)
isdivisible257	bool	1	isdivisible257(1)
isdivisible263	bool	1	isdivisible263(1)
isdivisible269	bool	1	isdivisible269(1)
isdivisible271	bool	1	isdivisible271(1)
isdivisible277	bool	1	isdivisible277(1)
isdivisible281	bool	1	isdivisible281(1)
isdivisible283	bool	1	isdivisible283(1)
isdivisible293	bool	1	isdivisible293(1)
isdivisible307	bool	1	isdivisible307(1)
isdivisible311	bool	1	isdivisible311(1)
isdivisible313	bool	1	isdivisible313(1)
isdivisible317	bool	1	isdivisible317(1)
isdivisible331	bool	1	isdivisible331(1)
isdivisible337	bool	1	isdivisible337(1)
isdivisible347	bool	1	isdivisible347(1)
isdivisible353	bool	1	isdivisible353(1)
isdivisible359	bool	1	isdivisible359(1)
isdivisible367	bool	1	isdivisible367(1)

Figure 1

The portion of each type should be estimated according to the actual requirements for operators, in order to use the address space most efficiently. Within the above FT design, the whole ISP IPv6 address space is divided into four parts: the network device address space (1/8 of total address space), the service platform address space (2/8 of total address space), the subscriber address space (3/8 of total address space), and a reserved address space (1/8 of total address space) for future usage.

## A.2. Network Device Type Bits within Network Device Address Space

Network Device Type (NDT) indicates different types of network devices. Normally, one operator may have multiple networks, e.g. backbone network, mobile network, ISP brokered service network, etc. Using NDT field to indicate specific network within an operator may help to apply some routing policies. Locating NDT bits in the left-most bits means that a single, simple access-control list implemented across all networking devices would be enough to enforce effective traffic segregation. The Locator field is followed behind NDT.



Network Device Type Bits Example

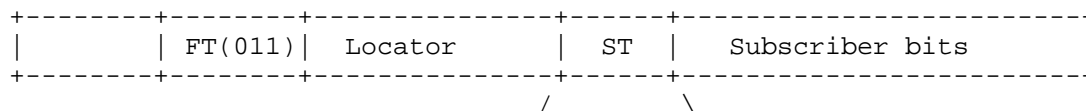
Figure 2

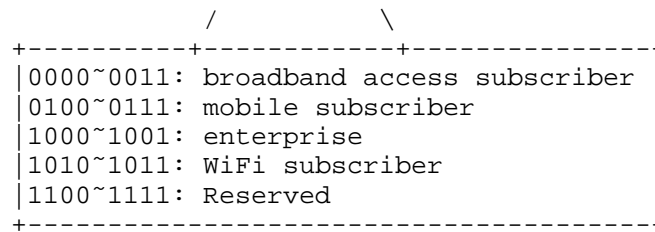
The portion of each subnet type should be estimated according to the actual requirements for operators, in order to use the address space most efficiently. Within the above NDT design, SubNet 1 is assigned 2/8 of the network device address space, SubNet 2 is assigned 5/8, and 1/8 is reserved.

## A.3. Subscriber Type Bits within Subscriber Address Space

Subscriber Type (ST) indicates different types of subscribers, e.g. wireline broadband subscriber, mobile subscriber, enterprise, WiFi, etc. This type of prefix is allocated to end users. Further, division may be taken on subscriber's priorities within a certain subscriber type.

The Locator field within subscriber address space is put before ST for better routing aggregation.





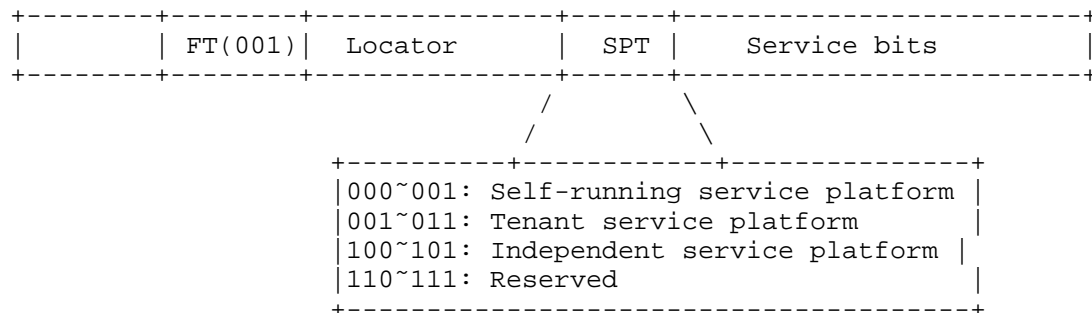
Subscriber Type Bits Example

Figure 3

The portion of each subscriber type should be estimated according to the actual requirements for operators, in order to use the address space most efficiently. Within the above ST design, the broadband access subscriber type is assigned 4/16 of the subscriber address space, the mobile subscriber is assigned 4/16, enterprise type and WiFi subscriber type are assigned 2/16 each, and 2/16 is reserved.

#### A.4. Service Platform Type Bits within Service Platform Address Space

Service Platform Type (SPT) indicates typical service platforms offered by operators. This field may have scalability problem since there are numerous types of services. It is recommended that only aggregated service platform types should be defined in this field. This type of prefix is usually allocated to service platforms in operator's data center.



Service Platform Type Bits Example

Figure 4

The portion of each subnet type should be estimated according to the actual requirements for operators, in order to use the address space most efficiently.

## Appendix B. An Enterprise Semantic Prefix example

This enterprise semantic prefix example is also abstracted from an ongoing enterprise address architecture design. This example is designed for a realtime video monitor network across a city region. The semantic prefix solution is planning to be deployed along with a strict authorization system.

Note: this example only covers unicast address within IP Version 6 Addressing Architecture [RFC4291].

For this example, the below semantics are important for the network operation and require different network behaviors.

- a. Terminal type: there are two terminal types only: monitor cameras or video receivers. They are estimated to have similar number. Network devices use another different address space.
- b. Geographic location: the city has been managed in a three-level hierarchical regionalism: district, area and street. Each level has less than 28 sub-regions. This can also be considered as a replacement of topology locator within this specific network.
- c. Authorization level: the network operator is planning to administrate the authorization in three or four levels. An receiver can access the cameras that are the same or lower authorization level.
- d. Civilian or police/government.
- e. Device attribute: this indicates the attribute of a camera device. The attribute is expressed in an abstract way, such as road traffic, hospital, nursery, bank, airport, etc. The abstracted attribute type is designed to be less than 64.
- f. Receiver Attribute: this indicates the attribute of a video receiver. The attribute is based on the receiver group, such as police, firefighter, local security, etc. The attribute/receiver group type is designed to be less than 128.

This example enterprise network has obtained a /32 address block from ISP. There is another /48 dedicated for network devices.

The first bit is Terminal type, which indicates terminal type.

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     ISP assigned block                                     |

```

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|T|   Geographic   Locator   | AL|C|Device Attr|   Device Bit |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

A semantic prefix design example for cameras

Figure 5

3-level hierarchical geographic locator takes 15 bits (each level 5 bits, 32 sub-regions). Authorization level takes 2 bits and 1 bit differentiates civilian or police/government. 6 bits is assigned for device attribute.

```

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     ISP assigned block                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|T| GeoLoc | AL|C|Receiver Attr|   Topology Locator |ReceiverBit|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

An semantic prefix design example for video receivers

Figure 6

The receiver is not as much as geographically distributed as cameras. Therefore, Geographic locator is only detailed to district level. Topology locator is needed for network forwarding and aggregation within a district. It is assigned 10 bits. Authorization level bits and civilian bit are the same with camera address space. Receive attribute takes 7 bits, giving it is designed to be up to 128.

#### Appendix C. A Multi-Prefix Semantic example

A multiple-site enterprise may have been assigned several prefixes of different lengths by its upstream ISPs. In this situation, in order to create a single, contiguous Semantic Prefix Domain, it is necessary to base the semantic prefix policy on the longest assigned prefix to ensure that there is enough addressing space to encode a consistent set of semantics across all of the assigned prefixes.

In this example, an enterprise has received a /38 address block for one site (A) and a /44 for a second site (B). They can be organized in the same Semantic Prefix Domain. The most-left 18 (site A) and 12 (site B) bits are allocated as locator. It provides topology based network aggregation. The 8 right-most bits (from bits 56 to 63) are assigned as the semantic field. In this design, the multiple-site enterprise that has been assigned two prefixes of different lengths can be organized as the same Semantic Prefix Domain. The semantic

and the Semantic Prefix Domain can traverse the intermediate ISP networks, or even public networks.

The similar situation may happen on ISPs in the future, when an ISP used up its assigned address space, or built up multiple networks in different places.

#### Authors' Addresses

Sheng Jiang (editor)  
Huawei Technologies Co., Ltd  
Q14, Huawei Campus, No.156 BeiQing Road  
Hai-Dian District, Beijing 100095  
P.R. China

Email: [jiangsheng@huawei.com](mailto:jiangsheng@huawei.com)

Qiong Sun  
China Telecom  
Room 708, No.118, Xizhimennei Street  
Beijing 100084  
P.R. China

Email: [sunqiong@ctbri.com.cn](mailto:sunqiong@ctbri.com.cn)

Ian Farrer  
Deutsche Telekom AG  
Bonn 53227  
Germany

Email: [ian.farrer@telekom.de](mailto:ian.farrer@telekom.de)

Yang Bo  
Huawei Technologies Co., Ltd  
Q21, Huawei Campus, No.156 BeiQing Road  
Hai-Dian District, Beijing 100095  
P.R. China

Email: [boyang.bo@huawei.com](mailto:boyang.bo@huawei.com)



Tianle Yang  
China Mobile  
32, Xuanwumenxi Ave. Xicheng District  
Beijing 100053  
China

Email: yangtianle@chinamobile.com

v6ops  
Internet-Draft  
Intended status: Informational  
Expires: January 15, 2014

D. Lopez  
Telefonica I+D  
Z. Chen  
China Telecom  
T. Tsou  
Huawei Technologies (USA)  
C. Zhou  
Huawei Technologies  
A. Servin  
LACNIC  
July 14, 2013

IPv6 Operational Guidelines for Datacenters  
draft-lopez-v6ops-dc-ipv6-05

Abstract

This document is intended to provide operational guidelines for datacenter operators planning to deploy IPv6 in their infrastructures. It aims to offer a reference framework for evaluating different products and architectures, and therefore it is also addressed to manufacturers and solution providers, so they can use it to gauge their solutions. We believe this will translate in a smoother and faster IPv6 transition for datacenters of these infrastructures.

The document focuses on the DC infrastructure itself, its operation, and the aspects related to DC interconnection through IPv6. It does not consider the particular mechanisms for making Internet services provided by applications hosted in the DC available through IPv6 beyond the specific aspects related to how their deployment on the Data Center (DC) infrastructure.

Apart from facilitating the transition to IPv6, the mechanisms outlined here are intended to make this transition as transparent as possible (if not completely transparent) to applications and services running on the DC infrastructure, as well as to take advantage of IPv6 features to simplify DC operations, internally and across the Internet.

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 January 15, 2014.

#### Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	4
2. Architecture and Transition Stages . . . . .	4
2.1. General Architecture . . . . .	5
2.2. Experimental Stage. Native IPv4 Infrastructure . . . . .	7
2.2.1. Off-shore v6 Access . . . . .	8
2.3. Dual Stack Stage. Internal Adaptation . . . . .	8
2.3.1. Dual-stack at the Aggregation Layer . . . . .	10
2.3.2. Dual-stack Extended OS/Hypervisor . . . . .	12
2.4. IPv6-Only Stage. Pervasive IPv6 Infrastructure . . . . .	12
3. Other Operational Considerations . . . . .	13
3.1. Addressing . . . . .	13
3.2. Management Systems and Applications . . . . .	14
3.3. Monitoring and Logging . . . . .	15
3.4. Costs . . . . .	15
4. Security Considerations . . . . .	15
4.1. Neighbor Discovery Protocol attacks . . . . .	16
4.2. Addressing . . . . .	16
4.3. Edge filtering . . . . .	17
4.4. Final Security Remarks . . . . .	17
5. IANA Considerations . . . . .	17
6. Acknowledgements . . . . .	17
7. Informative References . . . . .	17
Authors' Addresses . . . . .	19

## 1. Introduction

The need for considering the aspects related to IPv4-to-IPv6 transition for all devices and services connected to the Internet has been widely mentioned elsewhere, and it is not our intention to make an additional call on it. Just let us note that many of those services are already or will soon be located in Data Centers (DC), what makes considering the issues associated to DC infrastructure transition a key aspect both for these infrastructures themselves, and for providing a simpler and clear path to service transition.

All issues discussed here are related to DC infrastructure transition, and are intended to be orthogonal to whatever particular mechanisms for making the services hosted in the DC available through IPv6 beyond the specific aspects related to their deployment on the infrastructure. General mechanisms related to service transition have been discussed in depth elsewhere (see, for example [I-D.ietf-v6ops-icp-guidance] and [I-D.ietf-v6ops-enterprise-incremental-ipv6]) and are considered to be independent to the goal of this discussion. The applicability of these general mechanisms for service transition will, in many cases, depend on the supporting DC's infrastructure characteristics. However, this document intends to keep both problems (service vs. infrastructure transition) as different issues.

Furthermore, the combination of the regularity and controlled management in a DC interconnection fabric with IPv6 universal end-to-end addressing should translate in simpler and faster VM migrations, either intra- or inter-DC, and even inter-provider.

## 2. Architecture and Transition Stages

This document presents a transition framework structured along transition stages and operational guidance associated with the degree of penetration of IPv6 into the DC communication fabric. It is worth noting we are using these stages as a classification mechanism, and they have not to be associated with any a succession of steps from a v4-only infrastructure to full-fledged v6, but to provide a framework that operators, users, and even manufacturers could use to assess their plans and products.

There is no (explicit or implicit) requirement on starting at the stage describe in first place, nor to follow them in successive order. According to their needs and the available solutions, DC operators can choose to start or remain at a certain stage, and freely move from one to another as they see fit, without contravening this document. In this respect, the classification intends to

support the planning in aspects such as the adaptation of the different transition stages to the evolution of traffic patterns, or risk assessment in what relates to deploying new components and incorporating change control, integration and testing in highly-complex multi-vendor infrastructures.

Three main transition stages can be considered when analyzing IPv6 deployment in the DC infrastructure, all compatible with the availability of services running in the DC through IPv6:

- o Experimental. The DC keeps a native IPv4 infrastructure, with gateway routers (or even application gateways when services require so) performing the adaptation to requests arriving from the IPv6 Internet.
- o Dual stack. Native IPv6 and IPv4 are present in the infrastructure, up to whatever the layer in the interconnection scheme where L3 is applied to packet forwarding.
- o IPv6-Only. The DC has a fully pervasive IPv6 infrastructure, including full IPv6 hypervisors, which perform the appropriate tunneling or NAT if required by internal applications running IPv4.

## 2.1. General Architecture

The diagram in Figure 1 depicts a generalized interconnection schema in a DC.

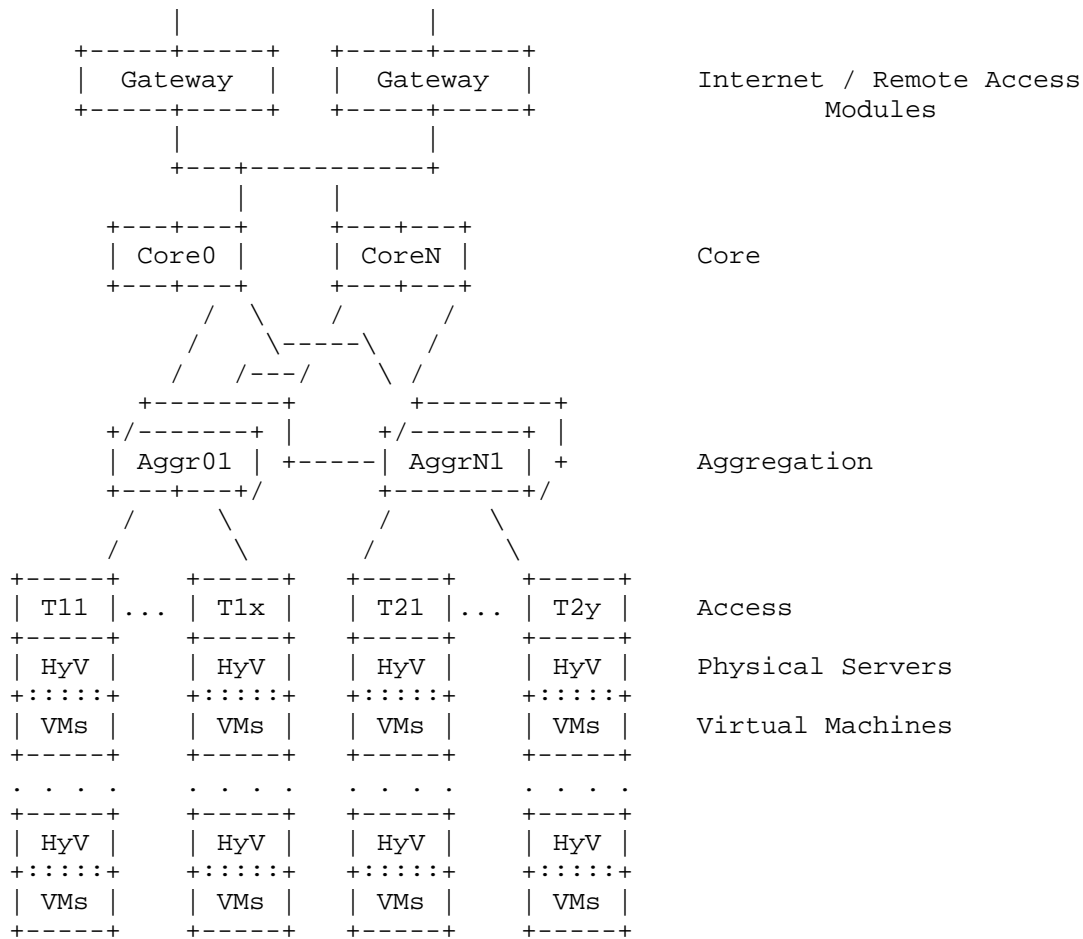


Figure 1: DC Interconnection Schema

- o Hypervisors provide connection services (among others) to virtual machines running on physical servers.
- o Access elements provide connectivity directly to/from physical servers. The access elements are typically placed either top-of-rack (ToR) or end-of-row(EoR).
- o Aggregation elements group several (many) physical racks to achieve local integration and provide as much structure as possible to data paths.
- o Core elements connect all aggregation elements acting as the DC backbone.

- o One or several gateways connecting the DC to the Internet, Branch Offices, Partners, Third-Parties, and/or other DCs. The interconnectivity to other DC may be in the form of VPNs, WAN links, metro links or any other form of interconnection.

In many actual deployments, depending on DC size and design decisions, some of these elements may be combined (core and gateways are provided by the same routers, or hypervisors act as access elements) or virtualized to some extent, but this layered schema is the one that best accommodates the different options to use L2 or L3 at any of the different DC interconnection layers, and will help us in the discussion along the document.

## 2.2. Experimental Stage. Native IPv4 Infrastructure

This transition stage corresponds to the first step that many datacenters may take (or have taken) in order to make their external services initially accessible from the IPv6 Internet and/or to evaluate the possibilities around it, and corresponds to IPv6 traffic patterns totally originated out of the DC or their tenants, being a small percentage of the total external requests. At this stage, DC network scheme and addressing do not require any important change, if any.

It is important to remark that in no case this can be considered a permanent stage in the transition, or even a long-term solution for incorporating IPv6 into the DC infrastructure. This stage is only recommended for experimentation or early evaluation purposes.

The translation of IPv6 requests into the internal infrastructure addressing format occurs at the outmost level of the DC Internet connection. This can be typically achieved at the DC gateway routers, that support the appropriate address translation mechanisms for those services required to be accessed through native IPv6 requests. The policies for applying adaptation can range from performing it only to a limited set of specified services to providing a general translation service for all public services. More granular mechanisms, based on address ranges or more sophisticated dynamic policies are also possible, as they are applied by a limited set of control elements. These provide an additional level of control to the usage of IPv6 routable addresses in the DC environment, which can be especially significant in the experimentation or early deployment phases this stage is applicable to.

Even at this stage, some implicit advantages of IPv6 application come into play, even if they can only be applied at the ingress elements:



- o Flow labels can be applied to enhance load-balancing, as described in [I-D.ietf-6man-flow-ecmp]. If the incoming IPv6 requests are adequately labeled the gateway systems can use the flow labels as a hint for applying load-balancing mechanisms when translating the requests towards the IPv4 internal network.
- o During VM migration (intra- or even inter-DC), Mobile IP mechanisms can be applied to keep service availability during the transient state.

#### 2.2.1. Off-shore v6 Access

This model is also suitable to be applied in an "off-shore" mode by the service provider connecting the DC infrastructure to the Internet, as described in [I-D.sunq-v6ops-contents-transition].

When this off-shore mode is applied, the original source address will be hidden to the DC infrastructure, and therefore identification techniques based on it, such as geolocation or reputation evaluation, will be hampered. Unless there is a specific trust link between the DC operator and the ISP, and the DC operator is able to access equivalent identification interfaces provided by the ISP as an additional service, the off-shore experimental stage cannot be considered applicable when source address identification is required.

#### 2.3. Dual Stack Stage. Internal Adaptation

This stage requires dual-stack elements in some internal parts of the DC infrastructure. This brings some degree of partition in the infrastructure, either in a horizontal (when data paths or management interfaces are migrated or left in IPv4 while the rest migrate) or a vertical (per tenant or service group), or even both.

Although it may seem an artificial case, situations requiring this stage can arise from different requirements from the user base, or the need for technology changes at different points of the infrastructure, or even the goal of having the possibility of experimenting new solutions in a controlled real-operations environment, at the price of the additional complexity of dealing with a double protocol stack, as noted in [I-D.ietf-v6ops-icp-guidance] and elsewhere.

This transition stage can accommodate different traffic patterns, both internal and external, though it better fits to scenarios of a clear differentiation of different types of traffic (external vs. internal, data vs management...), and/or a more or less even distribution of external requests. A common scenario would include native dual stack servers for certain services combined with single

stack ones for others (web server in dual stack and database servers only supporting v4, for example).

At this stage, the advantages outlined above on load balancing based on flow labels and Mobile IP mechanisms are applicable to any L3-based mechanism (intra- as well as inter-DC). They will translate into enhanced VM mobility, more effective load balancing, and higher service availability. Furthermore, the simpler integration provided by IPv6 to and from the L2 flat space to the structured L3 one can be applied to achieve simpler deployments, as well as alleviating encapsulation and fragmentation issues when traversing between L2 and L3 spaces. With an appropriate prefix management, automatic address assignment, discovery, and renumbering can be applied not only to public service interfaces, but most notably to data and management paths.

Other potential advantages include the application of multicast scopes to limit broadcast floods, and the usage of specific security headers to enhance tenant differentiation.

On the other hand, this stage requires a much more careful planning of addressing (please refer to ([RFC5375]) schemas and access control, according to security levels. While the experimental stage implies relatively few global routable addresses, this one brings the advantages and risks of using different kinds of addresses at each point of the IPv6-aware infrastructure.

## 2.3.1. Dual-stack at the Aggregation Layer

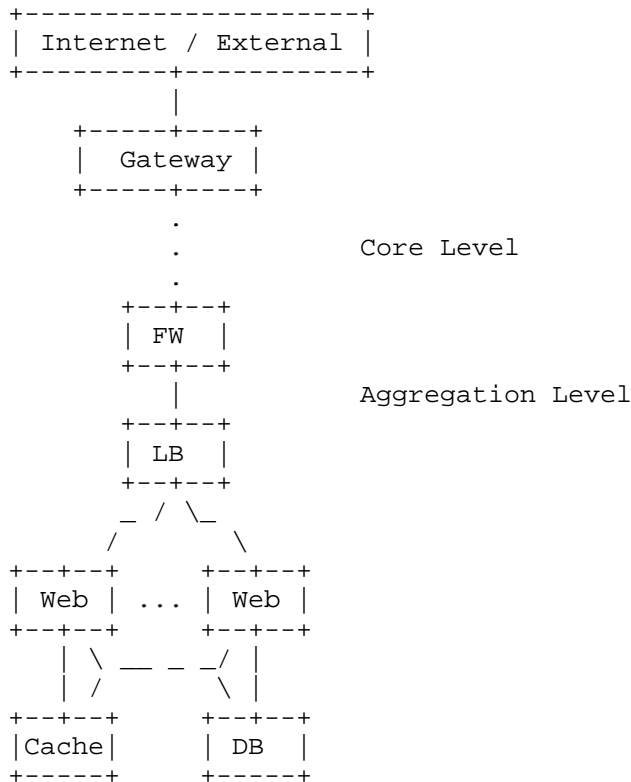


Figure 2: Data Center Application Scheme

An initial approach corresponding to this transition stage relies on taking advantage of specific elements at the aggregation layer described in Figure 1, and make them able to provide dual-stack gatewaying to the IPv4-based servers and data infrastructure.

Typically, firewalls (FW) are deployed as the security edge of the whole service domain and provides safe access control of this service domain from other function domains. In addition, some application optimization based on devices and security devices (e.g. Load Balancers, SSL VPN, IPS and etc.) may be deployed in the aggregation level to alleviate the burden of the server and to guarantee deep security, as shown in Figure 2.

The load balancer (LB) or some other boxes could be upgraded to support the data transmission. There may be two ways to achieve this

at the edge of the DC: Encapsulation and NAT. In the encapsulation case, the LB function carries the IPv6 traffic over IPv4 using an encapsulation (IPv6-in-IPv4). In the NAT case, there are already some technologies to solve this problem. For example, DNS and NAT device could be concatenated for IPv4/IPv6 translation if IPv6 host needs to visit IPv4 servers. However, this may require the concatenation of multiple network devices, which means the NAT tables needs to be synchronized at different devices. As described below, a simplified IPv4/IPv6 translation model can be applied, which could be implemented in the LB device. The mapping information of IPv4 and IPv6 will be generated automatically based on the information of the LB. The host IP address will be translated without port translation.

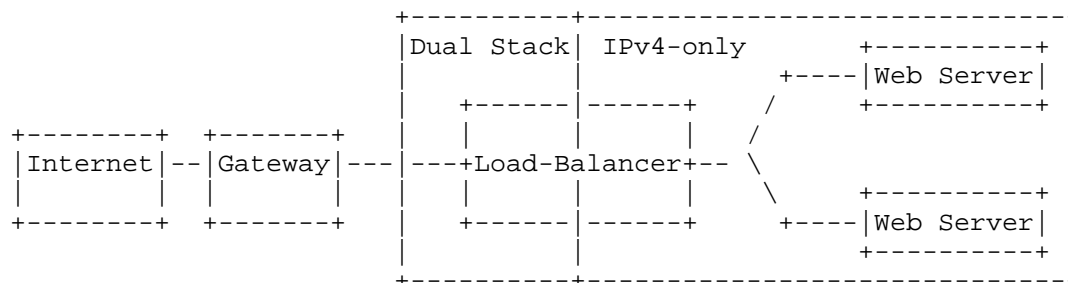


Figure 3: Dual Stack LB mechanism

As shown in Figure 3, the LB can be considered divided into two parts: The dual-stack part facing the external border, and the IPv4-only part which contains the traditional LB functions. The IPv4 DC is allocated an IPv6 prefix which is for the VSIPv6 (Virtual Service IPv6 Address). We suggest that the IPv6 prefix is not the well-known prefix in order to avoid the IPv4 routings of the services in different DCs spread to the IPv6 network. The VSIPv4 (Virtual Service IPv4 Address) is embedded in VSIPv6 using the allocated IPv6 prefix. In this way, the LB has the stateless IP address mapping between VSIPv6 and VSIPv4, and synchronization is not required between LB and DNS64 server.

The dual-stack part of the LB has a private IPv4 address pool. When IPv6 packets arrive, the dual-stack part does the one-on-one SIP (source IP address) mapping (as defined in [I-D.sunq-v6ops-contents-transition]) between IPv4 private address and IPv6 SIP. Because there will be too many UDP/TCP sessions between the DC and Internet, the IP addresses binding tables between IPv6 and IPv4 are not session-based, but SIP-based. Thus, the dual-stack part of LB builds IP binding stateful tables for the host IPv6 address and private IPv4 address of the pool. When the following

IPv6 packets of the host come from Internet to the LB, the dual stack part does the IP address translation for the packets. Thus, the IPv6 packets were translated to IPv4 packets and sent to the IPv4 only part of the LB.

#### 2.3.2. Dual-stack Extended OS/Hypervisor

Another option for deploying a infrastructure at the dual-stack stage would bring dual-stack much closer to the application servers, by requiring hypervisors, VMs and applications in the v6-capable zone of the DC to be able to operate in dual stack. This way, incoming connections would be dealt in a seamless manner, while for outgoing ones an OS-specific replacement for system calls like `gethostbyname()` and `getaddrinfo()` would accept a character string (an IPv4 literal, an IPv6 literal, or a domain name) and would return a connected socket or an error message, having executed a happy eyeballs algorithm ([RFC6555]).

If these hypothetical system call replacements were smart enough, they would allow the transparent interoperation of DCs with different levels of v6 penetration, either horizontal (internal data paths are not migrated, for example) or vertical (per tenant or service group). This approach requires, on the other hand, all the involved DC infrastructure to become dual-stack, as well as some degree of explicit application adaptation.

#### 2.4. IPv6-Only Stage. Pervasive IPv6 Infrastructure

We can consider a DC infrastructure at the final stage when all network layer elements, including hypervisors, are IPv6-aware and apply it by default. Conversely with the experimental stage, access from the IPv4 Internet is achieved, when required, by protocol translation performed at the edge infrastructure elements, or even supplied by the service provider as an additional network service.

There are different drivers that could motivate DC managers to transition to this stage. In principle the scarcity of IPv4 addresses may require to reclaim IPv4 resources from portions of the network infrastructure which no longer need them. Furthermore, the unavailability of IPv4 address would make dual-stack environments not possible anymore and careful assessments will be perfumed to asses where to use the remaining IPv4 resources.

Another important motivation to move DC operations from dual-stack to IPv6-only is to save costs and operation activities that managing a single-stack network could bring in comparison with managing two stacks. Today, besides of learning to manage two different stacks, network and system administrators require to duplicate other tasks

such as IP address management, firewalls configuration, system security hardening and monitoring among others. These activities are not just costly for the DC management, they may also may lead to configuration errors and security holes.

This stage can be also of interest for new deployments willing to apply a fresh start aligned with future IPv6 widespread usage, when a relevant amount of requests are expected to be using IPv6, or to take advantage of any of the potential benefits that an IPv6 support infrastructure can provide. Other, and probably more compelling in many cases, drivers for this stage may be either a lack of enough IPv4 resources (whether private or globally unique) or a need to reclaim IPv4 resources from portions of the network which no longer need them. In these circumstances, a careful evaluation of what still needs to speak IPv4 and what does not will need to happen to ensure judicious use of the remaining IPv4 resources.

The potential advantages mentioned for the previous stages (load balancing based on flow labels, mobility mechanisms for transient states in VM or data migration, controlled multicast, and better mapping of L2 flat space on L3 constructs) can be applied at any layer, even especially tailored for individual services. Obviously, the need for a careful planning of address space is even stronger here, though the centralized protocol translation services should reduce the risk of translation errors causing disruptions or security breaches.

[V6DCS] proposes an approach to a next generation DC deployment, already demonstrated in practice, and claims the advantages of materializing the stage from the beginning, providing some rationale for it based on simplifying the transition process. It relies on stateless NAT64 ([RFC6052], [RFC6145]) to enable access from the IPv4 Internet.

### 3. Other Operational Considerations

In this section we review some operation considerations related addressing and management issues in V6 DC infrastructure.

#### 3.1. Addressing

There are different considerations related on IPv6 addressing topics in DC. Many of these considerations are already documented in a variety of IETF documents and in general the recommendations and best practices mentioned on them apply in IPv6 DC environments. However we would like to point out some topics that we consider important to mention.

The first question that DC managers often have is the type of IPv6 address to use; that is Provider Aggregated (PA), Provider Independent (PI) or Unique Local IPv6 Addresses (ULAs) [RFC4193] Related to the use of PA vs. PI, we concur with [I-D.ietf-v6ops-icp-guidance] and [I-D.ietf-v6ops-enterprise-incremental-ipv6] that PI provides independence from the ISP and decreases renumbering issues, it may bring up other considerations as a fee for the allocation, a request process and allocation maintenance to the Regional Internet Registry, etc. In this respect, there is not a specific recommendation to use either PI vs. PA as it would depend also on business and management factors rather than pure technical.

ULAs should be used only in DC infrastructure that does not require access to the public Internet; such devices may be databases servers, application-servers, and management interfaces of web servers and network devices among others. This practice may decrease the renumbering issues when PA addressing is used, as only public faced devices would require an address change. Also we would like to know that although ULAs may provide some security the main motivation for it used should be address management.

Another topic to discuss is the length of prefixes within the DC. In general we recommend the use of subnets of 64 bits for each vlan or network segment used in the DC. Although subnet with prefixes longer than 64 bits may work, it is necessary that the reader understand that this may break stateless autoconfiguration and at least manual configuration must be employed. For details please read [RFC5375].

Address plans should follow the principles of being hierarchical and able to aggregate address space. We recommend at least to have a /48 for each data-center. If the DC provides services that require subassignment of address space we do not offer a single recommendation (i.e. request a /40 prefix from an RIR or ISP and assign /48 prefixes to customers), as this may depend on other no technical factors. Instead we refer the reader to [RFC6177].

For point-to-point links please refer to the recommendations in [RFC6164].

### 3.2. Management Systems and Applications

Data-centers may use Internet Protocol address management (IPAM) software, provisioning systems and other variety of software to document and operate. It is important that these systems are prepared and possibly modified to support IPv6 in their data models. In general, if IPv6 support for these applications has not been previously done, changes may take sometime as they may be not just

adding more space in input fields but also modifying data models and data migration.

### 3.3. Monitoring and Logging

Monitoring and logging are critical operations in any network environment and they should be carried at the same level for IPv6 and IPv4. Monitoring and management operations in V6 DC are by no means different than any other IPv6 networks environments. It is important to consider that the collection of information from network devices is orthogonal to the information collected. For example it is possible to collect data from IPv6 MIBs using IPv4 transport. Similarly it is possible to collect IPv6 data generated by Netflow9/IPFIX agents in IPv4 transport. In this way the important issue to address is that agents (i.e. network devices) are able to collect data specific to IPv6.

And as final note on monitoring, although IPv6 MIBs are supported by SNMP versions 1 and 2, we recommend to use SNMP version 3 instead.

### 3.4. Costs

It is very possible that moving from a single stack data-center infrastructure to any of the IPv6 stages described in this document may incur in capital expenditures. This may include but it is not confined to routers, load-balancers, firewalls and software upgrades among others. However the cost that most concern us is operational. Moving the DC infrastructure operations from a single-stack to a dual-stack may infer in a variety of extra costs such as application development and testing, operational troubleshooting and service deployment. At the same time, this extra cost may be seeing as saving when moving from a dual-stack DC to an IPv6-Only DC.

Depending of the complexity of the DC network, provisioning and other factors we estimate that the extra costs (and later savings) may be around between 15 to 20%.

## 4. Security Considerations

A thorough collection of operational security aspects for IPv6 network is made in [I-D.ietf-opsec-v6] . Most of them, with the probable exception of those specific to residential users, are applicable in the environment we consider in this document.



#### 4.1. Neighbor Discovery Protocol attacks

The first important issue that V6 DC manager should be aware is the attacks against Neighbor Discovery Protocol [RFC6583]. This attack is similar to ARP attacks [RFC4732] in IPv4 but exacerbated by the fact that the common size of an IPv6 subnet is /64. In principle an attacker would be able to fill the Neighbor Cache of the local router and starve its memory and processing resources by sending multiple ND packets requesting information of non-existing hosts. The result would be the inability of the router to respond to ND requests, to update its Neighbor Cache and even to forward packets. The attack does need to be launched with malicious purposes; it could be just the result of bad stack implementation behavior.

R[RFC6583] mentions some options to mitigate the effects of the attacks against NDP. For example filtering unused space, minimizing subnet size when possible, tuning rate limits in the NDP queue and to rely in router vendor implementations to better handle resources and to prioritize NDP requests.

#### 4.2. Addressing

Other important security considerations in V6 DC are related to addressing. Because of the large address space is commonly thought that IPv6 is not vulnerable to reconnaissance techniques such as scanning. Although that may be true to force brute attacks, [I-D.ietf-opsec-ipv6-host-scanning] shows some techniques that may be employed to speed up and improve results in order to discover IPv6 address in a subnet. The use of virtual machines and SLACC aggravate this problem due the fact that they tend to use automatically-generated MAC address well known patterns.

To mitigate address-scanning attacks it is recommended to avoid using SLAAC and if used stable privacy-enhanced addresses [I-D.ietf-6man-stable-privacy-addresses] should be the method of address generation. Also, for manually assigned addresses try to avoid IID low-byte address (i.e. from 0 to 256), IPv4-based addresses and wordy addresses especially for infrastructure without a fully qualified domain name.

In spite of the use of manually assigned addresses is the preferred method for V6 DC, SLACC and DHCPv6 may be also used for some special reasons. However we recommend paying special attention to RA [RFC6104] and DHCP [I-D.gont-opsec-dhcpv6-shield] hijack attacks. In these kinds of attacks the attacker deploys rogue routers sending RA messages or rogue DHCP servers to inject bogus information and possibly to perform a man in the middle attack. In order to mitigate this problem it is necessary to apply some techniques in access

switches such as RA-Guard [RFC6105] at least.

Another topic that we would like to mention related to addressing is the use of ULAs. As we previously mentioned, although ULAs may be used to hide host from the outside world we do not recommend to rely on them as a security tool but better as a tool to make renumbering easier.

#### 4.3. Edge filtering

In order to avoid being used as a source of amplification attacks is it important to follow the rules of BCP38 on ingress filtering. At the same time it is important to filter-in on the network border all the unicast traffic and routing announcement that should not be routed in the Internet, commonly known as "bogus prefixes".

#### 4.4. Final Security Remarks

Finally, let us just emphasize the need for careful configuration of access control rules at the translation points. This latter one is specially sensitive in infrastructures at the dual-stack stage, as the translation points are potentially distributed, and when protocol translation is offered as an external service, since there can be operational mismatches.

#### 5. IANA Considerations

None.

#### 6. Acknowledgements

We would like to thank Tore Anderson, Wes George, Ray Hunter, Joel Jaeggli, Fred Baker, Lorenzo Colitti, Dan York, Carlos Martinez, Lee Howard, Alejandro Acosta, Alexis Munoz, Nicolas Fiumarelli, Santiago Aggio and Hans Velez for their questions, suggestions, reviews and comments.

#### 7. Informative References

[I-D.gont-opsec-dhcpv6-shield]  
Gont, F. and W. Liu, "DHCPv6-Shield: Protecting Against Rogue DHCPv6 Servers", draft-gont-opsec-dhcpv6-shield-01 (work in progress), October 2012.

[I-D.ietf-6man-flow-ecmp]

Carpenter, B. and S. Amante, "Using the IPv6 flow label for equal cost multipath routing and link aggregation in tunnels", draft-ietf-6man-flow-ecmp-05 (work in progress), July 2011.

[I-D.ietf-6man-stable-privacy-addresses]

Gont, F., "A method for Generating Stable Privacy-Enhanced Addresses with IPv6 Stateless Address Autoconfiguration (SLAAC)", draft-ietf-6man-stable-privacy-addresses-10 (work in progress), June 2013.

[I-D.ietf-opsec-ipv6-host-scanning]

Gont, F. and T. Chown, "Network Reconnaissance in IPv6 Networks", draft-ietf-opsec-ipv6-host-scanning-01 (work in progress), April 2013.

[I-D.ietf-opsec-v6]

Chittimaneni, K., Kaeo, M., and E. Vyncke, "Operational Security Considerations for IPv6 Networks", draft-ietf-opsec-v6-02 (work in progress), February 2013.

[I-D.ietf-v6ops-enterprise-incremental-ipv6]

Chittimaneni, K., Chown, T., Howard, L., Kuarsingh, V., Pouffary, Y., and E. Vyncke, "Enterprise IPv6 Deployment Guidelines", draft-ietf-v6ops-enterprise-incremental-ipv6-03 (work in progress), July 2013.

[I-D.ietf-v6ops-icp-guidance]

Carpenter, B. and S. Jiang, "IPv6 Guidance for Internet Content and Application Service Providers", draft-ietf-v6ops-icp-guidance-05 (work in progress), January 2013.

[I-D.sunq-v6ops-contents-transition]

Sun, Q., Liu, D., Zhao, Q., Liu, Q., Xie, C., Li, X., and J. Qin, "Rapid Transition of IPv4 contents to be IPv6-accessible", draft-sunq-v6ops-contents-transition-03 (work in progress), March 2012.

[RFC4193] Hinden, R. and B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, October 2005.

[RFC4732] Handley, M., Rescorla, E., and IAB, "Internet Denial-of-Service Considerations", RFC 4732, December 2006.

[RFC5375] Van de Velde, G., Popoviciu, C., Chown, T., Bonness, O., and C. Hahn, "IPv6 Unicast Address Assignment

Considerations", RFC 5375, December 2008.

- [RFC6052] Bao, C., Huitema, C., Bagnulo, M., Boucadair, M., and X. Li, "IPv6 Addressing of IPv4/IPv6 Translators", RFC 6052, October 2010.
- [RFC6104] Chown, T. and S. Venaas, "Rogue IPv6 Router Advertisement Problem Statement", RFC 6104, February 2011.
- [RFC6105] Levy-Abegnoli, E., Van de Velde, G., Popoviciu, C., and J. Mohacsi, "IPv6 Router Advertisement Guard", RFC 6105, February 2011.
- [RFC6145] Li, X., Bao, C., and F. Baker, "IP/ICMP Translation Algorithm", RFC 6145, April 2011.
- [RFC6164] Kohno, M., Nitzan, B., Bush, R., Matsuzaki, Y., Colitti, L., and T. Narten, "Using 127-Bit IPv6 Prefixes on Inter-Router Links", RFC 6164, April 2011.
- [RFC6177] Narten, T., Huston, G., and L. Roberts, "IPv6 Address Assignment to End Sites", BCP 157, RFC 6177, March 2011.
- [RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.
- [RFC6583] Gashinsky, I., Jaeggli, J., and W. Kumari, "Operational Neighbor Discovery Problems", RFC 6583, March 2012.
- [V6DCS] "The case for IPv6-only data centres", <[https://ripe64.ripe.net/presentations/67-20120417-RIPE64-The\\_Case\\_for\\_IPv6\\_Only\\_Data\\_Centres.pdf](https://ripe64.ripe.net/presentations/67-20120417-RIPE64-The_Case_for_IPv6_Only_Data_Centres.pdf)>.

#### Authors' Addresses

Diego R. Lopez  
Telefonica I+D  
Don Ramon de la Cruz, 84  
Madrid 28006  
Spain

Phone: +34 913 129 041  
Email: [diego@tid.es](mailto:diego@tid.es)

Zhonghua Chen  
China Telecom  
P.R.China

Phone:  
Email: 18918588897@189.cn

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

Phone: +1 408 330 4424  
Email: Tina.Tsou.Zouting@huawei.com

Cathy Zhou  
Huawei Technologies  
Bantian, Longgang District  
Shenzhen 518129  
P.R. China

Phone:  
Email: cathy.zhou@huawei.com

Arturo Servin  
LACNIC  
Rambla Republica de Mexico 6125  
Montevideo 11300  
Uruguay

Phone: +598 2604 2222  
Email: aservin@lacnic.net



IPv6 Operations Working Group (v6ops)  
Internet-Draft  
Intended status: Experimental  
Expires: April 30, 2015

O. Nakamura  
Keio Univ./WIDE Project  
H. Hazeyama  
NAIST / WIDE Project  
Y. Ueno  
Keio Univ./WIDE Project  
A. Kato  
Keio Univ. / WIDE Project  
October 27, 2014

A Special Purpose TLD to resolve IPv4 Address Literal on DNS64/NAT64  
environments  
draft-osamu-v6ops-ipv4-literal-in-url-02

## Abstract

In an IPv6-only environment with DNS64/NAT64 based translation service, there is no way to get access a URL whose domain name part includes an IPv4 address literal. This memo proposes a special purpose TLD so that the IPv4 address literal is accessible from such a DNS64/NAT64 environments.

## 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 April 30, 2015.

## Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

## 1. Introduction and Overview

When a host in an IPv6 only environment (an IPv6-only host) has to access an IPv4-only destination, a translator-based approach is a powerful tool. The translator-based approach is usually composed of a DNS64 server [RFC6147] and a stateful NAT64 translator [RFC6146]. The DNS64 server responds with a AAAA record of an IPv4 embedded IPv6 address with a certain IPv6 prefix assigned to the NAT64 translator, for example, the well known NAT64 prefix (64:ff9b::) or a global IPv6 prefix. The IPv6-only host sends an IPv6 packet, which is translated by the NAT64 box to an IPv4 packet. In this memo, an IPv4 embedded IPv6 address with a NAT64 prefix is described as ``Pref64::/n address''. The translation of responded IPv4 packet back into an IPv6 packet is also performed in the NAT64 translator.

The NAT64 with DNS64 approach works well for most destinations. But it does not work well when the DNS response packet resulted NXDOMAIN or SERVFAIL to the AAAA query, partly described in [RFC4074]. Resolutions of this case are out of scope of this memo.

It is legitimate to embed an IPv4 address literal in an URL such as follows:

`http://192.0.2.10/index.html`

In the environment described above, the destination is not accessible from an IPv6-only host. This problem has already been reported in [RFC6586] and others.

The reason why the destination specified by above notation cannot be



accessible is that no DNS lookup is performed, and no DNS64 service is able to tell a Pref64::/n address to the host. To perform DNS64/NAT64 translation against such an IPv4 address literal notation, some mechanism will be required.

This memo proposes a special-purpose TLD and defines behaviors of resolvers and of the authoritative servers to treat the special-purpose TLD. This memo also considers implementation strategy of .TLD and side effects of .TLD usages to the current communications on the Internet. The special-purpose TLD is denoted as .TLD which will be replaced with an actual TLD allocated by IANA.

The concept of .TLD is simple: All IPv4 address literal notations are rewritten to ``<ipv4-address-literal>.TLD'' on a host. As ``<ipv4-address-literal>.TLD'' is seemed to be a regular FQDN, ``<ipv4-address-literal>.TLD'' lets DNS64 servers resolve IPv4 address literal as a regular FQDN and translate the A record of ``<ipv4-address-literal>.TLD'' to a corresponding Pref64::/n address on each leaf network. For example, 192.0.2.10.TLD in DNS64/NAT64 environment would be translated to a Pref64::c000:020a. In an IPv4 environment, 192.0.2.10.TLD would be resolved just as an A record about 192.0.2.10.

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

## 2. Scope of this memo

This memo focuses only on smooth migration to an IPv6-only environment with the DNS64/NAT64 solution. Therefore, this memo focuses on only ``IPv4 address literal'' problem mentioned in [RFC6586].

The ``IPv6 address literal'' is out of scope of this memo, because an URL including IPv6 address literal can be accessible in IPv6-only networks and in dual stack networks. The solutions to keep IPv4-only hosts or IPv4-only applications in IPv6 only environment are out of scope on this memo.

## 3. A special-purpose TLD for IPv4 Address Literal

When the part of IPv4 address literal is written to form a pseudo FQDN and the pseudo FQDN is resolved as an IPv4 address, a DNS64 server can return a AAAA record with the specified IPv4 address that is mapped to an appropriate NAT64 prefix.

Once a AAAA record is obtained, the IPv6-only host can send IPv6 packets to the destination. IPv6 packets will be translated back via NAT64 translator in exactly the same as a regular IPv4-only destination.

### 3.1. .TLD Authoritative DNS server behavior

The authoritative DNS server of .TLD SHOULD be operated only for a special purpose.

1. If a DNS query asks ``<ipv4-address-literal>.TLD '', .TLD authoritative server MUST return ``<ipv4-address-literal>'' as the A record of ``<ipv4-address-literal>.TLD ''.
2. Otherwise, .TLD authoritative server MUST return NXDOMAIN.

### 3.2. DNS64 behaviors

When a DNS64 receives a query of <ipv4-address-literal>.TLD, it SHOULD issue a DNS query to one of the .TLD authoritative servers. The response from .TLD authoritative server will be either an A record of the issued <ipv4-address-literal> or NXDOMAIN. If the response contains an A record, the DNS64 MUST translate the IPv4 address in the A record to the AAAA record by Pref64::/n address according to [RFC6147].

Taking into account of scalability, the DNS64 WOULD cache the AAAA record of <ipv4-address-literal>.TLD in a certain interval. As one of possible ways to get more scalability, the DNS64 CLOUD have the function of .TLD authoritative server.

### 3.3. Client behaviors

#### 3.3.1. Case 1: manual type-writing

When a client (human) wants to access an IPv4 only server by IPv4 address literal in a DNS64/NAT64 network, he / she manually attaches .TLD to the IPv4 address of the IPv4 only server. When the network has DNS64/NAT64 function, the AAAA record, that is Pref64::/n address of the issued <ipv4-address-literal> , will be return.

The client COULD attach .TLD to the IPv4 address of the IPv4 only server in an IPv4 only network or a dual stack network. When the network situation is IPv4 only or dual stack, the A record of the issued <ipv4-address-literal>.TLD will be returned.

If the client uses FQDN or IPv6 address literal, he / she MUST NOT attach .TLD.

### 3.3.2. Case 2: device or application

A client (device or application), that has a name resolution function, SHOULD attach .TLD when the input value of getaddrinfo is an IPv4 address literal. For example, <ipv4-address-literal> SHOULD be rewritten to <ipv4-address-literal>.TLD. If the input value of getaddrinfo is not IPv4 address literal, the client MUST NOT attach .TLD.

Of course, the client CAN take self-synthesizing of mapped address mentioned in [RFC7050], or MAY combine .TLD method and [RFC7050] self-synthesizing method.

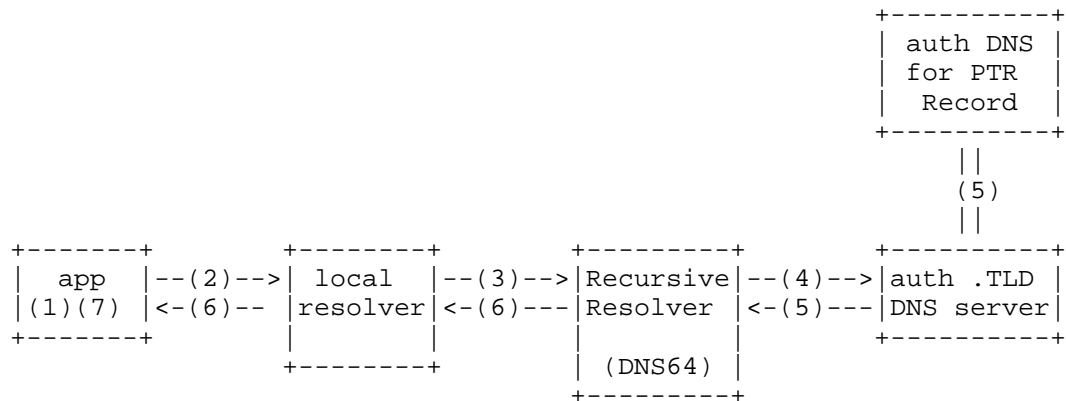
Some access authentication may not allow any external accesses until access authentication procedure is finished, and may use an IPv4 address literal on the redirected authentication web page. Taking into account such corner case, client WOULD check the reachability to the external network initially.

NOTE: migrating from IPv4 to IPv6, access authentication SHOULD avoid to use IPv4 address literal and SHOULD use FQDN for dual stack client or IPv6 only client.

### 3.4. DNS query flow

Figure 1 shows a DNS query flow on the .TLD.

1. An application on a client creates <ipv4-address-literal>.TLD.
2. The application inputs the query of AAAA or ANY about <ipv4-address-literal>.TLD. to its local resolver.
3. The local resolver forwards the query to a recursive resolver that would be a DNS64 server in DNS64/NAT64 environment.
4. The recursive resolver sends a recursive query of <ipv4-address-literal>.TLD.
5. .TLD authoritative server creates the A record of the issued <ipv4-address-literal>.TLD, and MAY check PTR record of the issued <ipv4-address-literal>. Then, .TLD authoritative server returns the DNS response to the recursive resolver.
6. When the recursive resolver has DNS64 function, it creates the AAAA record according to [RFC6147] and replies the AAAA record to the local resolver on the client. If the recursive resolver does not have DNS64 function, the recursive resolver returns the A record responded from .TLD authoritative server.
7. The application on the client gets the appropriate IP address (IPv4 address or Pref64::/n address), then creates an appropriate socket.



DNS Query Flow on .TLD

Figure 1

This solution would not require the modification of common shared libraries on any Operating Systems. The DNS implementations, SHOULD support .TLD. As the query flow mentioned above, .TLD authoritative server SHOULD be placed. The modification of NAT64 or DHCP are not required in this method.

### 3.5. Use cases

#### 3.5.1. Use case 1: manual type-writing

For example, consider living on an IPv6-only network with DNS64/NAT64, and receiving a message like ``please download a file foo.doc from a ftp server 192.0.2.10''. Usually, you may estimate the NAT64 prefix and calculate Pref64::/n address through [RFC7050] or [RFC7051]. Under the proposed mechanism on this memo, you can just type as follow;

```
% ftp 192.0.2.10.TLD
```

The packet would be transferred along with [RFC6384].

#### 3.5.2. Use case 2: browser plug-in

An IPv4 address literal is often used in URL for the lazy DNS operation, a temporary HTTP server or a hidden (private) server. Taking into account user convenience, a browser plug-in can be developed that it converts the <ipv4-address-literal> on the hostname

part of an URL to <ipv4-address-literal>.TLD. It may be suggested to turn this function on when the host is on IPv6-only network, however, it may not be easy to detect the situation of the network (IPv4 only, dual stack or DNS64/NAT64 environment). A sample of Google Chrome plug-in is attached in Appendix B

### 3.6. Recommendation

For usability in manual type-writing, the .TLD SHOULD be as short as possible, and SHOULD express the special purpose in the name space. ``.v4`` is recommended as a candidate of .TLD, because of the simplicity and the expression of IPv4.

## 4. Considerations

### 4.1. Attached the special-purpose TLD to a regular FQDN

Conceptually, the special-purpose TLD would be attached to only IPv4 address literals, however, the special-purpose TLD may be attached to a regular FQDN notation like ``foo.bar.com.TLD``. Such misuses SHOULD be avoided.

### 4.2. An embedded IP address literal in the content part of URL

In some case, <ipv4-address-literal> may be embedded into the content part of a URL, however, it may be difficult for users or browser plug-ins to recognize unambiguously that a string like <ipv4-address-literal> surely means some IPv4 address. From the point of view of IPv6 migration, embedded IP address literal in the content part of an URL MUST be avoided.

### 4.3. Prevention the leak of the special-purpose TLD

When .TLD is actually employed in the operation, .TLD may leak to the public DNS infrastructure including root DNS servers as seen in ``.local``. Therefore, once consensus is obtained, the relevant TLD SHOULD be delegated to a set of DNS servers.

Two possible DNS operation methods can be considered. One is to delegate the TLD to AS112 servers [as112-servers]. When one of the AS112 servers received a query with .TLD, it returns with NXDOMAIN.

The other possible DNS operation is to deploy a set of special purpose DNS servers which accept queries with .TLD and synthesize an A record corresponding to the IPv4 address in the QNAME when it is a legitimate IPv4 address. Otherwise, NXDOMAIN MUST be returned.

#### 4.4. Possibility to break connections with Apache VirtualHost concept

Changing the URL (swapping the DNS name or adding in a Pref64) frequently breaks the connections since the application is aware of the name it expects, and connecting correctly to the correct IP address is not sufficient, the name must also be the same in many cases.

For example, many websites use the Apache VirtualHost concept. When a web site that changes contents along with accessed IP address family like `http://www.kame.net/` or `http://dual.tlund.se/`, and if some client accesses such web site by `<ipv4-address-literal>.TLD` instead of FQDN, the VirtualHost may not work as intended.

Therefore, such web site, that uses the Apache VirtualHost concept, SHOULD NOT use `<ipv4-address-literal>` in URL and SHOULD use appropriate FQDN.

#### 4.5. Inaffinity with HTTP/HTTPS Cookie

This solution may not work with HTTP/HTTPS cookie. We should also consider the HTTP security considerations for the cases where someone puts one of the names into a URL. For example, consider `http://192.0.2.10.TLD/` to an origin that sets a cookie on the domain `"*.10.TLD"`.

There are likely already plenty of ways to do the same thing out there, so this may not be a major issue.

#### 4.6. TLD alternatives

In Section 3.6, we propose `.v4` as the TLD, and comparisons with other candidates are discussed as follows.

##### 4.6.1. `.v4.arpa`

```v4.arpa``` may be a candidate of `.TLD` that does not require new TLD, however, it may be confused with [RFC7050] ```ipv4only.arpa```, and the length (8 characters) of ```v4.arpa``` is bit longer than the length (3 characters) of ```v4``` for type-writing usages.

##### 4.6.2. `.host`

```.host``` has already been assigned as one of the new gTLDs, and not considered a candidate here unless the authority of `.host` offers 256 (or 356 -- see discussion in Section 4.6.3) delegations to this purpose.

#### 4.6.3. TLD less delegation

When it is feasible to "delegate" 256 TLDs (from ".0" through ".255") or 366 TLDs (".00", ".000", and others are added) for this particular purpose, it is possible to implement the functionality described in this memo without assigning a particular .TLD. It contributes 256 (or 356) extra TLDs in the Root zone.

It is known that DNS queries with such TLDs have been observed, and this delegation may interfere with undocumented usage of such TLDs.

If such 256 (or 366) delegations is suitable, bogus such queries to the root servers will be redirected to the DNS server described in Section 5.

#### 4.7. Usages of IPv6 address literal

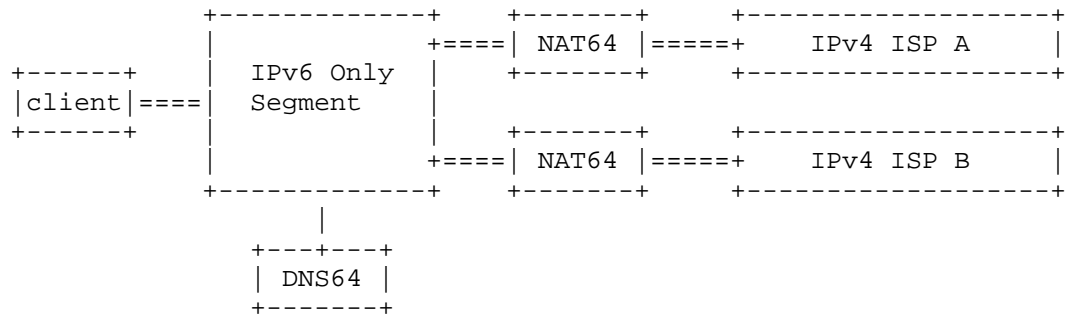
The special-purpose TLD may be applied to IPv6 address cases in same ways, however, such notation is not required in dual stack / IPv6-only environment, generally.

#### 4.8. RFC7050 ipv4only.arpa

[RFC7050] defines a method to estimate a NAT64 prefix by querying Well-Known IPv4-only Name ``ipv4only.arpa''. [RFC7050] does not cover several situations. .TLD method is aimed to solve such situations as follows:

##### 4.8.1. Multiple NAT64 prefixes for load balancing

One of situations is multihoming, illustrated in Figure 2. In this situation, the NAT64 prefix estimated by [RFC7050] method may be different from the one that the operator intends.

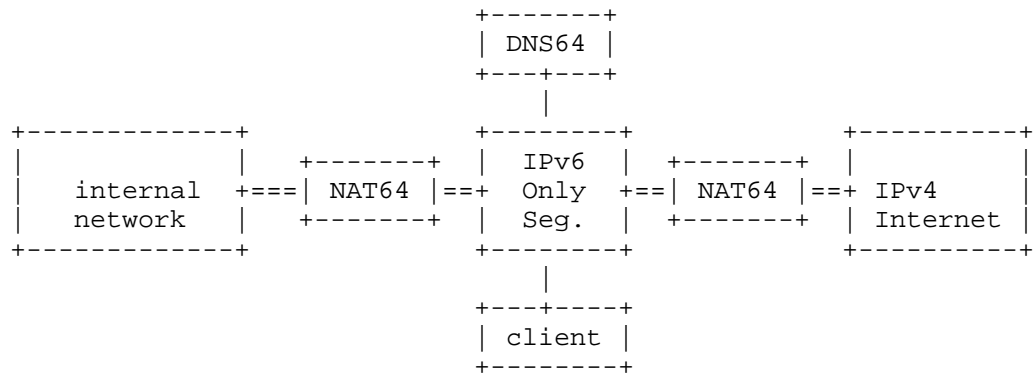


Situation A : multiple NAT64 prefixes for optimizing routes on multihoming

Figure 2

#### 4.8.2. Multiple NAT64 prefixes for external / internal IPv4 only networks

Another situation is where multiple NAT64 prefixes are operated for accessing the external IPv4 Internet and an internal private IPv4 only network from an internal IPv6 only network. Figure 3 draws this situation. In this situation, the NAT64 prefix estimated by [RFC7050] method could not be reached to the internal IPv4 only network.



Situation B : multiple NAT64 prefixes for internal / external

Figure 3



#### 4.8.3. Difficulty of conversion from octet expression to hex expression by human type-writing

As the initial motivation of this memo, IPv4 address literal is often used for a personal / private server that is not registered in DNS record because of lazy operation, temporal usage, or the intention to hide from DNS query scans. ``ipv4only.arpa`` solution can be available to synthesize the Pref64::/n address for the private server, however, the owner of the private server has to convert the octet expression of the IPv4 address on his/her private server to the hex expression by manual. Usually, conversion from octet expression to hex expression by manual is difficult or tiresome operation.

### 5. Implementation Strategy

It is suggested to implement the .TLD rewriting as in the following order:

1. Define .TLD  
Once the community agrees to accept the rewriting scheme described in this memo, it must fix the .TLD to be used. The .TLD WOULD require the update of [RFC6761].
2. .TLD delegation  
DNS queries with .TLD can leak to the DNS of the global Internet, it is highly suggested to delegate .TLD to a set of authoritative DNS servers as discussed in Section 4.3.
3. DNS64 modification  
DNS64 implementation is suggested to modify to respond corresponding AAAA record to a query with .TLD. This process can be done in parallel to the step 2 above.
4. Start using .TLD rewriting  
After, at least the step 2 is completed, the TLD rewriting may be used in manually described in Section 3.5.1 or automatically by browser plugins described in Section 3.5.2. While further discussions and observation is required, the use of an URL in IPv4 literal embedded might be discouraged. Instead, the use of .TLD notation as a legitimate URL might be encouraged even in the server side.

### 6. Security Considerations

The recommendation contains security considerations related to DNS. The special purpose DNS servers of this memo only treats the IPv4 address literal with .TLD. Therefore, the special DNS MAY use self-signed / authorized key for DNS responses.

When a client is to access an URL with IPv4 literal address embedded,

it triggers a DNS query, and the query may be sent over the Internet to the nearest authoritative .TLD DNS server. It may break the confidentiality against the DNS service.

TBD

## 7. IANA Considerations

This memo calls for ``.v4`` as the special-purpose TLD to the IANA registry.

## 8. Acknowledgments

Authors thank to WIDE Project members for their active discussion, implementations, and evaluations. Especially, we thank to Atsushi ONOE for the revision of this solution, Hirochika ASAI for the contribution of the prototype implementation of the special purpose authoritative DNS, and Hirotaka NAKAJIMA for the contribution of the Google chrome plug-in. We also thank to Yoshiaki KITAGUCHI, Yu-ya KAWAKAMI and others who evaluated our proof of concept special purpose DNS (.v4.wide.ad.jp) and the Google Chrome plugin-in at JANOG34 DNS64/NAT64 experiment networks. Teeme Savolainen, Cameron Byrne, Dan Wing, Erik Nygren gave us various considerations on the actual operation of .TLD.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4074] Morishita, Y. and T. Jinmei, "Common Misbehavior Against DNS Queries for IPv6 Addresses", RFC 4074, May 2005.
- [RFC6146] Bagnulo, M., Matthews, P., and I. van Beijnum, "Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers", RFC 6146, April 2011.
- [RFC6147] Bagnulo, M., Sullivan, A., Matthews, P., and I. van Beijnum, "DNS64: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers", RFC 6147, April 2011.
- [RFC6384] van Beijnum, I., "An FTP Application Layer Gateway (ALG)

for IPv6-to-IPv4 Translation", RFC 6384, October 2011.

- [RFC6586] Arkko, J. and A. Keranen, "Experiences from an IPv6-Only Network", RFC 6586, April 2012.
- [RFC6761] Cheshire, S. and M. Krochmal, "Special-Use Domain Names", RFC 6761, February 2013.
- [RFC7050] Savolainen, T., Korhonen, J., and D. Wing, "Discovery of the IPv6 Prefix Used for IPv6 Address Synthesis", RFC 7050, November 2013.
- [RFC7051] Korhonen, J. and T. Savolainen, "Analysis of Solution Proposals for Hosts to Learn NAT64 Prefix", RFC 7051, November 2013.

## 9.2. Informative References

- [as112-servers] AS112 Project, "AS112 Project", October 2009, <<https://www.as112.net/>>.

## Appendix A. A Test Server of the special TLD

We run a prototype implementation of the special-purpose DNS server in the WIDE backbone (AS 2500). We use ``v4.wide.ad.jp`` as .TLD.

## Appendix B. Sample extension for Google Chrome

We developed a sample plug-in code for Google Chrome ``IPv4 Address Literal Appender`` that automatically converts <ipv4-address-literal> in URL to <ipv4-address-literal>.TLD. The .TLD can be customized in the option. The ``IPv4 Address Literal Appender`` is freely available in Google Chrome Web Store, and also in github <https://github.com/nunnun/nat64-v4-literal-extension>.

```
var wr = chrome.webRequest;

var v4Suffix = ".TLD";
var ipAddrRegex = /^(\\d|[01]?\\d\\d|2[0-4]\\d|25[0-5])\\. (\\d|[01]?\\d\\d|2[0-4]\\d|25[0-5])\\. (\\d|[01]?\\d\\d|2[0-4]\\d|25[0-5])\\. (\\d|[01]?\\d\\d|2[0-4]\\d|25[0-5])$/;

function onBeforeRequest(details) {
  var tmpuri = new URI(details.url);
  var tmphost = tmpuri.host();
  var finalUri = '';
  tmphost.replace(ipAddrRegex,function(str,p1,p2,p3,p4,offset,s){
    finalUri=tmpuri.host(p1+"."+p2+"."+p3+"."+p4+v4Suffix).toString();
  });
  if('' != finalUri) {
    console.log(finalUri);
    return {redirectUrl: finalUri};
  }
};

wr.onBeforeRequest.addListener(onBeforeRequest,{urls: ["https://*/**",
"http://*/**", "ftp://*/**"]}, ["blocking"]);
```

#### Authors' Addresses

Osamu Nakamura  
Keio Univ./WIDE Project  
5322 Endo  
Fujisawa, Kanagawa 252-0882  
JP

Phone: +81 466 49 1100  
Email: osamu@wide.ad.jp

Hiroaki Hazeyama  
NAIST / WIDE Project  
8916-5 Takayama  
Ikoma, Nara 630-0192  
JP

Phone: +81 743 72 5111  
Email: hiroa-ha@is.naist.jp

Yukito Ueno  
Keio Univ./WIDE Project  
5322 Endo  
Fujisawa, Kanagawa 252-0882  
JP

Phone: +81 466 49 1100  
Email: eden@sfc.wide.ad.jp

Akira Kato  
Keio Univ. / WIDE Project  
Graduate School of Media Design, 4-1-1 Hiyoshi  
Kohoku, Yokohama 223-8526  
JP

Phone: +81 45 564 2490  
Email: kato@wide.ad.jp



v6ops  
Internet-Draft  
Intended status: Informational  
Expires: January 16, 2014

A. Servin  
LACNIC  
M. Rocha  
Redes de Interconexion  
Universitaria Asoc. Civil (ARIU)  
July 15, 2013

Monitoring Dual Stack/IPv6-only Networks and Services  
draft-servin-v6ops-monitor-ds-ipv6-01

Abstract

This document describes a set of recommendations and guidelines to help operators to monitor dual stack and IPv6-only networks. The document describes how to monitor these networks using SNMP, Flow Analyzers and other means.

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 January 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as

described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Network Monitoring . . . . .	3
2.1. Transport vs. Data . . . . .	3
2.2. Simple Network Management Protocol . . . . .	4
2.3. Flow Analyzers . . . . .	4
2.3.1. Netflow . . . . .	4
2.3.2. Sflow . . . . .	5
2.3.3. IPFIX . . . . .	5
2.3.4. Network/Traffic Analyzers . . . . .	5
2.4. Command line interface tools . . . . .	5
2.5. Software Defined Networks . . . . .	5
3. Addressing . . . . .	6
4. Application Monitoring . . . . .	6
4.1. Services . . . . .	6
4.2. FQDN as connection discriminator . . . . .	6
5. IPv6-Only Networks . . . . .	7
6. Operational Challenges . . . . .	7
7. Security Considerations . . . . .	8
8. IANA Considerations . . . . .	8
9. Acknowledgements . . . . .	8
10. Informative References . . . . .	8
Authors' Addresses . . . . .	10



## 1. Introduction

Network and services monitoring become more important as we rely more on them for our critical operations. Depending of the complexity of our monitor solution we would be able to have more control and information from our network and services. Among other things, a good monitor solution allows to::

- o Detect and avoid network incidents
- o Determine which actions may solve a network incident
- o Execute recovery and contingency plans

All these make sense when we monitor our network responsibly trying to cover all the variables. In the context of this memo, it means that we should monitor our services and networks running IPv6 as we have/had done in the IPv4 world..

There are many documents and guides explaining how to deploy IPv6 networks and services but there are so few that describe in detail how to monitor them. This document tries to encompass a set of recommendations and guidelines to help network and system administrators to monitor dual-stack/IPv6-only network and services.

## 2. Network Monitoring

In this section we describe SNMP and IPFIX as protocols able to manage IP devices and to monitor a variety of data from dual stack and IPv6-only networks. We also discuss traffic analyzers as other tools to monitor IP networks.

### 2.1. Transport vs. Data

It is important to understand the difference between IPv6 Transport vs. IPv6 data. In other words, protocols for monitor network infrastructure such as SNMP or IPFIX can send IPv6 collected data (e.g. the count of forwarded packets of an interface, flow information) using either IPv4 or IPv6 transport.

It is important to note that some node implementations would only send data (either IPv4 or IPv6) over IPv4 networks. Nevertheless these are implementation limitations not related to the monitoring protocol.

## 2.2. Simple Network Management Protocol

Simple Network Management Protocol (SNMP) defines the protocol suite to monitor and manage IP networks. SNMP works over UDP that allows it to work over IPv4 or IPv6 networks. However, the definitions that allow SNMP to collect data from IP devices known as "Management Information Base" (MIB) had to be modified from the original specifications. The most used versions of SNMP are Version 1 and Version 2 [RFC1441]. Version 3 is defined in [RFC3411].

SNMP MIB was defined in [RFC1156] and extended by [RFC1158]. Later it was modified by [RFC1213] in 1990 and in 1996 deprecated by RFCs [RFC2011], [RFC2012] and [RFC2013] that separated the MIB in IP, TCP and UDP. However all these modifications did not consider IPv6 yet. It was until [RFC2465] and [RFC2466] that MIB definitions were specified for IPv6 and ICMPv6. These RFCs described a dissociated definition for IPv4 and IPv6. The last MIB definitions came in 2006 when [RFC4292] (IP-Forwarding) and [RFC4293] (IP-MIB) defined a unified set of managed objects independent of the IP version.

Today there are many agent and collector implementations that support [RFC4292] and [RFC4293]. Nevertheless not all of them support them over IPv6 transport and IPv4 has to be used.

## 2.3. Flow Analyzers

Knowing the packet count that goes in and out from an interface it is very important but many times is not enough to detect faults or to get more detailed traffic information about the network. Netflow and IP Flow Information Export (IPFIX) [RFC5101] and [RFC5102] are protocols that monitor the IP flows passing through network devices. An IP flow is a sequence of packets identified by a common set of attributes such as IP Source Address, IP Destination Address, Source Port, Destination Port, Layer 3 protocol type, Class of service, etc.

### 2.3.1. Netflow

Netflow is a protocol developed by Cisco Systems and version 9 is described in the informational [RFC3594]. Other vendors have adopted equivalent technology such as Jflow (Juniper Networks), Cflowd (Alcatel-Lucent) and SFlow (sFlow.org consortium).

Netflow defines nine versions from which version 5 is the most common and only versions 9 and 10 support IPv6. Versions 9 and 10 are commonly known as the base of IPFIX. Although Netflow version 9 supports the collection of IPv6 flows, not all implementations of agents and collectors support IPv6 transport and IPv4 has to be used.

### 2.3.2. Sflow

Sflow is defined in [RFC3176] and it is very similar to netflow and IPFIX. It differs basically in the method to collect flow information. In the case of Sflow, it uses statistical packet-based sampling of switched flows and time-based sampling. The Sflow version described in [RFC3176] supports IPv4 and IPv6 address families.

### 2.3.3. IPFIX

IPFIX architecture and message format is defined in RFC5101 [RFC5101] and RFC5102 [RFC5102] defines its information model. From the operational standpoint of this document IPFIX and Netflow v9 are not very different and there is not much more to say besides that IPFIX as a relatively new protocol has not been widely implemented. For this reason finding an implementation supporting IPv6 transport may be hard to find.

### 2.3.4. Network/Traffic Analyzers

Besides SNMP and Flow analyzers IPv6 can be monitored using a variety of network/traffic analyzers. These devices come in a variety of flavors and some are open source or free and can be installed in commodity hardware, some other are expensive and run on specialized equipment. Commonly they are installed using promiscuous port that mirror all the network traffic or they are installed somewhere in the network where they can inspect most of the traffic.

Network/traffic analyzers are a quick way to inspect IPv6 traffic, however they may have scalability and privacy issues which make them unsuitable for large networks.

## 2.4. Command line interface tools

When SNMP and flow tools are not available in the network device and traffic analyzers are not suitable as a long term solution it may be possible to use in-house development or other tools to access networks devices and parse command line instructions that monitor IPv6 traffic. This solution could be used as well in IPv6 only networks when the device implementation does not support IPv6 transit to deliver monitoring data.

## 2.5. Software Defined Networks

TBD, In this section we will discuss the use of Software Defined Network (SDN) for the purposes of gathering data from the network.

### 3. Addressing

TBD. In this section we will discuss the implications to use of link-local, ULAs and Global Unicast Addresses for the purpose of monitor network infrastructure.

### 4. Application Monitoring

Beyond the traffic that goes through the network, network operators require to monitor other services such HTTP servers, email infrastructure, DNS, sensors, etc.

#### 4.1. Services

Besides network information, network operators require to know other variables that could affect the good operation of the network. Dual stack networks pose an important challenge to network and system administrators. In principle we are talking about two different networks that may have different paths and users may perceive a difference in quality. Furthermore, thanks to Happy Eye Balls RFC6555 [RFC6555] that improves the user experience, service operators may have no idea to which protocol users are connected. This impose the need to monitor two networks and two set of services such as HTTP servers, email infrastructure, DNS, etc. to guarantee the service expectations from users.

In order to monitor service uptime and performance, it is common to use service probes that frequently poll a specific service to verify its reachability. Most of the time this probes are configured to access a service using a Fully Qualified Domain Name (FQDN) but sometimes literals are used as well.

To monitor services using FQDNs with A and AAAA records network/system administrator must be aware that they do not have a guarantee that the probe is using IPv4 or IPv6 transport unless is forced to do so. Some tools provide configuration or execution flags to force the use of IPv4 or IPv6 transport. To guarantee a reliable monitoring strategy, we recommend using those flags to set up two monitor instances, one for each address family. Needless to say that in case of using literals instead of FQDNs, a new service monitor instance using an IPv6 address must be added.

#### 4.2. FQDN as connection discriminator

We mentioned that one possible solution to discriminate between IPv4 and IPv6 services is to use some of the flags provided by the monitoring tool to force a connection either in IPv4 or IPv6.

Depending of the tool used, this option may not be always available. To address this restriction it is possible to use a special FQDN with only an A record to force an IPv4 connection and a different FQDN with only an AAAA record for IPv6.

For example suppose that the main organization website has the name `www.example.com`. The name `www.example.com` would have A and AAAA records as normally, however it would also contain an A record of the form `www.v4-test.example.com` pointing to its IPv4 address and an AAAA record `www.v6-test.example.com` point to the IPv6 address of the service. Other variants may be `www.v6.example.com`, `www-v4.example.com`, etc. As these FQDNs are meant to be only internally the selection of which to use is left to the network operator.

Bear in mind that using this alternative may introduce an extra overhead related to DNS management and should be used only when strictly necessary.

## 5. IPv6-Only Networks

The critical path to monitor IPv6 data on dual-stack networks is the device support of the IPv6 only MIBs ([RFC2465], [RFC2466], [RFC2452] and [RFC2454]), the unified MIBs ([RFC4293], [RFC4022], [RFC4113] and [RFC4292] or flow tools as Netflow 9 or IPFIX. As long as these protocols are supported, the device can be monitor using IPv4 or IPv6 transport. However, in IPv6-only networks supporting IPv6 data monitoring is not enough. In order to work it is critical for the device or collector to support the delivery or polling data using IPv6 transport.

For SNMP data there are a variety of agents and collectors that support IPv6 MIBs (IPv6 and Protocol Independent) using IPv6 transport. Nevertheless still exist devices that do not support neither IPv6 MIBs nor IPv6 transport of monitoring data.

With respect of flow tools, the authors of this document are aware of only a few implementations that support IPv6 transport.

## 6. Operational Challenges

Even though the end of IPv4 is near, there are still many network devices that cannot provide any type of IPv6 monitor data. In other cases the device can provide some sort of data through command line interfaces or in the best scenario through out the old IPv6 MIBs and using only IPv4 transit for delivery.

Still many network devices do not support to collect or send data related to IPv6. Also, some implementations are not widely tested and they may not support IPv6 monitoring correctly. For example, there were in the past cases where network devices did not correctly reported data collected from interface counters as they only counted packets that were process switched. Eventually this bug was fixed to include hardware-processed packets. It will still possible to find more of these types of bugs whilst IPv6 support mature. For that reason we recommend to network operators to always double check the IPv6 data retrieved from SNMP agents and interface counters at least for a short period of time. As the IPv6 support moves forward and matures, this practice would be less important in the future.

## 7. Security Considerations

From the security stand point, monitoring IPv4, IPv6 or Dual Stack networks is no different and the same preventions have to be taken. In order to protect SNMP agents, Network Monitoring Systems (NMS), flow collectors, network analyzers, etc. operators are advised to use a variety of methods such as access list, separate networks for management and monitoring, avoid the use of clear text access, etc.

## 8. IANA Considerations

None.

## 9. Acknowledgements

We would like to thank Humberto Galiza, Alejandro Acosta, Sofia Silva, Diego Lopez, Ariel Weher, and Christian O'Flaherty for their questions, suggestions, reviews and comments. Also we would like to thank the LACNOG community for the informal comments that gave us during the meetings.

## 10. Informative References

- [RFC1156] McCloghrie, K. and M. Rose, "Management Information Base for network management of TCP/IP-based internets", RFC 1156, May 1990.
- [RFC1158] Rose, M., "Management Information Base for network management of TCP/IP-based internets: MIB-II", RFC 1158, May 1990.

- [RFC1213] McCloghrie, K. and M. Rose, "Management Information Base for Network Management of TCP/IP-based internets:MIB-II", STD 17, RFC 1213, March 1991.
- [RFC1441] Case, J., McCloghrie, K., Rose, M., and S. Waldbusser, "Introduction to version 2 of the Internet-standard Network Management Framework", RFC 1441, April 1993.
- [RFC2011] McCloghrie, K., "SNMPv2 Management Information Base for the Internet Protocol using SMIV2", RFC 2011, November 1996.
- [RFC2012] McCloghrie, K., "SNMPv2 Management Information Base for the Transmission Control Protocol using SMIV2", RFC 2012, November 1996.
- [RFC2013] McCloghrie, K., "SNMPv2 Management Information Base for the User Datagram Protocol using SMIV2", RFC 2013, November 1996.
- [RFC2452] Daniele, M., "IP Version 6 Management Information Base for the Transmission Control Protocol", RFC 2452, December 1998.
- [RFC2454] Daniele, M., "IP Version 6 Management Information Base for the User Datagram Protocol", RFC 2454, December 1998.
- [RFC2465] Haskin, D. and S. Onishi, "Management Information Base for IP Version 6: Textual Conventions and General Group", RFC 2465, December 1998.
- [RFC2466] Haskin, D. and S. Onishi, "Management Information Base for IP Version 6: ICMPv6 Group", RFC 2466, December 1998.
- [RFC3176] Phaal, P., Panchen, S., and N. McKee, "InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks", RFC 3176, September 2001.
- [RFC3411] Harrington, D., Presuhn, R., and B. Wijnen, "An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks", STD 62, RFC 3411, December 2002.
- [RFC3594] Duffy, P., "PacketCable Security Ticket Control Sub-Option for the DHCP CableLabs Client Configuration (CCC) Option", RFC 3594, September 2003.
- [RFC4022] Raghunarayan, R., "Management Information Base for the

Transmission Control Protocol (TCP)", RFC 4022,  
March 2005.

- [RFC4113] Fenner, B. and J. Flick, "Management Information Base for the User Datagram Protocol (UDP)", RFC 4113, June 2005.
- [RFC4292] Haberman, B., "IP Forwarding Table MIB", RFC 4292, April 2006.
- [RFC4293] Routhier, S., "Management Information Base for the Internet Protocol (IP)", RFC 4293, April 2006.
- [RFC5101] Claise, B., "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information", RFC 5101, January 2008.
- [RFC5102] Quittek, J., Bryant, S., Claise, B., Aitken, P., and J. Meyer, "Information Model for IP Flow Information Export", RFC 5102, January 2008.
- [RFC6555] Wing, D. and A. Yourtchenko, "Happy Eyeballs: Success with Dual-Stack Hosts", RFC 6555, April 2012.

#### Authors' Addresses

Arturo Servin  
LACNIC  
Rambla Republica de Mexico 6125  
Montevideo 11300  
Uruguay

Phone: +598 2604 2222  
Email: aservin@lacnic.net

Mariela Rocha  
Redes de Interconexion Universitaria Asoc. Civil (ARIU)  
Maipu 645 - 4to Piso  
Buenos Aires  
Argentina

Email: mrocha@riu.edu.ar





Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: June 6, 2014

J. Jaeggli  
Zynga  
L. Colitti  
W. Kumari  
Google  
E. Vyncke  
Cisco  
M. Kaeo  
Double Shot Security  
T. Taylor, Ed.  
Huawei Technologies  
December 3, 2013

Why Operators Filter Fragments and What It Implies  
draft-taylor-v6ops-fragdrop-02

Abstract

This memo was written to make application developers and network operators aware of the significant possibility that IPv6 packets containing fragmentation extension headers may fail to reach their destination. Some protocol or application assumptions about the ability to use messages larger than a single packet may accordingly not be supportable in all networks or circumstances.

This memo provides observational evidence for the dropping of IPv6 fragments along a significant number of paths, explores the operational impact of fragmentation and the reasons and scenarios where drops occur, and considers the effect of fragment drops on applications where fragmentation is known to occur, particularly including DNS.

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 June 6, 2014.

#### Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction . . . . .	2
2. Observations and Rationale . . . . .	3
2.1. Possible Causes . . . . .	3
2.1.1. Stateful inspection . . . . .	4
2.1.2. Stateless ACLs . . . . .	4
2.1.3. Performance considerations . . . . .	4
2.1.4. Other considerations . . . . .	4
2.1.5. Conclusions . . . . .	5
2.2. Impact on Applications . . . . .	5
3. Acknowledgements . . . . .	5
4. IANA Considerations . . . . .	5
5. Security Considerations . . . . .	6
6. Informative References . . . . .	6
Authors' Addresses . . . . .	6

#### 1. Introduction

Measurements of whether Internet Service Providers and edge networks deliver IPv6 fragments to their destination reveal that for IPv6 in particular, fragments are being dropped along a substantial number of paths. The filtering of IPv6 datagrams with fragmentation headers is presumed to be a non-issue in the core of the Internet, where fragments are routed just like any other IPv6 datagram. However, fragmentation can create operational issues at the edges of the Internet that may lead to administratively imposed filtering or inadvertent failure to deliver the fragment to the end-system or application.

Section 2 begins with some observations on how often IPv6 fragment loss occurs in practice. We go on to look at the operational reasons for filtering fragments, a key aspect of which is the limitations they expose in the application of security policy, at resource bottlenecks and in forwarding decisions. Section 2.2 then looks at the impact on key applications, particularly DNS.

In the longer run, as network operators gain a better understanding of the risks and non-risks of fragmentation and as middlebox, customer premise equipment (CPE), and host implementations improve, we believe that some incidence of fragment dropping currently required will diminish. Some of the justifications for filtering will persist in the long-term, and application developers and network operators must remain aware of the implications.

This document deliberately refrains from discussing possible responses to the problem posed by the dropping of IPv6 fragments. Such a discussion will quickly turn up a number of possibilities, application-specific or more general; but the amount of time needed to specify and deploy a given resolution will be a major constraint in choosing amongst them. In any event, that discussion is likely to proceed in multiple directions, occur in different areas and is therefore considered beyond the scope of this memo.

## 2. Observations and Rationale

[Blackhole] is a good public reference for some empirical data on IPv6 fragment filtering. It describes experiments run to determine the incidence and location of ICMP Packet Too Big and fragment filtering. The authors used fragmented DNS packets to determine the latter, setting the servers to an IPv6 minimum of 1280 bytes to avoid any PMTU issues. The tests found for IPv6 that filtering appeared to be occurring on some 10% of the tested paths. The filtering appeared to be located at the edge (enterprise and customer networks) rather than in the core.

### 2.1. Possible Causes

Why does such filtering happen? One cause is non-conforming implementations in CPE and low-end routers. Some network managers filter fragments on principle, thinking this is an easier way to deter realizable attacks utilizing IPv6 fragments without thinking of other network impacts, similar to the practice of filtering ICMP Packet Too Big. Both implementations and management should improve over time, reducing the problem somewhat.

Some filtering and dropping of fragments is known to be done for hardware, performance, or topological considerations.

#### 2.1.1. Stateful inspection

Stateful inspection devices or destination hosts can readily experience resource exhaustion if they are flooded with fragments that are not followed in a timely manner by the remaining fragments of the original datagram. Holding fragments for reassembly even on end-system firewalls can readily result in an effective denial of service by memory and CPU exhaustion even if techniques, such as virtual re-assembly exist.

#### 2.1.2. Stateless ACLs

Stateless ACLs at layer 4 and up may be difficult to apply to fragments other than the first one in which enough of the upper layer header is present. As [Attacks] demonstrates, inconsistencies in reassembly logic between middleboxes or CPEs and hosts can cause fragments to be wrongfully discarded, or can allow exploits to pass undetected through middleboxes. Stateless load balancing schemes may hash fragmented datagrams from the same flow to different paths because the 5-tuple may be available on only the initial fragment. While rehashing has the possibility of reordering packets in ISP cores it is not disastrous. However, in front of a stateful inspection device, load balancer tier, or anycast service instance, where headers other than the L3 header -- for example, the L4 header, interface index (for traffic already rehashed onto different paths), DS fields -- are considered as part of the hash, rehashing may result in the fragments being delivered to different end-systems

#### 2.1.3. Performance considerations

Leaving aside these incentives towards fragment dropping, other considerations may weigh on the operator's mind. One example cited on the NANOG list was that of a router where fragment processing was done by the control plane processor rather than in the forwarding plane hardware, with a consequent hit on performance.

#### 2.1.4. Other considerations

Another incentive toward dropping of fragments is the disproportionate number of software errors still being encountered in fragment processing. Since this code is exercised less frequently than the rest of the stack, bugs remain longer in the code before they are detected. Some of these software errors can introduce vulnerabilities subject to exploitation. It is common practice [RFC6192] to recommend that control-plane ACLs protecting routers and network devices be configured to drop all fragments.

### 2.1.5. Conclusions

Operators weigh the risks associated with each of the considerations just enumerated, and come up with the most suitable policy for their circumstances. It is likely that at least some operators will find it desirable to drop fragments in at least some cases.

The IETF and operators can help this effort by identifying specific classes of fragments that do not represent legitimate use cases and hence should always be dropped. Examples of this work are given by [RFC6946] and [I-D.ietf-6man-oversized-header-chain]. The problem of inconsistent implementations may also be mitigated by providing further advice on the more difficult points. However, some cases will remain where legitimate fragments are discarded for legitimate reasons. The potential problems these cases pose for applications is our next topic.

### 2.2. Impact on Applications

Some applications can live without fragmentation, some cannot. UDP DNS is one application that has the potential to be impacted when fragment dropping occurs. EDNS0 extensions [RFC2671] allow for responses in UDP PDUs that are greater than 512 bytes. Particularly with DNSSEC [RFC4033], responses may be larger than the link MTU and fragmentation would therefore occur at the sending host in order to respond using UDP. The current choices open to the operators of DNS servers in this situation are to defer deployment of DNSSEC, fragment responses, or use TCP if there are cases where the rreset would be expected to exceed the MTU. The use of fallback to TCP will impose a major resource and performance hit and increases vulnerability to denial of service attacks.

Other applications, such as the Network File System, NFS, are also known to fragment large UDP packets for datagrams larger than the MTU. NFS is most often restricted to the internal networks of organizations. In general, managing NFS connectivity should not be impacted by decisions managing fragment drops at network borders or end-systems.

### 3. Acknowledgements

The authors of this document would like to thank the RIPE Atlas project and NLNetlabs whose conclusions ignited this document.

### 4. IANA Considerations

This memo includes no request to IANA.

## 5. Security Considerations

The potential for denial of service attacks, as well as limitations inherent in upper-layer filtering when dealing with non-initial fragments are significant issues under consideration by operators and end-users filtering fragments. This document does not offer alternative solutions to that problem, it does describe the impact of those filtering practices.

## 6. Informative References

- [Attacks] Atlasis, A., "Attacking IPv6 Implementation Using Fragmentation", March 2012.
- [http://media.blackhat.com/bh-eu-12/Atlasis/bh-eu-12-Atlasis-Attacking\\_IPv6-WP.pdf](http://media.blackhat.com/bh-eu-12/Atlasis/bh-eu-12-Atlasis-Attacking_IPv6-WP.pdf)
- [Blackhole] de Boer, M. and J. Bosma, "Discovering Path MTU black holes on the Internet using RIPE Atlas", July 2012.
- <http://www.nlnetlabs.nl/downloads/publications/pmtu-black-holes-msc-thesis.pdf>
- [I-D.ietf-6man-oversized-header-chain] Gont, F., Manral, V., and R. Bonica, "Implications of Oversized IPv6 Header Chains", draft-ietf-6man-oversized-header-chain-08 (work in progress), October 2013.
- [RFC2671] Vixie, P., "Extension Mechanisms for DNS (EDNS0)", RFC 2671, August 1999.
- [RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.
- [RFC6192] Dugal, D., Pignataro, C., and R. Dunn, "Protecting the Router Control Plane", RFC 6192, March 2011.
- [RFC6946] Gont, F., "Processing of IPv6 "Atomic" Fragments", RFC 6946, May 2013.

## Authors' Addresses

Joel Jaeggli  
Zynga  
630 taylor ct #10  
Mountain View, CA 94043  
USA

Email: [jjaeggli@zynga.com](mailto:jjaeggli@zynga.com)

Lorenzo Colitti  
Google

Email: [lorenzo@google.com](mailto:lorenzo@google.com)

Warren Kumari  
Google  
1600 Amphitheatre Parkway  
Mountain View, CA 94043  
USA

Email: [warren@kumari.net](mailto:warren@kumari.net)

Eric Vyncke  
Cisco  
De Kleetlaan 6A  
Diegem 1831  
Belgium

Email: [evyncke@cisco.com](mailto:evyncke@cisco.com)

Merike Kaeo  
Double Shot Security

Email: [merike@doubleshotsecurity.com](mailto:merike@doubleshotsecurity.com)

Tom Taylor (editor)  
Huawei Technologies  
Ottawa, Ontario  
Canada

Email: [tom.taylor.stds@gmail.com](mailto:tom.taylor.stds@gmail.com)



IPv6 Operations  
Internet-Draft  
Intended status: Informational  
Expires: January 16, 2014

M. Gysi  
G. Leclanche  
Swisscom  
E. Vyncke, Ed.  
Cisco Systems  
R. Anfinson  
Altibox  
July 15, 2013

Balanced Security for IPv6 CPE  
draft-v6ops-vyncke-balanced-ipv6-security-01.txt

Abstract

This document describes how an IPv6 residential Customer Premise Equipment (CPE) can have a balanced security policy that allows for a mostly end-to-end connectivity while keeping the major threats outside of the home. It is based on an actual IPv6 deployment by Swisscom and proposes to allow all packets inbound/outbound EXCEPT for some layer-4 ports where attacks and vulnerabilities (such as weak passwords) are well-known.

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 January 16, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Threats . . . . .	2
3. Overview . . . . .	3
3.1. Rules for Balanced Security Policy . . . . .	3
3.2. Rules example for Layer-4 Protection as Used by Swisscom . . . . .	4
4. IANA Considerations . . . . .	5
5. Security Considerations . . . . .	5
6. Acknowledgements . . . . .	6
7. Informative References . . . . .	6
Authors' Addresses . . . . .	6

## 1. Introduction

Internet access in residential IPv4 deployments generally consist of a single IPv4 address provided by the service provider for each home. Residential CPE then translates the single address into multiple private IPv4 addresses allowing more than one device in the home, but at the cost of losing end-to-end reachability. IPv6 allows all devices to have a unique, global, IP address, restoring end-to-end reachability directly between any device. Such reachability is very powerful for ubiquitous global connectivity, and is often heralded as one of the significant advantages to IPv6 over IPv4. Despite this, concern about exposure to inbound packets from the IPv6 Internet (which would otherwise be dropped by the address translation function if they had been sent from the IPv4 Internet) remain. This document describes firewall functionality for an IPv6 CPE which departs from the "simple security" model described in [RFC6092]. The intention is to provide an example of a security model which allows most traffic, including incoming unsolicited packets and connections, to traverse the CPE unless the CPE identifies the traffic as potentially harmful based on a set of rules. This model has been deployed successfully in Switzerland by Swisscom without any known security incident.

This document is applicable to off-the-shelves CPE as well to managed Service Provider CPE or for mobile Service Providers (where it can be centrally implemented).

## 2. Threats

For a typical residential network connected to the Internet over a broadband connection, the threats can be classified into:

- o denial of service by packet flooding: overwhelming either the access bandwidth or the bandwidth of a slower link in the residential network (like a slow home automation network) or the CPU power of a slow IPv6 host (like networked thermostat or any other sensor type nodes);
- o denial of service by Neighbor Discovery cache exhaustion [RFC6583]: the outside attacker floods the inside prefix(es) with packets with a random destination address forcing the CPE to exhaust its memory and its CPU in useless Neighbor Solicitations;
- o denial of service by service requests: like sending print jobs from the Internet to an ink jet printer until the ink cartridge is empty or like filing some file server with junk data;
- o unauthorized use of services: like accessing a webcam or a file server which are open to anonymous access within the residential network but should not be accessed from outside of the home network or accessing to remote desktop or SSH with weak password protection;
- o exploiting a vulnerability in the host in order to get access to data or to execute some arbitrary code in the attacked host such as several against old versions of Windows;
- o trojanized host (belonging to a Botnet) can communicate via a covert channel to its master and launch attacks to Internet targets.

### 3. Overview

The basic goal is to provide a pre-defined security policy which aims to block known harmful traffic and allow the rest, restoring as much of end-to-end communication as possible. This pre-defined policy can be centrally updated and could also be a member of a security policy menu for the subscriber.

#### 3.1. Rules for Balanced Security Policy

These are an example set of generic rules to be applied. Each would normally be configurable, either by the user directly or on behalf of the user by a subscription service.

If we name all nodes on the residential side of the CPE as 'inside' and all nodes on the Internet as 'outside', and any packet sent from

outside to inside as being 'inbound' and 'outbound' in the other direction, then the behavior of the CPE is described by a small set of rules:

1. Rule RejectBogon: apply ingress filtering in both directions per [RFC3704] and [RFC2827] for example with unicast reverse path forwarding (uRPF) checks (anti-spoofing) for all inbound and outbound traffic (implicitly blocking link-local and ULA in the same shot), this is basically the Section 2.1 Basic Sanitation and Section 3.1 Stateless Filters of [RFC6092];
2. Rule ProtectWeakServices: drop all inbound and outbound packets whose layer-4 destination is part of a limited set (see Section 3.2), the intent is to protect against the most common unauthorized access and avoid propagation of worms (even if the latter is questionable in IPv6); an advanced residential user should be able to modify this pre-defined list;
3. Rule Openess: allow all unsolicited inbound packets with rate limiting the initial packet of a new connection (such as TCP SYN, SCTP INIT or DCCP-request not applicable to UDP) to provide very basic protection against SYN port and address scanning attacks. All transport protocols and all non-deprecated extension headers are accepted. This is the major deviation from REC-11, REC-17 and REC-33 of [RFC6092].
4. All requirements of [RFC6092] except REC-11, REC-18 and REC-33 must be supported.

### 3.2. Rules example for Layer-4 Protection as Used by Swisscom

The rule ProtectWeakService can be implemented by using the following suggestions as implemented by Swisscom in 2013:

Transport	Port	Description
tcp	22	Secure Shell (SSH)
tcp	23	Telnet
tcp	80	HTTP
tcp	3389	Microsoft Remote Desktop Protocol
tcp	5900	VNC remote desktop protocol

Table 1: Drop Inbound

Transport	Port	Description
-----------	------	-------------

tcp-udp	88	Kerberos
tcp	111	SUN Remote Procedure Call
tcp	135	MS Remote Procedure Call
tcp	139	NetBIOS Session Service
tcp	445	Microsoft SMB Domain Server
tcp	513	Remote Login
tcp	514	Remote Shell
tcp	548	Apple Filing Protocol over TCP
tcp	631	Internet Printing Protocol
udp	1900	Simple Service Discovery Protocol
tcp	2869	Simple Service Discovery Protocol
udp	3702	Web Services Dynamic Discovery
udp	5353	Multicast DNS
udp	5355	Link-Lcl Mcast Name Resolution

Table 2: Drop Inbound and Outbound

This list should evolve with the time as new protocols and new threats appear, [DSHIELD] is used by Swisscom to keep those filters up to date. Another source of information could be the appendix A of [TR124]. The above proposal does not block GRE tunnels ([RFC2473]) so this is a deviation from [RFC6092].

Note: the authors believe that with this set the usual residential subscriber, the proverbial grand-ma, is protected. Of course, technical subscribers should be able to open other applications (identified by their TCP or UDP ports) through their CPE through some kind of user interface or even select a completely different security policy such as the open or 'closed' policies defined by [RFC6092].

#### 4. IANA Considerations

There are no extra IANA consideration for this document.

#### 5. Security Considerations

The authors of the documents believe and the Swisscom deployment shows that the following attack are mostly stopped:

- o Unauthorized access because vulnerable ports are blocked

This proposal cannot help with the following attacks:

- o Flooding of the CPE access link;

- o Malware which is fetched by inside hosts on a hostile web site (which is in 2012 the majority of infection sources).

## 6. Acknowledgements

The authors would like to thank several people who initiated the discussion on the `ipv6-ops@lists.cluonet.de` mailing list, notably: Tore Anderson, Lorenzo Colitti, Merike Kaeo, Simon Leinen, Eduard Metz, Martin Millnert, Benedikt Stockebrand.

## 7. Informative References

- [DSHIELD] DShield, "Port report: DShield", , <<https://secure.dshield.org/portreport.html?sort=records>>.
- [RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, December 1998.
- [RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.
- [RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.
- [RFC6092] Woodyatt, J., "Recommended Simple Security Capabilities in Customer Premises Equipment (CPE) for Providing Residential IPv6 Internet Service", RFC 6092, January 2011.
- [RFC6583] Gashinsky, I., Jaeggli, J., and W. Kumari, "Operational Neighbor Discovery Problems", RFC 6583, March 2012.
- [TR124] Broadband Forum, "Functional Requirements for Broadband Residential Gateway Devices", December 2006, <<http://www.broadband-forum.org/technical/download/TR-124.pdf>>.

## Authors' Addresses

Martin Gysi  
Swisscom  
Switzerland

Email: [Martin.Gysi@swisscom.com](mailto:Martin.Gysi@swisscom.com)

Guillaume Leclanche  
Swisscom  
Switzerland

Email: [Guillaume.Leclanche@swisscom.com](mailto:Guillaume.Leclanche@swisscom.com)

Eric Vyncke (editor)  
Cisco Systems  
De Kleetlaan 6a  
Diegem 1831  
Belgium

Phone: +32 2 778 4677  
Email: [evyncke@cisco.com](mailto:evyncke@cisco.com)

Ragnar Anfinssen  
Altibox  
Norway

Email: [Ragnar.Anfinssen@altibox.no](mailto:Ragnar.Anfinssen@altibox.no)