

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 27, 2022

J. Dong  
Z. Li  
Huawei Technologies  
C. Xie  
C. Ma  
China Telecom  
G. Mishra  
Verizon Inc.  
October 24, 2021

Carrying Virtual Transport Network (VTN) Identifier in IPv6 Extension  
Header  
draft-dong-6man-enhanced-vpn-vtn-id-06

Abstract

Virtual Private Networks (VPNs) provide different customers with logically separated connectivity over a common network infrastructure. With the introduction and evolvement of 5G and other network scenarios, some existing or new customers may require connectivity services with advanced characteristics comparing to traditional VPNs. Such kind of network service is called enhanced VPNs (VPN+).

A Virtual Transport Network (VTN) is a virtual underlay network which consists of a set of dedicated or shared network resources allocated from the physical underlay network, and is associated with a customized logical network topology. VPN+ services can be delivered by mapping one or a group of overlay VPNs to the appropriate VTNs as the virtual underlay. In packet forwarding, some fields in the data packet needs to be used to identify the VTN the packet belongs to, so that the VTN-specific processing can be performed on each node the packet traverses.

This document proposes a new Hop-by-Hop option of IPv6 extension header to carry the VTN Resource ID, which is used to identify the set of network resources allocated to a VTN for packet processing. The procedure for processing the VTN option is also specified.

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 <https://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 27, 2022.

#### Copyright Notice

Copyright (c) 2021 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 (<https://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
1.1. Requirements Language . . . . .	4
2. New IPv6 Extension Header Option for VTN . . . . .	4
3. Procedures . . . . .	5
3.1. VTN Option Insertion . . . . .	5
3.2. VTN based Packet Forwarding . . . . .	5
4. Operational Considerations . . . . .	6
5. IANA Considerations . . . . .	6
6. Security Considerations . . . . .	6
7. Contributors . . . . .	6
8. Acknowledgements . . . . .	7
9. References . . . . .	7
9.1. Normative References . . . . .	7
9.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

#### 1. Introduction

Virtual Private Networks (VPNs) provide different customers with logically isolated connectivity over a common network infrastructure. With the introduction and evolvement of 5G and other network

scenarios, some existing or new customers may require connectivity services with advanced characteristics comparing to traditional VPNs, such as resource isolation from other services or guaranteed performance. Such kind of network service is called enhanced VPN (VPN+). VPN+ service requires the coordination and integration between the overlay VPNs and the network characteristics of the underlay.

[I-D.ietf-teas-enhanced-vpn] describes a framework and the candidate component technologies for providing VPN+ services. It also introduces the concept of Virtual Transport Network (VTN). A Virtual Transport Network (VTN) is a virtual underlay network which consists of a set of dedicated or shared network resources allocated from the physical underlay network, and is associated with a customized logical network topology. VPN+ services can be delivered by mapping one or a group of overlay VPNs to the appropriate VTNs as the underlay, so as to provide the network characteristics required by the customers. In packet forwarding, traffic of different VPN+ services need to be processed separately based on the network resources and the logical topology associated with the corresponding VTN.

[I-D.dong-teas-enhanced-vpn-vtn-scalability] describes the scalability considerations and the possible optimizations for providing a relatively large number of VTNs for VPN+ services. One approach to improve the data plane scalability of VTN is to introduce a dedicated VTN Resource Identifier (VTN Resource ID) in the data packet to identify the set of network resources allocated to a VTN, so that VTN-specific packet processing can be performed using that set of resources, which avoids the possible resource competition with services in other VTNs. This is called Resource Independent (RI) VTN. A VTN Resource ID represents a subset of the resources (e.g. bandwidth, buffer and queuing resources) allocated on a given set of links and nodes which constitute a logical network topology. The logical topology associated with a VTN could be defined using mechanisms such as Multi-Topology [RFC4915], [RFC5120] or Flex-Algo [I-D.ietf-lsr-flex-algo], etc.

This document proposes a mechanism to carry the VTN resource ID in a new Hop-by-Hop option of IPv6 extension header [RFC8200] of IPv6 packet, so that on each network node along the packet forwarding path, the VTN option in the packet is parsed, and the obtained VTN Resource ID is used to instruct the network node to use the set of network resources allocated to the corresponding VTN to process and forward the packet. The procedure for processing the VTN Resource ID is also specified. This provides a scalable solution to support a relatively large number of VTNs in an IPv6 network.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP14 RFC 2119 [RFC2119] RFC 8174 [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. New IPv6 Extension Header Option for VTN

A new Hop-by-Hop option type "VTN" is defined to carry the VTN related Identifier in an IPv6 packet. Its format is shown as below:

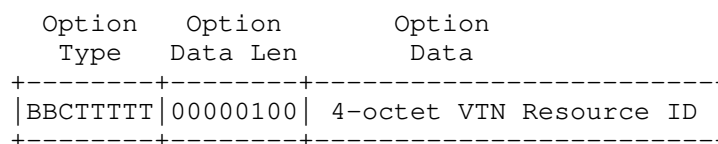


Figure 1. The format of VTN Option

**Option Type:** 8-bit identifier of the type of option. The type of VTN option is to be assigned by IANA. The highest-order bits of the type field are defined as below:

- o BB 00 The highest-order 2 bits are set to 00 to indicate that a node which does not recognize this type will skip over it and continue processing the header.
- o C 0 The third highest-order bit are set to 0 to indicate this option does not change en route.

**Opt Data Len:** 8-bit unsigned integer indicates the length of the option Data field of this option, in octets. The value of Opt Data Len of VTN option SHOULD be set to 4.

**VTN Resource ID:** 4-octet identifier which uniquely identifies the set of network resources allocated to a VTN.

**Editor's note:** The length of the VTN Resource ID is defined as 4-octet in correspondence to the 4-octet Single Network Slice Selection Assistance Information (S-NSSAI) defined in 3GPP [TS23501].

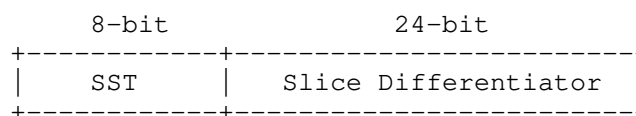


Figure 2. The format of S-NSSAI

### 3. Procedures

As the VTN option needs to be processed by each node along the path for VTN-specific forwarding, it SHOULD be carried in IPv6 Hop-by-Hop options header when the Hop-by-Hop options header can be either processed or ignored in forwarding plane by all the nodes along the path.

#### 3.1. VTN Option Insertion

When an ingress node of an IPv6 domain receives a packet, according to the traffic classification or mapping policy, the packet is steered into one of the VTNs in the network, then the packet SHOULD be encapsulated in an outer IPv6 header, and the Resource ID of the VTN which the packet is mapped to SHOULD be carried in the VTN option of the Hop-by-Hop options header associated with the outer IPv6 header.

#### 3.2. VTN based Packet Forwarding

On receipt of a packet with the VTN option, each network node which can process the VTN option in fast path SHOULD use the VTN Resource ID to determine the set of local network resources allocated to the VTN for packet processing. The packet forwarding behavior is based on both the destination IP address and the VTN Resource ID. More specifically, the destination IP address is used to determine the next-hop and the outgoing interface, and VTN Resource ID is used to determine the set of network resources on the outgoing interface which are reserved to the VTN for processing and sending the packet. The Traffic Class field of the outer IPv6 header MAY be used to provide Diffserv treatment for packets which belong to the same VTN. The egress node of the IPv6 domain SHOULD decapsulate the outer IPv6 header which includes the VTN option.

In the forwarding plane, there can be different approaches of partitioning the local network resources and allocating them to different VTNs. For example, on one physical interface, a subset of the forwarding plane resources (e.g. the bandwidth and the associated buffer and queuing resources) can be allocated to a particular VTN and represented as a virtual sub-interface with reserved bandwidth resource. In packet forwarding, the IPv6 destination address of the received packet is used to identify the next-hop and the outgoing layer-3 interface, and the VTN Resource ID is used to further identify the virtual sub-interface which is associated with the VTN on the outgoing interface.

Network nodes which do not support the processing of Hop-by-Hop options header SHOULD ignore the Hop-by-Hop options header and

forward the packet only based on the destination IP address. Network nodes which support Hop-by-Hop Options header, but do not support the VTN option SHOULD ignore the VTN option and continue to forward the packet based on the destination IP address and MAY also based on the rest of the Hop-by-Hop Options.

#### 4. Operational Considerations

As described in [RFC8200], network nodes may be configured to ignore the Hop-by-Hop Options header, and in some implementations a packet containing a Hop-by-Hop Options header may be dropped or assigned to a slow processing path. The proposed modification to the processing of IPv6 Hop-by-Hop options header is specified in [I-D.hinden-6man-hbh-processing]. Operator needs to make sure that all the network nodes involved in a VTN can either process Hop-by-Hop Options header in the fast path, or ignore the Hop-by-Hop Option header. Since a VTN is associated with a logical network topology, it is practical to ensure that all the network nodes involved in that logical topology support the processing of the HBH options header and the VTN option. In other word, packets steered into a VTN MUST NOT be dropped due to the existence of the Hop-by-Hop Options header. It is RECOMMENDED to configure all the network nodes involved in a VTN to process the Hop-by-Hop Options header and the VTN option if there is a nob for this.

#### 5. IANA Considerations

This document requests IANA to assign a new option type from "Destination Options and Hop-by-Hop Options" registry.

Value	Description	Reference
TBD	VTN Option	this document

#### 6. Security Considerations

The security considerations with IPv6 Hop-by-Hop options header are described in [RFC8200], [RFC7045] and [I-D.hinden-6man-hbh-processing]. This document introduces a new IPv6 Hop-by-Hop option which is either processed in the fast path or ignored by network nodes, thus it does not introduce additional security issues.

#### 7. Contributors

Zhibo Hu  
Email: huzhibo@huawei.com

Lei Bao  
Email: baolei7@huawei.com

## 8. Acknowledgements

The authors would like to thank Juhua Xu, James Guichard, Joel Halpern and Tom Petch for their review and valuable comments.

## 9. References

### 9.1. Normative References

- [I-D.ietf-teas-enhanced-vpn]  
Dong, J., Bryant, S., Li, Z., Miyasaka, T., and Y. Lee, "A Framework for Enhanced Virtual Private Network (VPN+) Services", draft-ietf-teas-enhanced-vpn-08 (work in progress), July 2021.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.

### 9.2. Informative References

- [I-D.dong-teas-enhanced-vpn-vtn-scalability]  
Dong, J., Li, Z., Gong, L., Yang, G., Guichard, J. N., Mishra, G., and F. Qin, "Scalability Considerations for Enhanced VPN (VPN+)", draft-dong-teas-enhanced-vpn-vtn-scalability-03 (work in progress), July 2021.
- [I-D.hinden-6man-hbh-processing]  
Hinden, R. M. and G. Fairhurst, "IPv6 Hop-by-Hop Options Processing Procedures", draft-hinden-6man-hbh-processing-01 (work in progress), June 2021.

- [I-D.ietf-lsr-flex-algo]  
Psenak, P., Hegde, S., Filsfils, C., Talaulikar, K., and  
A. Gulko, "IGP Flexible Algorithm", draft-ietf-lsr-flex-  
algo-17 (work in progress), July 2021.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P.  
Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF",  
RFC 4915, DOI 10.17487/RFC4915, June 2007,  
<<https://www.rfc-editor.org/info/rfc4915>>.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi  
Topology (MT) Routing in Intermediate System to  
Intermediate Systems (IS-IS)", RFC 5120,  
DOI 10.17487/RFC5120, February 2008,  
<<https://www.rfc-editor.org/info/rfc5120>>.
- [RFC7045] Carpenter, B. and S. Jiang, "Transmission and Processing  
of IPv6 Extension Headers", RFC 7045,  
DOI 10.17487/RFC7045, December 2013,  
<<https://www.rfc-editor.org/info/rfc7045>>.
- [TS23501] "3GPP TS23.501", 2016,  
<[https://portal.3gpp.org/desktopmodules/Specifications/  
SpecificationDetails.aspx?specificationId=3144](https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144)>.

#### Authors' Addresses

Jie Dong  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Road  
Beijing 100095  
China

Email: [jie.dong@huawei.com](mailto:jie.dong@huawei.com)

Zhenbin Li  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Road  
Beijing 100095  
China

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



Chongfeng Xie  
China Telecom  
China Telecom Beijing Information Science & Technology, Beiqijia  
Beijing 102209  
China

Email: xiechf@chinatelecom.cn

Chenhao Ma  
China Telecom  
China Telecom Beijing Information Science & Technology, Beiqijia  
Beijing 102209  
China

Email: machh@chinatelecom.cn

Gyan Mishra  
Verizon Inc.

Email: gyan.s.mishra@verizon.com

6MAN Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: October 30, 2022

G. Fioccola  
T. Zhou  
Huawei  
M. Cociglio  
Telecom Italia  
F. Qin  
China Mobile  
R. Pang  
China Unicom  
April 28, 2022

IPv6 Application of the Alternate Marking Method  
draft-ietf-6man-ipv6-alt-mark-14

Abstract

This document describes how the Alternate Marking Method can be used as a passive performance measurement tool in an IPv6 domain. It defines a new Extension Header Option to encode Alternate Marking information in both the Hop-by-Hop Options Header and Destination Options Header.

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 <https://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 October 30, 2022.

Copyright Notice

Copyright (c) 2022 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 (<https://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
1.1. Terminology . . . . .	3
1.2. Requirements Language . . . . .	3
2. Alternate Marking application to IPv6 . . . . .	3
2.1. Controlled Domain . . . . .	5
2.1.1. Alternate Marking Measurement Domain . . . . .	6
3. Definition of the AltMark Option . . . . .	7
3.1. Data Fields Format . . . . .	7
4. Use of the AltMark Option . . . . .	8
5. Alternate Marking Method Operation . . . . .	10
5.1. Packet Loss Measurement . . . . .	10
5.2. Packet Delay Measurement . . . . .	12
5.3. Flow Monitoring Identification . . . . .	13
5.4. Multipoint and Clustered Alternate Marking . . . . .	15
5.5. Data Collection and Calculation . . . . .	16
6. Security Considerations . . . . .	16
7. IANA Considerations . . . . .	20
8. Acknowledgements . . . . .	20
9. References . . . . .	20
9.1. Normative References . . . . .	20
9.2. Informative References . . . . .	21
Authors' Addresses . . . . .	22

## 1. Introduction

[I-D.ietf-ippm-rfc8321bis] and [I-D.ietf-ippm-rfc8889bis] describe a passive performance measurement method, which can be used to measure packet loss, latency and jitter on live traffic. Since this method is based on marking consecutive batches of packets, the method is often referred to as the Alternate Marking Method.

This document defines how the Alternate Marking Method can be used to measure performance metrics in IPv6. The rationale is to apply the Alternate Marking methodology to IPv6 and therefore allow detailed packet loss, delay and delay variation measurements both hop-by-hop and end-to-end to exactly locate the issues in an IPv6 network.

The Alternate Marking is an on-path telemetry technique and consists of synchronizing the measurements in different points of a network by

switching the value of a marking bit and therefore dividing the packet flow into batches. Each batch represents a measurable entity recognizable by all network nodes along the path. By counting the number of packets in each batch and comparing the values measured by different nodes, it is possible to precisely measure the packet loss. Similarly, the alternation of the values of the marking bits can be used as a time reference to calculate the delay and delay variation. The Alternate Marking operation is further described in Section 5.

The format of IPv6 addresses is defined in [RFC4291] while [RFC8200] defines the IPv6 Header, including a 20-bit Flow Label and the IPv6 Extension Headers.

This document introduces a new TLV (type-length-value) that can be encoded in the Options Headers (Hop-by-Hop or Destination) for the purpose of the Alternate Marking Method application in an IPv6 domain.

The threat model for the application of the Alternate Marking Method in an IPv6 domain is reported in Section 6. As with all on-path telemetry techniques, the only definitive solution is that this methodology MUST be applied in a controlled domain.

### 1.1. Terminology

This document uses the terms related to the Alternate Marking Method as defined in [I-D.ietf-ippm-rfc8321bis] and [I-D.ietf-ippm-rfc8889bis].

### 1.2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Alternate Marking application to IPv6

The Alternate Marking Method requires a marking field. Several alternatives could be considered such as IPv6 Extension Headers, IPv6 Address and Flow Label. But, it is necessary to analyze the drawbacks for all the available possibilities, more specifically:

Reusing existing Extension Header for Alternate Marking leads to a non-optimized implementation;

Using the IPv6 destination address to encode the Alternate Marking processing is very expensive;

Using the IPv6 Flow Label for Alternate Marking conflicts with the utilization of the Flow Label for load distribution purpose ([RFC6438]).

In the end, a new Hop-by-Hop or a new Destination Option is the best choice.

The approach for the Alternate Marking application to IPv6 specified in this memo is compliant with [RFC8200]. It involves the following operations:

- o The source node is the only one that writes the Option Header to mark alternately the flow (for both Hop-by-Hop and Destination Option). The intermediate nodes and destination node MUST only read the marking values of the option without modifying the Option Header.
- o In case of Hop-by-Hop Option Header carrying Alternate Marking bits, it is not inserted or deleted, but can be read by any node along the path. The intermediate nodes may be configured to support this Option or not and the measurement can be done only for the nodes configured to read the Option. As further discussed in Section 4, the presence of the hop-by-hop option should not affect the traffic throughput both on nodes that do not recognize this option and on the nodes that support it. However, it is worth mentioning that there is a difference between theory and practice. Indeed, in a real implementation it can happen that packets with hop-by-hop option could also be skipped or processed in the slow path. While some proposals are trying to address this problem and make Hop-by-Hop Options more practical ([I-D.peng-v6ops-hbh], [I-D.hinden-6man-hbh-processing]), these aspects are out of the scope for this document.
- o In case of Destination Option Header carrying Alternate Marking bits, it is not processed, inserted, or deleted by any node along the path until the packet reaches the destination node. Note that, if there is also a Routing Header (RH), any visited destination in the route list can process the Option Header.

Hop-by-Hop Option Header is also useful to signal to routers on the path to process the Alternate Marking. However, as said, routers will only examine this option if properly configured.

The optimization of both implementation and scaling of the Alternate Marking Method is also considered and a way to identify flows is

required. The Flow Monitoring Identification field (FlowMonID), as introduced in Section 5.3, goes in this direction and it is used to identify a monitored flow.

The FlowMonID is different from the Flow Label field of the IPv6 Header ([RFC6437]). The Flow Label field in the IPv6 header is used by a source to label sequences of packets to be treated in the network as a single flow and, as reported in [RFC6438], it can be used for load-balancing/equal cost multi-path (LB/ECMP). The reuse of Flow Label field for identifying monitored flows is not considered because it may change the application intent and forwarding behavior. Also, the Flow Label may be changed en route and this may also invalidate the integrity of the measurement. Furthermore, since the Flow Label is pseudo-random, there is always a finite probability of collision. Those reasons make the definition of the FlowMonID necessary for IPv6. Indeed, the FlowMonID is designed and only used to identify the monitored flow. Flow Label and FlowMonID within the same packet are totally disjoint, have different scope, are used to identify flows based on different criteria, and are intended for different use cases.

The rationale for the FlowMonID is further discussed in Section 5.3. This 20 bit field allows easy and flexible identification of the monitored flow and enables improved measurement correlation and finer granularity since it can be used in combination with the traditional TCP/IP 5-tuple to identify a flow. An important point that will be discussed in Section 5.3 is the uniqueness of the FlowMonID and how to allow disambiguation of the FlowMonID in case of collision.

The following section highlights an important requirement for the application of the Alternate Marking to IPv6. The concept of the controlled domain is explained and it is considered an essential precondition, as also highlighted in Section 6.

## 2.1. Controlled Domain

[RFC8799] introduces the concept of specific limited domain solutions and, in this regard, it is reported the IPv6 Application of the Alternate Marking Method as an example.

IPv6 has much more flexibility than IPv4 and innovative applications have been proposed, but for a number of reasons, such as the policies, options supported, the style of network management and security requirements, it is suggested to limit some of these applications to a controlled domain. This is also the case of the Alternate Marking application to IPv6 as assumed hereinafter.

Therefore, the IPv6 application of the Alternate Marking Method MUST be deployed in a controlled domain. It is RECOMMENDED that an implementation filters packets that carry Alternate Marking data and are entering or leaving the controlled domains.

A controlled domain is a managed network where it is required to select, monitor and control the access to the network by enforcing policies at the domain boundaries in order to discard undesired external packets entering the domain and check the internal packets leaving the domain. It does not necessarily mean that a controlled domain is a single administrative domain or a single organization. A controlled domain can correspond to a single administrative domain or can be composed by multiple administrative domains under a defined network management. Indeed, some scenarios may imply that the Alternate Marking Method involves more than one domain, but in these cases, it is RECOMMENDED that the multiple domains create a whole controlled domain while traversing the external domain by employing IPsec [RFC4301] authentication and encryption or other VPN technology that provides full packet confidentiality and integrity protection. In a few words, it must be possible to control the domain boundaries and eventually use specific precautions if the traffic traverse the Internet.

The security considerations reported in Section 6 also highlight this requirement.

#### 2.1.1. Alternate Marking Measurement Domain

The Alternate Marking measurement domain can overlap with the controlled domain or may be a subset of the controlled domain. The typical scenarios for the application of the Alternate Marking Method depend on the controlled domain boundaries, in particular:

the user equipment can be the starting or ending node, only in case it is fully managed and if it belongs to the controlled domain. In this case the user generated IPv6 packets contain the Alternate Marking data. But, in practice, this is not common due to the fact that the user equipment cannot be totally secured in the majority of cases.

the CPE (Customer Premises Equipment) is most likely to be the starting or ending node since it connects the user's premises with the service provider's network and therefore belongs to the operator's controlled domain. Typically the CPE encapsulates a received packet in an outer IPv6 header which contains the Alternate Marking data. The CPE can also be able to filter and drop packets from outside of the domain with inconsistent fields to make effective the relevant security rules at the domain

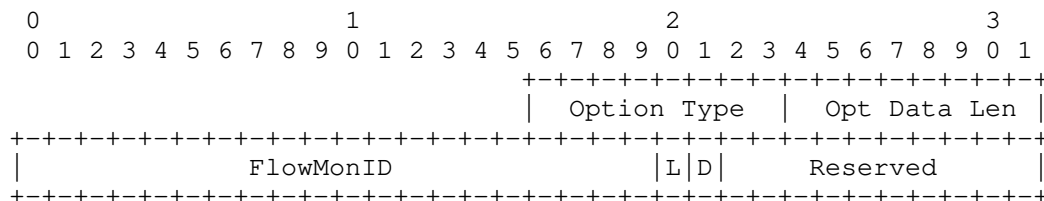
boundaries, for example a simple security check can be to insert the Alternate Marking data if and only if the destination is within the controlled domain.

### 3. Definition of the AltMark Option

The definition of a new TLV for the Options Extension Headers, carrying the data fields dedicated to the Alternate Marking method, is reported below.

#### 3.1. Data Fields Format

The following figure shows the data fields format for enhanced Alternate Marking TLV (AltMark). This AltMark data can be encapsulated in the IPv6 Options Headers (Hop-by-Hop or Destination Option).



where:

- o Option Type: 8-bit identifier of the type of Option that needs to be allocated. Unrecognized Types MUST be ignored on processing. For Hop-by-Hop Options Header or Destination Options Header, [RFC8200] defines how to encode the three high-order bits of the Option Type field. The two high-order bits specify the action that must be taken if the processing IPv6 node does not recognize the Option Type; for AltMark these two bits MUST be set to 00 (skip over this Option and continue processing the header). The third-highest-order bit specifies whether the Option Data can change en route to the packet's final destination; for AltMark the value of this bit MUST be set to 0 (Option Data does not change en route). In this way, since the three high-order bits of the AltMark Option are set to 000, it means that nodes can simply skip this Option if they do not recognize and that the data of this Option do not change en route, indeed the source is the only one that can write it.
- o Opt Data Len: 4. It is the length of the Option Data Fields of this Option in bytes.



- o FlowMonID: 20-bit unsigned integer. The FlowMon identifier is described in Section 5.3. As further discussed below, it has been picked as 20 bits since it is a reasonable value and a good compromise in relation to the chance of collision. It MUST be set pseudo randomly by the source node or by a centralized controller.
- o L: Loss flag for Packet Loss Measurement as described in Section 5.1;
- o D: Delay flag for Single Packet Delay Measurement as described in Section 5.2;
- o Reserved: is reserved for future use. These bits MUST be set to zero on transmission and ignored on receipt.

#### 4. Use of the AltMark Option

The AltMark Option is the best way to implement the Alternate Marking method and it is carried by the Hop-by-Hop Options header and the Destination Options header. In case of Destination Option, it is processed only by the source and destination nodes: the source node inserts and the destination node processes it. While, in case of Hop-by-Hop Option, it may be examined by any node along the path, if explicitly configured to do so.

It is important to highlight that the Option Layout can be used both as Destination Option and as Hop-by-Hop Option depending on the Use Cases and it is based on the chosen type of performance measurement. In general, it is needed to perform both end to end and hop by hop measurements, and the Alternate Marking methodology allows, by definition, both performance measurements. In many cases the end-to-end measurement is not enough and it is required the hop-by-hop measurement, so the most complete choice can be the Hop-by-Hop Options Header.

IPv6, as specified in [RFC8200], allows nodes to optionally process Hop-by-Hop headers. Specifically the Hop-by-Hop Options header is not inserted or deleted, but may be examined or processed by any node along a packet's delivery path, until the packet reaches the node (or each of the set of nodes, in the case of multicast) identified in the Destination Address field of the IPv6 header. Also, it is expected that nodes along a packet's delivery path only examine and process the Hop-by-Hop Options header if explicitly configured to do so.

Another scenario that can be mentioned is the presence of a Routing Header, in particular it is possible to consider SRv6. A new type of Routing Header, referred as Segment Routing Header (SRH), has been defined in [RFC8754] for SRv6. Like any other use case of IPv6, Hop-

by-Hop and Destination Options are usable when SRv6 header is present. Because SRv6 is implemented through a Segment Routing Header (SRH), Destination Options before the Routing Header are processed by each destination in the route list, that means, in case of SRH, by every SR node that is identified by the SR path. More details about the SRv6 application are described in [I-D.fz-spring-srv6-alt-mark].

In summary, it is possible to list the alternative possibilities:

- o Destination Option not preceding a Routing Header => measurement only by node in Destination Address.
- o Hop-by-Hop Option => every router on the path with feature enabled.
- o Destination Option preceding a Routing Header => every destination node in the route list.

In general, Hop-by-Hop and Destination Options are the most suitable ways to implement Alternate Marking.

It is worth mentioning that new Hop-by-Hop Options are not strongly recommended in [RFC7045] and [RFC8200], unless there is a clear justification to standardize it, because nodes may be configured to ignore the Options Header, drop or assign packets containing an Options Header to a slow processing path. In case of the AltMark data fields described in this document, the motivation to standardize a new Hop-by-Hop Option is that it is needed for OAM (Operations, Administration, and Maintenance). An intermediate node can read it or not, but this does not affect the packet behavior. The source node is the only one that writes the Hop-by-Hop Option to mark alternately the flow, so, the performance measurement can be done for those nodes configured to read this Option, while the others are simply not considered for the metrics.

The Hop-by-Hop Option defined in this document is designed to take advantage of the property of how Hop-by-Hop options are processed. Nodes that do not support this Option SHOULD ignore them. This can mean that, in this case, the performance measurement does not account for all links and nodes along a path. The definition of the Hop-by-Hop Options in this document is also designed to minimize throughput impact both on nodes that do not recognize the Option and on node that support it. Indeed, the three high-order bits of the Options Header defined in this draft are 000 and, in theory, as per [RFC8200] and [I-D.hinden-6man-hbh-processing], this means "skip if do not recognize and data do not change en route". [RFC8200] also mentions that the nodes only examine and process the Hop-by-Hop Options header

if explicitly configured to do so. For these reasons, this Hop-by-Hop Option should not affect the throughput. However, in practice, it is important to be aware that the things may be different in the implementation and it can happen that packets with Hop-by-Hop are forced onto the slow path, but this is a general issue, as also explained in [I-D.hinden-6man-hbh-processing]. It is also worth mentioning that the application to a controlled domain should avoid the risk of arbitrary nodes dropping packets with Hop-by-Hop Options.

## 5. Alternate Marking Method Operation

This section describes how the method operates.

[I-D.ietf-ippm-rfc8321bis] introduces several applicable methods which are reported below, and a new field is introduced to facilitate the deployment and improve the scalability.

### 5.1. Packet Loss Measurement

The measurement of the packet loss is really straightforward in comparison to the existing mechanisms, as detailed in [I-D.ietf-ippm-rfc8321bis]. The packets of the flow are grouped into batches, and all the packets within a batch are marked by setting the L bit (Loss flag) to a same value. The source node can switch the value of the L bit between 0 and 1 after a fixed number of packets or according to a fixed timer, and this depends on the implementation. The source node is the only one that marks the packets to create the batches, while the intermediate nodes only read the marking values and identify the packet batches. By counting the number of packets in each batch and comparing the values measured by different network nodes along the path, it is possible to measure the packet loss occurred in any single batch between any two nodes. Each batch represents a measurable entity recognizable by all network nodes along the path.

Both fixed number of packets and fixed timer can be used by the source node to create packet batches. But, as also explained in [I-D.ietf-ippm-rfc8321bis], the timer-based batches are preferable because they are more deterministic than the counter-based batches. There is no definitive rule for counter-based batches, differently from timer-based batches. Using a fixed timer for the switching offers better control over the method, indeed the length of the batches can be chosen large enough to simplify the collection and the comparison of the measures taken by different network nodes. In the implementation the counters can be sent out by each node to the controller that is responsible for the calculation. It is also possible to exchange this information by using other on-path techniques. But this is out of scope for this document.

Packets with different L values may get swapped at batch boundaries, and in this case, it is required that each marked packet can be assigned to the right batch by each router. It is important to mention that for the application of this method there are two elements to consider: the clock error between network nodes and the network delay. These can create offsets between the batches and out-of-order of the packets. The mathematical formula on timing aspects, explained in section 5 of [I-D.ietf-ippm-rfc8321bis], must be satisfied and it takes into considerations the different causes of reordering such as clock error and network delay. The assumption is to define the available counting interval where to get stable counters and to avoid these issues. Specifically, if the effects of network delay are ignored, the condition to implement the methodology is that the clocks in different nodes MUST be synchronized to the same clock reference with an accuracy of  $\pm B/2$  time units, where B is the fixed time duration of the batch, which refers to the original marking interval at the source node considering that this interval could fluctuate along the path. In this way each marked packet can be assigned to the right batch by each node. Usually the counters can be taken in the middle of the batch period to be sure to take still counters. In a few words this implies that the length of the batches MUST be chosen large enough so that the method is not affected by those factors. The length of the batches can be determined based on the specific deployment scenario.

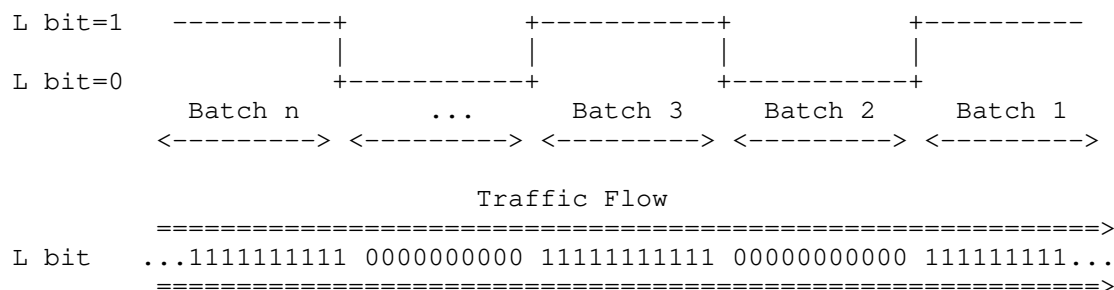


Figure 1: Packet Loss Measurement and Single-Marking Methodology using L bit

It is worth mentioning that the duration of the batches is considered stable over time in the previous figure. In theory, it is possible to change the length of batches over time and among different flows for more flexibility. But, in practice, it could complicate the correlation of the information.

## 5.2. Packet Delay Measurement

The same principle used to measure packet loss can be applied also to one-way delay measurement. Delay metrics MAY be calculated using the two possibilities:

1. **Single-Marking Methodology:** This approach uses only the L bit to calculate both packet loss and delay. In this case, the D flag MUST be set to zero on transmit and ignored by the monitoring points. The alternation of the values of the L bit can be used as a time reference to calculate the delay. Whenever the L bit changes and a new batch starts, a network node can store the timestamp of the first packet of the new batch, that timestamp can be compared with the timestamp of the first packet of the same batch on a second node to compute packet delay. But this measurement is accurate only if no packet loss occurs and if there is no packet reordering at the edges of the batches. A different approach can also be considered and it is based on the concept of the mean delay. The mean delay for each batch is calculated by considering the average arrival time of the packets for the relative batch. There are limitations also in this case indeed, each node needs to collect all the timestamps and calculate the average timestamp for each batch. In addition, the information is limited to a mean value.
2. **Double-Marking Methodology:** This approach is more complete and uses the L bit only to calculate packet loss and the D bit (Delay flag) is fully dedicated to delay measurements. The idea is to use the first marking with the L bit to create the alternate flow and, within the batches identified by the L bit, a second marking is used to select the packets for measuring delay. The D bit creates a new set of marked packets that are fully identified over the network, so that a network node can store the timestamps of these packets; these timestamps can be compared with the timestamps of the same packets on a second node to compute packet delay values for each packet. The most efficient and robust mode is to select a single double-marked packet for each batch, in this way there is no time gap to consider between the double-marked packets to avoid their reorder. Regarding the rule for the selection of the packet to be double-marked, the same considerations in Section 5.1 apply also here and the double-marked packet can be chosen within the available counting interval that is not affected by factors such as clock errors. If a double-marked packet is lost, the delay measurement for the considered batch is simply discarded, but this is not a big problem because it is easy to recognize the problematic batch and skip the measurement just for that one. So in order to have more

information about the delay and to overcome out-of-order issues this method is preferred.

In summary the approach with double marking is better than the approach with single marking. Moreover, the two approaches provide slightly different pieces of information and the data consumer can combine them to have a more robust data set.

Similar to what said in Section 5.1 for the packet counters, in the implementation the timestamps can be sent out to the controller that is responsible for the calculation or could also be exchanged using other on-path techniques. But this is out of scope for this document.

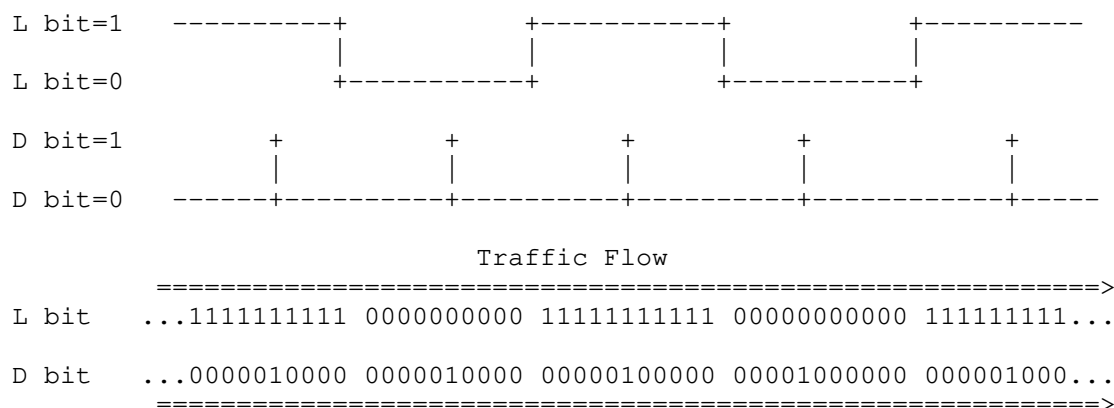


Figure 2: Double-Marking Methodology using L bit and D bit

Likewise to packet delay measurement (both for Single Marking and Double Marking), the method can also be used to measure the inter-arrival jitter.

### 5.3. Flow Monitoring Identification

The Flow Monitoring Identification (FlowMonID) identifies the flow to be measured and is required for some general reasons:

First, it helps to reduce the per node configuration. Otherwise, each node needs to configure an access-control list (ACL) for each of the monitored flows. Moreover, using a flow identifier allows a flexible granularity for the flow definition, indeed, it can be used together with other identifiers (e.g. 5-tuple).

Second, it simplifies the counters handling. Hardware processing of flow tuples (and ACL matching) is challenging and often incurs into performance issues, especially in tunnel interfaces.

Third, it eases the data export encapsulation and correlation for the collectors.

The FlowMonID MUST only be used as a monitored flow identifier in order to determine a monitored flow within the measurement domain. This entails not only an easy identification but improved correlation as well.

The value of 20 bits has been selected for the FlowMonID since it is a good compromise and implies a low rate of ambiguous FlowMonIDs that can be considered acceptable in most of the applications. The disambiguation issue can be solved by tagging the pseudo randomly generated FlowMonID with additional flow information. In particular, it is RECOMMENDED to consider the 3-tuple FlowMonID, source and destination addresses:

- o If the 20 bit FlowMonID is set independently and pseudo randomly in a distributed way there is a chance of collision. Indeed, by using the well-known birthday problem in probability theory, if the 20 bit FlowMonID is set independently and pseudo randomly without any additional input entropy, there is a 50% chance of collision for 1206 flows. So, for more entropy, FlowMonID is combined with source and destination addresses. Since there is a 1% chance of collision for 145 flows, it is possible to monitor 145 concurrent flows per host pairs with a 1% chance of collision.
- o If the 20 bits FlowMonID is set pseudo randomly but in a centralized way, the controller can instruct the nodes properly in order to guarantee the uniqueness of the FlowMonID. With 20 bits, the number of combinations is 1048576, and the controller should ensure that all the FlowMonID values are used without any collision. Therefore, by considering source and destination addresses together with the FlowMonID, it can be possible to monitor 1048576 concurrent flows per host pairs.

A consistent approach MUST be used in the Alternate Marking deployment to avoid the mixture of different ways of identifying. All the nodes along the path and involved into the measurement SHOULD use the same mode for identification. As mentioned, it is RECOMMENDED to use the FlowMonID for identification purpose in combination with source and destination addresses to identify a flow. By considering source and destination addresses together with the FlowMonID it can be possible to monitor 145 concurrent flows per host pairs with a 1% chance of collision in case of pseudo randomly

generated FlowMonID, or 1048576 concurrent flows per host pairs in case of centralized controller. It is worth mentioning that the solution with the centralized control allows finer granularity and therefore adds even more flexibility to the flow identification.

The FlowMonID field is set at the source node, which is the ingress point of the measurement domain, and can be set in two ways:

- a. It can be algorithmically generated by the source node, that can set it pseudo-randomly with some chance of collision. This approach cannot guarantee the uniqueness of FlowMonID since conflicts and collisions are possible. But, considering the recommendation to use FlowMonID with source and destination addresses the conflict probability is reduced due to the FlowMonID space available for each endpoint pair (i.e. 145 flows with 1% chance of collision).
- b. It can be assigned by the central controller. Since the controller knows the network topology, it can allocate the value properly to avoid or minimize ambiguity and guarantee the uniqueness. In this regard, the controller can verify that there is no ambiguity between different pseudo-randomly generated FlowMonIDs on the same path. The conflict probability is really small given that the FlowMonID is coupled with source and destination addresses and up to 1048576 flows can be monitored for each endpoint pair. When all values in the FlowMonID space are consumed, the centralized controller can keep track and reassign the values that are not used any more by old flows.

If the FlowMonID is set by the source node, the intermediate nodes can read the FlowMonIDs from the packets in flight and act accordingly. While, if the FlowMonID is set by the controller, both possibilities are feasible for the intermediate nodes which can learn by reading the packets or can be instructed by the controller.

#### 5.4. Multipoint and Clustered Alternate Marking

The Alternate Marking method can also be extended to any kind of multipoint to multipoint paths, and the network clustering approach allows a flexible and optimized performance measurement, as described in [I-D.ietf-ippm-rfc8889bis].

The Cluster is the smallest identifiable subnetwork of the entire Network graph that still satisfies the condition that the number of packets that goes in is the same that goes out. With network clustering, it is possible to use the partition of the network into clusters at different levels in order to perform the needed degree of detail. So, for Multipoint Alternate Marking, FlowMonID can identify



in general a multipoint-to-multipoint flow and not only a point-to-point flow.

#### 5.5. Data Collection and Calculation

The nodes enabled to perform performance monitoring collect the value of the packet counters and timestamps. There are several alternatives to implement Data Collection and Calculation, but this is not specified in this document.

There are documents on the control plane mechanisms of Alternate Marking, e.g. [I-D.ietf-idr-sr-policy-ifit], [I-D.chen-pce-pcep-ifit].

#### 6. Security Considerations

This document aims to apply a method to perform measurements that does not directly affect Internet security nor applications that run on the Internet. However, implementation of this method must be mindful of security and privacy concerns.

There are two types of security concerns: potential harm caused by the measurements and potential harm to the measurements.

Harm caused by the measurement: Alternate Marking implies modifications on the fly to an Option Header of IPv6 packets by the source node, but this must be performed in a way that does not alter the quality of service experienced by the packets and that preserves stability and performance of routers doing the measurements. As already discussed in Section 4, it is RECOMMENDED that the AltMark Option does not affect the throughput and therefore the user experience.

Harm to the measurement: Alternate Marking measurements could be harmed by routers altering the fields of the AltMark Option (e.g. marking of the packets, FlowMonID) or by a malicious attacker adding AltMark Option to the packets in order to consume the resources of network devices and entities involved. As described above, the source node is the only one that writes the Option Header while the intermediate nodes and destination node only read it without modifying the Option Header. But, for example, an on-path attacker can modify the flags, whether intentionally or accidentally, or deliberately insert a new option to the packet flow or delete the option from the packet flow. The consequent effect could be to give the appearance of loss or delay or invalidate the measurement by modifying option identifiers, such as FlowMonID. The malicious implication can be to cause actions from the network administrator where an intervention is not necessary or to hide real issues in the

network. Since the measurement itself may be affected by network nodes intentionally altering the bits of the AltMark Option or injecting Options headers as a means for Denial of Service (DoS), the Alternate Marking MUST be applied in the context of a controlled domain, where the network nodes are locally administered and this type of attack can be avoided. For this reason, the implementation of the method is not done on the end node if it is not fully managed and does not belong to the controlled domain. Packets generated outside the controlled domain may consume router resources by maliciously using the HbH Option, but this can be mitigated by filtering these packets at the controlled domain boundary. This can be done because, if the end node does not belong to the controlled domain, it is not supposed to add the AltMark HbH Option, and it can be easily recognized.

An attacker that does not belong to the controlled domain can maliciously send packets with AltMark Option. But if Alternate Marking is not supported in the controlled domain, no problem happens because the AltMark Option is treated as any other unrecognized option and will not be considered by the nodes since they are not configured to deal with it, so the only effect is the increased MTU (by 48 bits). While if Alternate Marking is supported in the controlled domain, it is also necessary to avoid that the measurements are affected and external packets with AltMark Option MUST be filtered. As any other Hop-by-Hop Options or Destination Options, it is possible to filter AltMark Options entering or leaving the domain e.g. by using ACL extensions for filtering.

The flow identifier (FlowMonID) composes the AltMark Option together with the two marking bits (L and D). As explained in Section 5.3, there is a chance of collision if the FlowMonID is set pseudo randomly and a solution exists. In general this may not be a problem and a low rate of ambiguous FlowMonIDs can be acceptable, since this does not cause significant harm to the operators or their clients and this harm may not justify the complications of avoiding it. But, for large scale measurements, a big number of flows could be monitored and the probability of a collision is higher, thus the disambiguation of the FlowMonID field can be considered.

The privacy concerns also need to be analyzed even if the method only relies on information contained in the Option Header without any release of user data. Indeed, from a confidentiality perspective, although AltMark Option does not contain user data, the metadata can be used for network reconnaissance to compromise the privacy of users by allowing attackers to collect information about network performance and network paths. AltMark Option contains two kinds of metadata: the marking bits (L and D bits) and the flow identifier (FlowMonID).

The marking bits are the small information that is exchanged between the network nodes. Therefore, due to this intrinsic characteristic, network reconnaissance through passive eavesdropping on data-plane traffic is difficult. Indeed, an attacker cannot gain information about network performance from a single monitoring point. The only way for an attacker can be to eavesdrop on multiple monitoring points at the same time, because they have to do the same kind of calculation and aggregation as Alternate Marking requires.

The FlowMonID field is used in the AltMark Option as the identifier of the monitored flow. It represents a more sensitive information for network reconnaissance and may allow a flow tracking type of attack because an attacker could collect information about network paths.

Furthermore, in a pervasive surveillance attack, the information that can be derived over time is more. But, as further described hereinafter, the application of the Alternate Marking to a controlled domain helps to mitigate all the above aspects of privacy concerns.

At the management plane, attacks can be set up by misconfiguring or by maliciously configuring AltMark Option. Thus, AltMark Option configuration MUST be secured in a way that authenticates authorized users and verifies the integrity of configuration procedures. Solutions to ensure the integrity of AltMark Option are outside the scope of this document. Also, attacks on the reporting of the statistics between the monitoring points and the network management system (e.g. centralized controller) can interfere with the proper functioning of the system. Hence, the channels used to report back flow statistics MUST be secured.

As stated above, the precondition for the application of the Alternate Marking is that it MUST be applied in specific controlled domains, thus confining the potential attack vectors within the network domain. [RFC8799] analyzes and discusses the trend towards network behaviors that can be applied only within a limited domain. This is due to the specific set of requirements especially related to security, network management, policies and options supported which may vary between such limited domains. A limited administrative domain provides the network administrator with the means to select, monitor and control the access to the network, making it a trusted domain. In this regard it is expected to enforce policies at the domain boundaries to filter both external packets with AltMark Option entering the domain and internal packets with AltMark Option leaving the domain. Therefore, the trusted domain is unlikely subject to hijacking of packets since packets with AltMark Option are processed and used only within the controlled domain.

As stated, the application to a controlled domain ensures the control over the packets entering and leaving the domain, but despite that, leakages may happen for different reasons, such as a failure or a fault. In this case, nodes outside the domain MUST simply ignore packets with AltMark Option since they are not configured to handle it and should not process it.

Additionally, it is to be noted that the AltMark Option is carried by the Options Header and it may have some impact on the packet sizes for the monitored flow and on the path MTU, since some packets might exceed the MTU. However, the relative small size (48 bit in total) of these Option Headers and its application to a controlled domain help to mitigate the problem.

It is worth mentioning that the security concerns may change based on the specific deployment scenario and related threat analysis, which can lead to specific security solutions that are beyond the scope of this document. As an example, the AltMark Option can be used as Hop-by-Hop or Destination Option and, in case of Destination Option, multiple administrative domains may be traversed by the AltMark Option that is not confined to a single administrative domain. In this case, the user, aware of the kind of risks, may still want to use Alternate Marking for telemetry and test purposes but the controlled domain must be composed by more than one administrative domains. To this end, the inter-domain links need to be secured (e.g., by IPsec, VPNs) in order to avoid external threats and realize the whole controlled domain.

It might be theoretically possible to modulate the marking or the other fields of the AltMark Option to serve as a covert channel to be used by an on-path observer. This may affect both the data and management plane, but, here too, the application to a controlled domain helps to reduce the effects.

The Alternate Marking application described in this document relies on a time synchronization protocol. Thus, by attacking the time protocol, an attacker can potentially compromise the integrity of the measurement. A detailed discussion about the threats against time protocols and how to mitigate them is presented in [RFC7384]. Network Time Security (NTS), described in [RFC8915], is a mechanism that can be employed. Also, the time, which is distributed to the network nodes through the time protocol, is centrally taken from an external accurate time source, such as an atomic clock or a GPS clock. By attacking the time source it can be possible to compromise the integrity of the measurement as well. There are security measures that can be taken to mitigate the GPS spoofing attacks and a network administrator should certainly employ solutions to secure the network domain.

## 7. IANA Considerations

The Option Type should be assigned in IANA's "Destination Options and Hop-by-Hop Options" registry.

This draft requests the following IPv6 Option Type assignment from the Destination Options and Hop-by-Hop Options sub-registry of Internet Protocol Version 6 (IPv6) Parameters (<https://www.iana.org/assignments/ipv6-parameters/>).

Hex Value	Binary Value act chg rest			Description	Reference
TBD	00	0	tbd	AltMark	[This draft]

## 8. Acknowledgements

The authors would like to thank Bob Hinden, Ole Troan, Martin Duke, Lars Eggert, Roman Danyliw, Alvaro Retana, Eric Vyncke, Warren Kumari, Benjamin Kaduk, Stewart Bryant, Christopher Wood, Yoshifumi Nishida, Tom Herbert, Stefano Previdi, Brian Carpenter, Greg Mirsky, Ron Bonica for the precious comments and suggestions.

## 9. References

### 9.1. Normative References

- [I-D.ietf-ippm-rfc8321bis]  
Fioccola, G., Cociglio, M., Mirsky, G., Mizrahi, T., and T. Zhou, "Alternate-Marking Method", draft-ietf-ippm-rfc8321bis-01 (work in progress), April 2022.
- [I-D.ietf-ippm-rfc8889bis]  
Fioccola, G., Cociglio, M., Sapio, A., Sisto, R., and T. Zhou, "Multipoint Alternate-Marking Clustered Method", draft-ietf-ippm-rfc8889bis-01 (work in progress), April 2022.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.

## 9.2. Informative References

- [I-D.chen-pce-pcep-ifit]  
Yuan, H., Zhou, T., Li, W., Fioccola, G., and Y. Wang, "Path Computation Element Communication Protocol (PCEP) Extensions to Enable IFIT", draft-chen-pce-pcep-ifit-06 (work in progress), February 2022.
- [I-D.fz-spring-srv6-alt-mark]  
Fioccola, G., Zhou, T., and M. Cociglio, "Segment Routing Header encapsulation for Alternate Marking Method", draft-fz-spring-srv6-alt-mark-02 (work in progress), February 2022.
- [I-D.hinden-6man-hbh-processing]  
Hinden, R. M. and G. Fairhurst, "IPv6 Hop-by-Hop Options Processing Procedures", draft-hinden-6man-hbh-processing-01 (work in progress), June 2021.
- [I-D.ietf-idr-sr-policy-ifit]  
Qin, F., Yuan, H., Zhou, T., Fioccola, G., and Y. Wang, "BGP SR Policy Extensions to Enable IFIT", draft-ietf-idr-sr-policy-ifit-03 (work in progress), January 2022.
- [I-D.peng-v6ops-hbh]  
Peng, S., Li, Z., Xie, C., Qin, Z., and G. Mishra, "Processing of the Hop-by-Hop Options Header", draft-peng-v6ops-hbh-06 (work in progress), August 2021.
- [RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, DOI 10.17487/RFC4291, February 2006, <<https://www.rfc-editor.org/info/rfc4291>>.
- [RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <<https://www.rfc-editor.org/info/rfc4301>>.
- [RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, DOI 10.17487/RFC6437, November 2011, <<https://www.rfc-editor.org/info/rfc6437>>.

- [RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, DOI 10.17487/RFC6438, November 2011, <<https://www.rfc-editor.org/info/rfc6438>>.
- [RFC7045] Carpenter, B. and S. Jiang, "Transmission and Processing of IPv6 Extension Headers", RFC 7045, DOI 10.17487/RFC7045, December 2013, <<https://www.rfc-editor.org/info/rfc7045>>.
- [RFC7384] Mizrahi, T., "Security Requirements of Time Protocols in Packet Switched Networks", RFC 7384, DOI 10.17487/RFC7384, October 2014, <<https://www.rfc-editor.org/info/rfc7384>>.
- [RFC8754] Filsfils, C., Ed., Dukes, D., Ed., Previdi, S., Leddy, J., Matsushima, S., and D. Voyer, "IPv6 Segment Routing Header (SRH)", RFC 8754, DOI 10.17487/RFC8754, March 2020, <<https://www.rfc-editor.org/info/rfc8754>>.
- [RFC8799] Carpenter, B. and B. Liu, "Limited Domains and Internet Protocols", RFC 8799, DOI 10.17487/RFC8799, July 2020, <<https://www.rfc-editor.org/info/rfc8799>>.
- [RFC8915] Franke, D., Sibold, D., Teichel, K., Dansarie, M., and R. Sundblad, "Network Time Security for the Network Time Protocol", RFC 8915, DOI 10.17487/RFC8915, September 2020, <<https://www.rfc-editor.org/info/rfc8915>>.

## Authors' Addresses

Giuseppe Fioccola  
Huawei  
Riesstrasse, 25  
Munich 80992  
Germany

Email: [giuseppe.fioccola@huawei.com](mailto:giuseppe.fioccola@huawei.com)

Tianran Zhou  
Huawei  
156 Beiqing Rd.  
Beijing 100095  
China

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

Mauro Cociglio  
Telecom Italia  
Via Reiss Romoli, 274  
Torino 10148  
Italy

Email: [mauro.cociglio@telecomitalia.it](mailto:mauro.cociglio@telecomitalia.it)

Fengwei Qin  
China Mobile  
32 Xuanwumenxi Ave.  
Beijing 100032  
China

Email: [qinfengwei@chinamobile.com](mailto:qinfengwei@chinamobile.com)

Ran Pang  
China Unicom  
9 Shouti South Rd.  
Beijing 100089  
China

Email: [pangran@chinaunicom.cn](mailto:pangran@chinaunicom.cn)



Network Working Group  
Internet-Draft  
Updates: RFC8200, RFC8201, RFC4443, RFC1191 (if approved)  
Intended status: Standards Track  
Expires: 30 September 2022

F. L. Templin, Ed.  
Boeing Research & Technology  
29 March 2022

IPv6 Fragment Retransmission and Path MTU Discovery Soft Errors  
draft-templin-6man-fragrep-07

Abstract

Internet Protocol version 6 (IPv6) provides a fragmentation and reassembly service for end systems allowing for the transmission of packets that exceed the path MTU. However, loss of individual fragments requires retransmission of original packets in their entirety leading to cascading reassembly failures. This document specifies an IPv6 fragment retransmission scheme that matches the loss unit to the retransmission unit. The document further specifies an update to Path MTU Discovery that distinguishes hard link size restrictions from reassembly congestion events.

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 <https://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 30 September 2022.

Copyright Notice

Copyright (c) 2022 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 (<https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Terminology . . . . .	3
3. Common Use Cases . . . . .	4
4. IPv6 Fragmentation . . . . .	4
5. IPv6 Fragment Retransmission . . . . .	5
6. Packet Too Big (PTB) Soft Errors . . . . .	8
7. Implementation Status . . . . .	9
8. IANA Considerations . . . . .	9
9. Security Considerations . . . . .	10
10. Acknowledgements . . . . .	10
11. References . . . . .	10
11.1. Normative References . . . . .	10
11.2. Informative References . . . . .	11
Author's Address . . . . .	11

## 1. Introduction

Internet Protocol version 6 (IPv6) [RFC8200] provides a fragmentation and reassembly service similar to that found in IPv4 [RFC0791], with the exception that only the source host (i.e., and not routers on the path) may perform fragmentation. When an IPv6 packet is fragmented, the loss unit (i.e., a single IPv6 fragment) becomes smaller than the retransmission unit (i.e., the entire packet) which even under moderate loss conditions could result in cascading reassembly failures that degrade forward progress [RFC8900].

The presumed drawbacks of fragmentation are tempered by the fact that performance increases can often be realized when the source sends packets larger than the path MTU. This is due to the fact that larger packets result in fewer application system calls, plus transmission of a single large packet results in a burst of multiple IPv6 fragments separated by minimal inter-packet delays. These bursts yield high network utilization for the burst duration, while modern reassembly implementations have proven capable of accommodating the bursts. If the loss unit can somehow be made to match the retransmission unit, the performance benefits of IPv6 fragmentation can be realized.

This document therefore proposes an IPv6 fragment retransmission service where the source marks fragments as retransmission-eligible while the destination may request retransmission of lost fragments. The service provides opportunistic best-effort retransmissions over an imaginary "link" extending from the source to the destination consistent with the Automatic Repeat Request (ARQ) function of common data links [RFC3366]. The service does not attempt to replace true end-to-end reliability, but instead often allows the destination to recover missing individual fragments of partial reassemblies before true end-to-end timers would cause retransmission of the entire packet.

The original packet source may be either co-located with or many IP network hops before the IPv6 fragmentation source. In the same fashion, the IPv6 reassembly destination may be either co-located with or many IP network hops before the final destination. When conditions suggest that an original source should begin sending smaller packets, the fragmentation source and/or reassembly destination can return a new type of ICMPv6/ICMPv4 Packet Too Big (PTB) message termed a PTB "soft error".

PTB "soft errors" are distinguished from classic "hard errors" by a non-zero PTB Code (ICMPv6) or unused (ICMPv4) field value. The fragmentation source can return rate-limited soft errors to recommend smaller packet sizes to the original source while fragmentation of large packets is producing excessive numbers of fragments. Similarly, the reassembly destination can return rate-limited soft errors (i.e., via the fragmentation source to the original source) while reassembly of large packets is causing excessive reassembly congestion. Original sources that receive these soft errors should reduce their packet sizes until the errors subside, but can begin to increase packet sizes again without delay until further soft or hard errors arrive.

The following sections discuss common use cases and operational considerations for applying IPv6 fragment retransmission and path MTU discovery soft errors. They further specify new codings for the IPv6 fragment header Reserved field, a new ICMPv6 message type and updates to ICMPv6/ICMPv4 PTB messages. This document therefore updates existing standards where necessary.

## 2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119][RFC8174] when, and only when, they appear in all capitals, as shown here.



- \* Fragment Offset is a 13-bit field that provides the offset (in 8-octet units) of the data portion that follows from the beginning of the packet.
- \* Res is a 2-bit field set to 0 on transmission and ignored on reception.
- \* M is the "More Fragments" bit telling whether additional fragments follow.
- \* Identification is a 32 bit numerical identification value for the entire IPv6 packet. The value is copied into each fragment of the same IPv6 packet.

The fragmentation and reassembly specification in [RFC8200] can be considered as the standard method which adheres to the details of that RFC. This document presents an enhanced method that allows for retransmissions of individual fragments.

## 5. IPv6 Fragment Retransmission

Fragmentation implementations that follow this specification reuse the (formerly) Reserved field of the IPv6 Fragment Header. For first fragments (i.e., those with zero Fragment Offset) the 8-bit Reserved field is replaced with a 7-bit Parcel ID followed by a 1-bit A(RQ) flag, and the 2-bit Res field is replaced with a 1-bit P(parcel) flag followed by a 1-bit S(ub-parcels) flag as shown below:

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Next Header | Parcel ID |A|      Fragment Offset      |P|S|M|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Identification                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

For non-first fragments (i.e., those with non-zero Fragment Offset), the Reserved field is replaced with a 7-bit "Ordinal" field followed by a 1-bit A(RQ) flag as shown below:

```

+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Next Header | Ordinal  |A|      Fragment Offset      |Res|M|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                     Identification                                     |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

When a source that follows this specification fragments an IPv6 packet it sets the first fragment A flag to 1, then for IP parcels sets Parcel ID, P and S according to the processing and transmission procedures found in [I-D.templin-intarea-parcels] and [I-D.templin-6man-omni]. For non-parcels, the source instead sets Parcel ID, P and S to 0.

The source then sets the Ordinal value for each successive non-first fragment to a monotonically-increasing value beginning with 1, i.e., it sets Ordinal to '1' for the first non-first fragment, '2' for the second non-first fragment, '3' for the third non-first fragment, etc. up to either Ordinal '127' or the final fragment (whichever comes first) while also setting the A flag to 1. (If there are additional non-first fragments beyond Ordinal '127', the source instead sets their Ordinals to '0' to indicate that the fragment is not eligible for retransmission.)

When a destination that follows this specification receives IPv6 fragments with the A flag set, it infers that the source participates in the protocol and maintains a checklist of all Ordinal fragments received for a specific Identification number. (Note that receipt of any IPv6 fragments with the A flag set provides an implicit assertion that any lost Ordinals of the same packet are also eligible for retransmission.)

If the destination notices one or more Ordinals missing after most other Ordinals for the same Identification have arrived, it can prepare an ICMPv6 Fragmentation Report (FRAGREP) message [RFC4443] to send back to the source. The message is formatted as follows:

[illegible]

In this format, the destination prepares the FRAGREP message as a list of 20-octet (Identification(i), Bitmap(i)) pairs. The first 4 octets in each pair encode the Identification value for the IPv6 packet that is subject of the report, while the remaining 16 octets encode a 128-bit Bitmap of Ordinal fragments received for this Identification. For example, if the destination receives the first fragment (i.e., Ordinal number 0) plus non-first fragment Ordinals 1, 3, 4, 6, and 8 it sets Bitmap bits 0, 1, 3, 4, 6 and 8 to '1' and sets all other bits to '0'. The destination may include as many (Identification, Bitmap) pairs as necessary without causing the entire message to exceed the minimum IPv6 MTU (i.e., 1280 octets); if additional pairs are necessary, the destination may prepare and send multiple messages.

The destination next transmits the FRAGREP message to the IPv6 fragment source. When the source receives the message, it examines each entry to determine the per-Identification Ordinal fragments that require retransmission. For example, if the source receives a Bitmap for Identification 0x12345678 with bits 0, 1, 3, 4, 6 and 8 set to '1', it would retransmit Ordinal fragments (0x12345678, 2), (0x12345678, 5) and (0x12345678, 7).

This implies that the source should retain a cache of recently transmitted fragments for a time that determines "link persistence" [RFC3366]. The link persistence should be at least as long as the round-trip time from the fragmentation source to the reassembly destination, plus an additional small delay to allow for processing overhead and/or delay variance. Then, if the source receives a FRAGREP message requesting retransmission of one or more Ordinals, it can retransmit any still in its cache. Otherwise, the Ordinal will incur a cache miss and the original source will eventually retransmit the original packet in its entirety. After processing all entries in the FRAGREP, the source discards the message.

The maximum-sized IPv6 packet that a source can submit for fragmentation is 65535 octets, and the minimum IPv6 path MTU is 1280 octets. Assuming the minimum IPv6 path MTU as the nominal size for non-final fragments, the number of Ordinals for each IPv6 packet should therefore easily fit within the available Bitmap bits when the fragments are transmitted over IPv6-only network paths. However, when the path may traverse one or more IPv4 networks (e.g., via tunneling) the path MTU may be significantly smaller. In that case, the number of IPv6 fragments needed may exceed the maximum number of Ordinal retransmission candidates.

When the number of IPv6 fragments exceeds 128, the source assigns an Ordinal value in the first 127 non-first fragments, but sets Ordinal to 0 in any remaining non-first fragments then transmits all

fragments. When the destination receives the fragments, it may return a FRAGREP to request retransmission of the first fragment and/or any missing Ordinal non-first fragments, but may not request retransmission of non-first fragments with zero Ordinals for which the default behavior of best-effort delivery applies. However, all fragments are presented equally to the reassembly cache regardless of the (formerly) Reserved field settings, where the Reserved values are ignored and successful reassembly is likely.

Finally, transmission of IPv6 fragments over IPv6-only paths can be safely conducted without a fragmentation-layer integrity check since IPv6 includes reassembly safeguards and a 32-bit Identification value. Conversely, transmission of IPv6 fragments over IPv4-only or mixed IPv6/IPv4 paths requires a fragmentation-layer integrity check inserted by the source before fragmentation and verified by the destination following reassembly since IPv4 provides only a 16-bit Identification and no reassembly safeguards. (In cases where the full path cannot be determined a priori, an integrity check should always be included as specified in AERO [I-D.templin-6man-aero] and OMNI [I-D.templin-6man-omni].)

## 6. Packet Too Big (PTB) Soft Errors

When an IPv6 fragmentation source forwards packets that produce what it considers as excessive numbers fragments (e.g., 32, 48, 64, more), the fragmentation source can also return PTB "soft errors" to the original source (subject to rate limiting). Either the fragmentation source or reassembly destination may also return PTB soft errors if the frequency of retransmissions or reassembly failures exceeds acceptable thresholds.

PTB soft errors are distinguished from ordinary "hard errors" through non-zero values in the ICMPv6 "Code" [RFC8201][RFC4443] or ICMPv4 "unused" [RFC1191] fields. The following values are currently defined:

- \* 0 - "PTB hard error" - Original sources that receive these messages obey the classic Path MTU Discovery (PMTUD) specifications found in [RFC8201][RFC1191].
- \* 1 - "PTB soft error (packet lost)" - Original sources that receive these messages should reduce their packet sizes while retransmitting the lost packet data, but need not wait the prescribed 10 minutes before attempting to again increase packet sizes.



- \* 2 - "PTB soft error (packet forwarded)" - Original sources that receive these messages should reduce their packet sizes without invoking retransmission, and also need not wait the prescribed 10 minutes before attempting to again increase packet sizes.
- \* 3-255 - reserved for future use.

PTB soft errors include as much of the invoking packet as possible without the message exceeding the minimum MTU (i.e., 1280 octets for IPv6 or 576 octets for IPv4). Original sources that recognize PTB soft errors should follow common logic to dynamically tune their packet sizes to obtain the best performance. In particular, an original source can gradually increase its packet sizes while PTB soft errors are suppressed then again reduce packet sizes when excessive soft errors arrive.

Original sources that do not recognize PTB soft errors (i.e., that do not examine the Code/unused field value) follow the same standards as for hard errors as described above and may therefore miss performance improvement opportunities.

## 7. Implementation Status

TBD.

## 8. IANA Considerations

A new ICMPv6 Message Type code for "Fragmentation Report (FRAGREP)" is requested. The registration procedure is "IETF Review" and the reference is this document [RFCXXXX].

The IANA is instructed to create new registries for "ICMPv6 Packet Too Big Code field" and "ICMPv4 Fragmentation Needed unused field" values. Both registries should have the following initial values:

Value	Sub-Type name	Reference
-----	-----	-----
0	PTB hard error	[RFCXXXX]
1	PTB soft error (loss)	[RFCXXXX]
2	PTB soft error (no loss)	[RFCXXXX]
3-252	Unassigned	
253-254	Reserved for Experimentation	[RFCXXXX]
255	Reserved by IANA	[RFCXXXX]

Figure 1: Packet Too Big Code/unused Values

## 9. Security Considerations

Communications networking security is necessary to preserve confidentiality, integrity and availability.

## 10. Acknowledgements

This work was inspired by ongoing AERO/OMNI/DTN investigations along with recent innovations with IP Parcels.

.

## 11. References

### 11.1. Normative References

- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, DOI 10.17487/RFC0791, September 1981, <<https://www.rfc-editor.org/info/rfc791>>.
- [RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <<https://www.rfc-editor.org/info/rfc1191>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <<https://www.rfc-editor.org/info/rfc4443>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <<https://www.rfc-editor.org/info/rfc8201>>.

## 11.2. Informative References

- [I-D.templin-6man-aero]  
Templin, F. L., "Automatic Extended Route Optimization (AERO)", Work in Progress, Internet-Draft, draft-templin-6man-aero-40, 7 March 2022, <<https://www.ietf.org/archive/id/draft-templin-6man-aero-40.txt>>.
- [I-D.templin-6man-omni]  
Templin, F. L., "Transmission of IP Packets over Overlay Multilink Network (OMNI) Interfaces", Work in Progress, Internet-Draft, draft-templin-6man-omni-55, 7 March 2022, <<https://www.ietf.org/archive/id/draft-templin-6man-omni-55.txt>>.
- [I-D.templin-intarea-parcels]  
Templin, F. L., "IP Parcels", Work in Progress, Internet-Draft, draft-templin-intarea-parcels-09, 10 February 2022, <<https://www.ietf.org/archive/id/draft-templin-intarea-parcels-09.txt>>.
- [RFC2473] Conta, A. and S. Deering, "Generic Packet Tunneling in IPv6 Specification", RFC 2473, DOI 10.17487/RFC2473, December 1998, <<https://www.rfc-editor.org/info/rfc2473>>.
- [RFC3366] Fairhurst, G. and L. Wood, "Advice to link designers on link Automatic Repeat reQuest (ARQ)", BCP 62, RFC 3366, DOI 10.17487/RFC3366, August 2002, <<https://www.rfc-editor.org/info/rfc3366>>.
- [RFC8900] Bonica, R., Baker, F., Huston, G., Hinden, R., Troan, O., and F. Gont, "IP Fragmentation Considered Fragile", BCP 230, RFC 8900, DOI 10.17487/RFC8900, September 2020, <<https://www.rfc-editor.org/info/rfc8900>>.

## Author's Address

Fred L. Templin (editor)  
Boeing Research & Technology  
P.O. Box 3707  
Seattle, WA 98124  
United States of America  
Email: [fltemplin@acm.org](mailto:fltemplin@acm.org)

IPv6 Maintenance (6man) Working Group  
Internet Draft  
Updates: 4861, 4862 (if approved)  
Intended status: Standards Track  
Expires: September 2022

E. Vasilenko  
P. Volpato  
Huawei Technologies  
Olorunloba Olopade  
Virgin Media  
March 4, 2022

ND Prefix Robustness Improvements  
draft-vv-6man-nd-prefix-robustness-02

Abstract

IPv6 prefixes could become invalid abruptly as a result of outages, network administrator actions, or particular product shortcomings.

That could lead to connectivity problems for the hosts attached to the subtended network.

This document has two targets: on one hand, to analyze the cases that may lead to network prefix invalidity; on the other to develop a root cause analysis for those cases and propose a solution.

This may bring to extensions of the protocols used to convey prefix information and other options.

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 <https://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 2022.

## Copyright Notice

Copyright (c) 2022 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. Terminology and pre-requisite.....	3
2. Introduction.....	4
3. Problem Scenarios.....	4
3.1. Reference architectures.....	5
3.2. Discussion on the scenarios.....	5
3.2.1. Non-graceful reload due to unexpected events .....	5
3.2.2. Graceful reload without precautions .....	6
3.2.3. Abrupt hardware replacement without the possibility for graceful prefix deprecation.....	7
3.2.4. Non-graceful configuration change .....	8
3.2.5. An uplink breaks connectivity without a relevant notification to the connected hosts.....	8
4. Root cause analysis.....	10
4.1. What to protect.....	10
4.2. Where to protect.....	12
4.3. When to protect: technology scenarios.....	12
5. Solutions.....	13
5.1. Multi-homing multi-prefix (MHMP) environment.....	13
5.2. A provider is not reachable in MHMP environment.....	16
5.3. Administrator abruptly replaces PA prefix.....	17
5.4. Planned router outage.....	18
5.5. Prefix information lost because of abrupt router outage..	19
5.6. Prefix information lost after hardware replacement.....	19
5.7. Link layer address of the router should be changed.....	20
5.8. Dependency between solutions and extensions.....	20
6. Extensions of the existing standards.....	20

6.1. Default router choice by host.....	20
6.2. Prefixes become dynamic.....	21
6.3. Do not forget to deprecate prefixes on renumbering.....	22
6.4. Do not forget to deprecate prefixes on shutdown.....	23
6.5. Store prefixes in non-volatile memory.....	23
6.6. Find lost information by "Synchronization".....	24
6.7. Default router announcement rules.....	26
6.8. Faster detection of the stale default router.....	26
6.9. Clean orphaned prefixes after default router list change.	27
7. Interoperability analysis.....	27
8. Applicability analysis.....	28
9. Security Considerations.....	28
10. IANA Considerations.....	29
11. References.....	29
11.1. Normative References.....	29
11.2. Informative References.....	30
12. Acknowledgments.....	31

## 1. Terminology and pre-requisite

[ND] and [SLAAC] are pre-requisite to understand this document.  
The terms are inherited from these standards.

Additional terms:

Home Gateway - a small consumer-grade router that provides network access between hosts on the local area network (LAN) and the Internet behind the wide area network (WAN)

PA - Provider-Aggregatable addresses leased to the client or subscriber

MHMP - Multi-Homing Multi-Prefix. An environment with hosts connected to different PA providers (multi-homing) through different address spaces announced from different providers (multi-prefix)

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 2. Introduction

It has been reported that some number of cases could lead to loss of information (primarily prefixes) by [ND]. Current [ND] protocol's default timers may lead to many days of outage for hosts. This is not acceptable.

This document analyses all potential cases when an outage could happen and proposes solutions. Discussion is restricted to potential [ND] extensions only.

MHMP environment has been considered. It has been discovered that [ND] problems could be isolated from the overall complex [MHMP] environment, and could be fixed separately.

The document is organized to introduce, in section 3, the scenarios where the issue of prefix invalidity may happen and the cases of invalidity.

Section 4 provides a root cause analysis for the cases of invalidity and identifies the corner-cases which are subject of our discussion.

Section 5 proposes a solution for the cases identified.

Section 6 brings the discussion forward, proposing extensions to [ND].

## 3. Problem Scenarios

[ND] distributes prefixes as Prefix Information Options (PIOs) in Router Advertisements (RA) messages from routers.

Once a router assigns a prefix to a host, this prefix is assumed to be stable so that hosts can employ it to configure the IPv6 addresses associated with their interfaces [SLAAC] or to forward packets to the network.

Prefix changes may happen and are governed by the rules of [ND], [SLAAC].

Yet, cases exist where prefix instability may happen. An example is provided by the so-called "flash-renumbering" event: when flash-renumbering happens a network prefix in use suddenly becomes invalid because it is replaced by a new prefix.

The router causing or forced to cause the network renumbering may not be able to cope with the effects of this sudden change (for example, deprecating the previously assigned prefixes). Another possibility is that the subtended hosts do not have the means of overcoming the effects of renumbering.

This section describes problems that were found in live networks. Most of the information in this section comes from [Flash-Renumbering], [SLAAC Robustness]. Their contributions are greatly acknowledged.

### 3.1. Reference architectures

Home broadband networks, SOHO (Small Office Home Office) networks are the typical scenarios affected by renumbering. Some problems discussed below applicable on the more general basis.

In typical case a router (e.g. Home Gateway, Customer Premise Equipment (CPE), Customer Edge (CE), etc.) is deployed to provide connectivity to a Service Provider network for the attached devices. A second router may be deployed for redundancy, especially for business scenarios.

Two reference architecture can be considered:

Architecture #1. Hosts are directly connected to the router. For example, a Home Gateway embeds the functions of L2 device (Ethernet switch, WiFi AP) and L3 device (router).

Architecture #2. Hosts connect to an intermediate L2 device (e.g. a wired Ethernet switch or a Wi-Fi access point) that, in turn, connects to the router (or routers, if uplink redundancy is requested).

### 3.2. Discussion on the scenarios

The discussion provided here is introductory to both the root cause analysis provided in section 4. and the solutions proposed in section 5.

#### 3.2.1. Non-graceful reload due to unexpected events

A router could be reloaded abruptly for many reasons: hardware or software bug, power outage, manual intervention. This last one is very probable for home broadband subscribers that tend to fix every problem with power recycle.



Usually, it does not create additional problems for [ND] and [SLAAC] because the same PIO information would be advertised by the router in RA messages after each reload. In such cases, a Home Gateway would initialize its Ethernet and WiFi connections, clearing all stale information on directly connected hosts.

It should not create problems for proper home network design where all CPEs are routers - see [HomeNet Architecture]. The delegated prefix would not be changed in the case of subtended CPE reload. Prefix change in the case of upstream CPE reload should be properly discontinued by subtended CPE. There is the need for a special protocol for prefix distribution that is out of the scope of this document - see [HNCP].

For architecture #2 implemented in home environments, there is a corner case when Home Gateway's abrupt reload would not be visible to hosts connected to subtended "bridged" CPE. If it would coincide with the situation when a different prefix would be delegated from Carrier (at 37% probability according to [Residential practices]), it would lead to the situation that hosts would receive a new prefix without deprecation of the previous one. Hosts do not have any standard mechanism to choose only the new prefix for communication. That would lead to a connectivity problem.

How long a non-preferred prefix would be kept in a stale state on the host is not important (default AdvValidLifetime is 30 days in section 6.2.1 of [ND]), because according to [Default Address] section 5 rule#3, it should have a lower priority to be chosen. [SLAAC] section 5.5.4 is another good reference highlighting that address should be avoided after it would reach the deprecated status.

How long an address would stay in the preferred state is important. [ND] instructs hosts to prefer certain prefix for 7 days - see default AdvPreferredLifetime in section 6.2.1.

It is not realistic for the subscriber to wait for 7 days.

It practically means that the subscriber in this corner case would have a few options to fix the problem: (1) reload all hosts, or (2) reconnect the physical link of every host, or (3) reload subtended bridge, or (4) manually delete the prefix on the hosts to clear stale information.

### 3.2.2. Graceful reload without precautions

Specifically this scenario may happen when developers don't apply precautions in case previous prefixes are not deprecated. It may happen in both architectures.

The router could be reloaded by graceful procedure (reboot or shutdown that would use "init 6" in Unix). It is still possible that software would not send RA with prefix Preferred Lifetime zero to inform hosts about prefix deprecation. This practice prevails because IPv4's centralized address assignments by DHCP does not need similar precautions.

Again, like in the previous section, it would not create a problem in the majority of the cases for directly connected hosts (architecture #1) because link layer would be reinitialized too. The same corner case (architecture #2) would lead to the same result: a connectivity problem that could be resolved only by 4 types of manual intervention mentioned in the previous section.

### 3.2.3. Abrupt hardware replacement without the possibility for graceful prefix deprecation

Such type of an outage is again may happen only for architecture #2. It would lead to up to 30 minutes (including time for hardware replacement) outage in all cases (to detect missing router) and up to 1 week additional outage if a different prefix would be announced after the hardware replacement.

The hardware could fail or be replaced with an abrupt power disconnect. The latter is very probable for the home environment. Graceful notification of hosts may not happen.

The new hardware may have a different link layer address and a different link local address as a result. The router would look like a new one on the link. Any communication with it could not be the reason to deprecate announcements made early by the router perceived as a different one.

[ND] section 6.2.1 has recommended the AdvDefaultLifetime as  $3 * \text{MaxRtrAdvInterval}$ . Hosts would send traffic to a non-existent router for up to 30 minutes.

According to section 4.2 of [ND] "Router Lifetime" is related only to router default status. PIO announced early may be preferred up to 7 days according to AdvPreferredLifetime in section 6.2.1 of [ND] even after the router default status is deprecated. The probability for such a situation is the same low as discussed in section 3.2.1. because a different prefix should be announced after hardware reload and a switch should be present between the host and the router. The same corner case would lead to the same result: a connectivity

problem that could be resolved only by 4 types of manual intervention mentioned in section 3.2.1. .

#### 3.2.4. Non-graceful configuration change

This situation may happen due to abrupt prefix change on the router (in both architectures) or VLAN change on the switch (it may happen in architecture #2).

Router configuration could be changed manually, by automation tools, or by protocols (for example, prefix distribution).

Additionally for architecture #2, L2 domain could be abruptly changed by configuration (for example, VLAN change from "quarantine" to "production" without any chance for the router to send a message).

It could lead to the situation that prefix would change abruptly, without any notification to hosts about the necessity to deprecate the previous prefix. Hosts should be notified by prefix announcement with Preferred Lifetime set to zero.

It should not happen for residential CPE because [CPE Requirements] section 4.3 requirement L-13 clearly instructs: "If the delegated prefix changes, i.e., the current prefix is replaced with a new prefix without any overlapping period of time, then the IPv6 CE router MUST immediately advertise the old prefix with a Preferred Lifetime of zero".

But it is perfectly possible for other environments (except residential CPEs) because other routers are not required to do the same: [Node Requirements] does not clarify the exact router behavior in the case of abrupt prefix change. [SLAAC] does not have any recommendations either.

#### 3.2.5. An uplink breaks connectivity without a relevant notification to the connected hosts

It may happen in both architectures #1 and #2.

A router could lose uplink. The probability for such an event is much bigger for a mobile uplink (modem). It would invalidate the possibility to use a PA prefix advertised from this carrier even in the case that another carrier uplink is available on this or redundant router (connectivity to the Internet is not lost). Some mechanism is needed to inform hosts not to use address space from

the disconnected carrier because another carrier would filter it out by anti-spoofing security protection.

The multi-homing multi-prefix PA environment has been properly explained in [MHMP]. The discussion of how traffic should be source-routed by routers in [MHMP] environment is not relevant to our [ND] discussion. Unfortunately, an improper address used as a source would cause a traffic drop as soon as traffic gets to the different carrier.

[Default Address] section 5 (source address selection) rule 5 (for different interfaces on the host) and rule 5.5 (for the same interface) partially prepare hosts for such situation: "Prefer addresses in a prefix advertised by the next-hop. If SA or SA's prefix is assigned by the selected next-hop that will be used to send to D [...] then prefer SA". This algorithm has an assumption that the source address should be chosen after the next hop.

Unfortunately, the rules mentioned above in [Default Address] section 5 would work only if the default router would cease to be default after it loses route to its carrier. It would work only in simplified topology where all hosts connect by L2 to different CPEs, each leading to its separate carrier prefix. It could be called a "common-link environment for all hosts and routers". It is not possible in practice because hosts on the most popular link layer technology (WiFi) are rooted to only one CPE (with AP inside) - they would not switch automatically to different CPE where the Internet connectivity may be still available.

[CPE Requirements] have G-3/4/5 specifically for this simplified multi-homing residential design. It recommends announcing Router Lifetime as zero on LAN if CPE does not have "default router from the uplink" - it would push the host to use another source address by the mentioned above source address selection algorithm.

It is not explained in [CPE Requirements] what should happen with PA delegated prefix after the respective uplink is disconnected. Probably, this is because it was not needed to deprecate stale prefix for the above mentioned mechanism (based on default router withdrawal) to work.

The local residential network could be left without any default router as a result of using the above mechanism - it is especially probable in the single CPE environment. Hence, [CPE Requirements] promotes [ULA] addresses for local connectivity. Default router functionality is returned specifically for [ULA] addresses by

requirement L-3: use "Route Information Option" from [Route Preferences]. It needs hosts' participation in routing through the RIO option.

Unfortunately, this long chain of fixes explained above is strictly optimized for the environment "common-link for all hosts and routers". It is not the case for single WiFi inside any CPE or other topologies.

Neither [ND] nor [SLAAC] instruct the router what to do when the PA delegated prefix is withdrawn abruptly.

[Multi-Homing] section 3 has a good discussion about the proper relationship between default routers and prefixes advertised by respective routers in a stable situation. This would be discussed in more details in section 5.1. . [Multi-Homing] does not discuss what to do in the situation when the router is available, but some uplinks (with delegated prefixes) are lost.

[MHMP] discusses the problem in deep detail with two tools proposed to regulate [ND] behavior: [Policy by DHCP] to change [Default Address] algorithm and [Route Preferences] to inform about appropriate exit points. There are more details later in section 5.1.

#### 4. Root cause analysis

Let's further analyze to be sure that all corner cases are found.

It is assumed in all discussions below that [RA-Guard] is implemented, and all messages are from routers under legitimate administrative control. Security issues are considered as resolved by [RA-Guard], and possibly with extensions in [RA-Guard+].

DHCP is almost as vulnerable as SLAAC for cases found below. DHCP's typical lease time (hours) is shorter than SLAAC's prefix lifetime (days), but is too long for users to accept self-repairing time. Root cause analysis below applies to all possible environments: DHCP, SLAAC, and mixed.

##### 4.1. What to protect

[ND] Router Advertisements deliver configuration information to hosts. Such information could become inaccurate in two different periods of time:

- a) "Recoverable". Time is needed for some process to finish and update information (example: router reload or uplink re-connect).
- b) "Non-recoverable". Time, dependent on some timer expiration (example: complete loss of prefix or default router).

A careful look at the information distributed by RA would give us the understanding that the most problematic is the information that is already protected by deprecation timers: Prefix Information Option and Default Router. Section 3 discusses that the handling of this information is still susceptible to recoverable and non-recoverable periods of inaccuracy.

For example, in the case of abrupt router reload described in sections 3.2.1. -3.2.3. , the recoverable part is the time spent by router and hosts to update their cache after the router reload. The non-recoverable part is related to the setting of the AdvPreferredLifetime timer which would probably force a user to solve the issue with manual intervention.

The next problematic case is the abrupt change of source link-layer address. This problem is not discovered yet in production because it has a low probability. Indeed, a router with a different link-layer address would be treated as a new router, the old router would just disappear from the link. It would affect primarily default router information because all other information should be immediately re-advertised from the new link layer address. Section 6.2.8 of [ND] already discusses how to properly deprecate the default router status of the old link layer address, but no recommendation is given in [ND] for prefix deprecation in this situation. A corner case is possible that software would not treat the new virtual interface as identical concerning the prefix information that should be announced. Different prefixes may be announced. Some additional precautions are needed.

Other information in RA (Hop Limit, MTU, DHCP flags, Reachable timer, and Retransmit timer) are not so sensitive because (1) it is typically static and (2) it does not affect connectivity for respective parameters change in the wide range.

Flag "A" in PIO deserves special attention. It could be cleared abruptly (signaling that hosts should not use this prefix for [SLAAC] anymore). That should not create any problem, because the prefix is still available from a respected PA provider - traffic could be routed to the global Internet. Therefore, it is not vitally important for the host to immediately deprecate the address from

this prefix.

A similar situation is with flag "M" in RA: DHCP address should be deprecated. It should not create a connectivity problem because prefixes could be routed to the global Internet.

#### 4.2. Where to protect

[ND] is the protocol for first-hop connection between host and router. It is designed for one link only. One link could have more than one router.

It is assumed below that a more complex topology (many other routers) is shielded from this link by some other protocol that would deliver all necessary information to those routers.

[HomeNet Architecture] discusses many types of information that should be distributed to every home router. Let's focus on delegated prefixes for our discussion.

The number of uplinks on every router is not important, as long as proper information about prefixes is up to date on the router.

Hence, all our topologies could be simplified into the following scenarios:

- I. L2 device (switch, WiFi AP) and L3 device (router) are in the same device (sharing the fate for power, reboot) (refer to architecture #1 in section 3.1. ).
- II. Separate L2 device (probably a switch) and an arbitrary number of L3 devices (routers) are connected to the same IPv6 link (refer to architecture #2 in section 3.1. ).

#### 4.3. When to protect: technology scenarios

Let's reorder scenarios discussed in section 3. in the way that it would be better to map to the technology modifications and account for some corner cases found in root cause analysis:

1. Proper prefix usage for Multi-Homing Multi-Prefix environment.  
Hosts should be capable of choosing in a coordinated way  
(1) a source address (from proper PA prefix) and (2) a next hop:
  - A.1. In a normal situation: all providers and prefixes are available

A.2. In a faulty situation: one provider is not reachable, but some hosts and links on the routed path to this provider may still be reachable

A.3. In the case when an administrator abruptly replaces delegated prefix

## 2. Proper prefix usage for the case of router outage that:

A.4. Planned for this interface  
(reboot, shutdown, or ceasing to be a router)

A.5. Abrupt (power outage, software or hardware bug)

A.6. Abrupt (power outage, hardware fault) with hardware replacement

## 3. Proper prefix usage for the case of link layer address of the router.

These cases are discussed from section 5.1. to section 5.6.

There is no big difference for [ND] between ULA and GUA at the considered link because both could be disjoined at any routed hop upstream. It would need the same invalidation mechanisms on the link. ULA could be invalidated too for the case that ULA spans many sites in a big company. The residential network would probably have a separate ULA for every household that would decrease the probability of ULA prefixes invalidation. It is the responsibility of another protocol (for example, [HNCP]) to decide when ULA should be invalidated, if ever.

## 5. Solutions

Let's look at the solutions for scenarios listed in section 4.3.

### 5.1. Multi-homing multi-prefix (MHMP) environment

Let's consider here host capability to choose a proper PA prefix and next hop router in a stable multi-homing multi-prefix (MHMP) environment.



The complex MHMP situation is properly discussed in [MHMP] section 3.1 - it is critical to read it to understand the rest of this section. Our discussion is restricted to [ND] protocol only (one link) - it would cut the number of topologies discussed in section 4.2. MHMP may need additional complex routing interactions that are out of the scope of this document.

It is possible to introduce one additional classification to clearly separate what it is possible to implement now from what needs additional standardization efforts:

1. Case "equal prefixes": Announced prefixes are fully equal by scope and value, all resources interested for hosts could be reachable through any announced PA prefix; additionally, traffic distribution between carriers could be round-robin (no any traffic engineering or policing).
2. Case "non-equal prefixes": Announced prefixes are not equal because (1) some resources could be accessed only through a particular prefix (for example walled garden of one carrier) or (2) it is desirable to have some policy for traffic distribution between PA prefixes (cost of traffic, delay, packet loss, jitter, proportional load).

There are two reminders before the discussion of the above cases:

- o [ND] section 6.3.6 recommends next hop choice between default routers in a round-robin style. Traffic policy or even reachability of particular resources through a particular default router is not considered at the [ND] level.
- o [Default Address] section 7 assumes that source and destination address selection should happen after the next hop (or interface) choice by [ND] or routing, source address is chosen after this.

Case "equal prefixes" does not create any requirement on what prefix should be used for the source address. It is only needed that the source address would be chosen to be compatible with the next hop that should be in the direction of the respective carrier.

No problem is possible for the topology with only one router on the link. The router itself may need source routing to choose next hop properly but it is out of the scope of ND protocol and this document.

Host on a multi-homing link would better be compliant to [Default Address] section 5 (source address selection) rule 5 (for different interfaces on the host) or rule 5.5 (for different next hops on the

same interface). It would help to properly choose a source address compliant to the next hop chosen first. Moreover, if the source address would be chosen wrongly then it is still possible to reroute the packet later by source routing. Hence, it is possible to satisfy the "equal prefixes" case on the current level of standardization developed.

Case "non-equal prefixes" is more complicated. It would be too late to try to solve this problem on a router, because the wrong source address may be already chosen by the host - it would not be possible to contact the appropriate resource in the "walled garden". Only NAT could be left as an option, but that is not a valid choice for IPv6.

There are 2 methods to resolve the case of "non-equal prefixes":

1. The same policies could be formatted differently and fed to the host by two mechanisms: 1) "Routing Information Options" of [Route Preferences] and 2) [Policy by DHCP] to modify policies in [Default Address] selection algorithm. Then current priority of mechanisms could be preserved the same: initially [ND] or routing would choose the next hop, then [Default Address] would choose a source address (and destination if multiple answers from DNS are available). It is the method that is assumed in [MHMP].
2. Alternatively, policies could be supplied only by [Policy by DHCP] to [Default Address] selection algorithm. [Default Address] discusses potential capability in section 7 to reverse algorithm's order: source address may be chosen first, only then to choose next hop (default router). Source address selected from proper carrier is potentially the complete information needed for the host to choose the next hop, but not for the default round-robin distribution between available routers that specified in [ND]. [ND] extension is needed for this method for the host to prioritize default routers that have announced prefixes used for the source address of the considered flow. It is this method that is assumed in [Multi-Homing] section 3.2. This document is different in that the same rules are formulated not as the general advice, but as the particular extension to [ND] - see section 6.1 of this document.

The second method has the advantage that there is no need to download RIO policies by [Route Preferences]. It would simplify the implementation of the MHMP environment. Only the second method is universal and extendable because some policies may not be translated as RIO of [Route Preferences]. For example, dynamic policies (packet loss, delay, and jitter) could

be measured on hosts. Hence, the decision about source address and next hop should be local.

## 5.2. A provider is not reachable in MHMP environment

Let's assume the fault situation when one provider is not reachable in the [MHMP] environment. A prefix may be very dynamic for a few reasons. It could be received from some protocols (DHCP-PD, HNCP). The prefix could become invalid (at least for the global Internet connectivity) as a result of the abrupt link loss in the upstream direction to the carrier that distributed this prefix.

Additionally, consider the more complicated case when some hosts on the upstream routed path to this provider may still be reachable using a particular prefix but Internet connectivity is broken later.

Let's consider the last problem. Because Internet connectivity is lost for this prefix, it should be announced to hosts by zero Preferred Lifetime. [Route Preferences] gives the possibility to inform hosts that particular a prefix (RIO) is still available on-site but it would be an automation challenge to dynamically calculate and announce prefix. Additionally, [Route Preferences] should be supported by hosts.

In general, it is not a good idea to involve [ND] in routing. Hence, it is better to support on-site connectivity by PI GUA or ULA that may not be invalidated. There are many reasons to promote [ULA] for internal site connectivity: (1) hosts may not have GUA address at all without initial connection to the provider, (2) PA addresses would be invalidated in 30 days of disconnect anyway, (3) it is not a good idea to use addresses from PA pool that is disconnected from global Internet - hosts may have a better option to get global reachability. ULA has better security (open transport ports that are not accessible from the Internet) which is an additional bonus. It is effectively the request to join current [CPE Requirements] and [HomeNet Architecture] requirements in sections 2.2, 2.4, 3.4.2 that subscriber's network should have local ULA addresses.

Prefix deprecation should be done by RA with zero Lifetime for this prefix. It will put the prefix on hosts to the deprecated status that according to many standards ([ND], [SLAAC], and [Default Address]) would prioritize other addresses. Global communication would be disrupted for this prefix anyway. Local communication for deprecated addresses would continue till normal resolution because the default Valid Lifetime is 30 days. Moreover, if it would happen that this delegated prefix was the only one in the local network (no [ULA] for the same reason), then new sessions would be opened on

deprecated prefix because it is the only address available. If connectivity would be re-established and the same prefix would be delegated to the link - it would be announced again with proper preferred lifetime. If a different prefix could be delegated by the PA provider, then the old prefix would stay in deprecated status. It is an advantage for the host that would know about global reachability on this prefix (by deprecated status) because the host may use other means for communication at that time.

Such dynamic treatment of prefixes may have the danger of [ND] messages flood if the link on the path to PA provider would be oscillating.

[HNCP] section 1.1 states: "it is desirable for ISPs to provide large enough valid and preferred lifetimes to avoid unnecessary HNCP state churn in homes".

It makes sense to introduce dampening for the rate of prefix announcements.

Such conceptual change in the treatment of prefixes would not affect current enterprise installations where prefixes are static.

It is important to mention again that it is the responsibility of the respective protocol (that has delivered prefix to the considered router) to inform the router that prefix is not routed anymore to the respective carrier. It is easy to do it in the simplified topology when the only router could correlate uplink status with the DHCP-PD prefix delegated early. Some additional protocols like [HNCP] are needed for a more complex topology.

There is nothing in [ND] or [SLAAC] that prevents us from treating prefixes as something more dynamic than "renumbering" to reflect the dynamic path status to the PA provider. Section 6.2. proposes extensions to [CPE Requirements] and [SLAAC] that follow the logic of this section.

### 5.3. Administrator abruptly replaces PA prefix

This is the case when the network administrator (maybe from another domain) replaces prefix much faster than 2 hours or the remaining preferred lifetime (as per section 5.5.3 of [SLAAC] on router advertisement processing). The reason for abrupt replacement is probably not related to networking.

Abrupt prefix change may be caused by improper configuration, for example, VLAN change at the switch.

Standards recommend deprecating old prefixes but do not recommend for developers and system designers to additionally check abrupt

configuration changes to mitigate human mistakes. IPv4 cannot mitigate such type of mistake, IPv6 has an advantage here.

Section 6.3. proposes a recommendation for the additional check to make sure that prefix would be deprecated.

This problem could be exacerbated by the low reliability of multicast delivery in a wireless environment - the only packet sent (for example before VLAN change) could be lost. A long-term solution for this problem is proposed in section 6.6 that permits synchronizing host states with a new flag in router announcements.

#### 5.4. Planned router outage

A router could be planned to be put out of service for a link (reboot, shutdown, or ceasing to be a router).

The primary Operating System for routers is LINUX. The following discussion is based on LINUX as an example - other developers can find an analogy for their operating system.

Some LINUX shutdown commands are not graceful in principle (like Halt or Poweroff). It would need extraordinary efforts to send messages discussed in this section before the system would be stopped. It is better to restrict network administrators from such tools on routers.

Other LINUX shutdown commands are safe (Reboot is safe for a long time, Shutdown and "Init 6" have been safe). It would execute shutdown scripts that would give the developer the chance to comply with requirements in this section.

It is up to the developer how reboot and shutdown should be mapped to particular OS commands in graphical user interface (GUI), command line interface (CLI), or automation interface (Netconf/YANG), and what particular actions should be taken. It SHOULD guarantee that section 6.2.5 of [ND] with updates in section 6.4 of this document properly inform hosts that the router is going out of service.

The same procedure SHOULD be automatically activated for cases when an administrator tries manually (via CLI or GUI) or automatically (via Netcong/YANG/Other) to change Link Layer Address on this router interface or disable router functionality in [ND] for this link.

### 5.5. Prefix information lost because of abrupt router outage

PIO could be lost because of the abrupt reload - the router may not have a chance to warn hosts, but the router could receive a different prefix after reload. Reasons could be (1) power outage, (2) software bug, or (3) hardware problem.

[HomeNet Architecture] section 3.4.3 (Delegated Prefixes) has already recommended usage of non-volatile memory:  
"Provisioning such persistent prefixes may imply the need for stable storage on routing devices and also a method for a home user to 'reset' the stored prefix should a significant reconfiguration be required (though ideally the home user should not be involved at all)".

[SLAAC] section 5.7 has recommended storing acquired addresses on hosts in non-volatile memory too.

This document joins these requests and propose adding similar requirements to [CPE Requirements] and [SLAAC] - see section 6.5.

The best long-term solution is to inform the host by [ND] protocol that RA has all information in one announcement. Any missing information SHOULD be considered deprecated. It is possible to do it with the new flag in RA - see section 6.6.

"Complete" flag would become useful only when implemented on both: host and router. It is proposed to rely on storage improvements in non-volatile memory till the "Complete" flag would be supported on many hosts.

### 5.6. Prefix information lost after hardware replacement

Hardware fault or power outage may follow by hardware replacement.

Prefix storage in non-volatile memory and a "complete" flag would not protect in such a situation. The new router would not have the old prefix information and the "complete" flag would be sourced from a different LLA.

Initially, it would be good to speed up the detection of hardware replacement to delete the stale hardware from the default router list of hosts. It is proposed to request all routers availability by RS all-routers multicast address after new router detection on the link- see section 6.8. It would permit to detect that old hardware is not active in 13 seconds (see section 6.3.7 of [ND] for timers  $\text{MAX\_RTR\_SOLICITATIONS} * \text{RTR\_SOLICITATION\_INTERVAL} + \text{MAX\_RTR\_SOLICITATION\_DELAY}$ ). 13 seconds is considered a short enough outage compare to hardware replacement and reload.

Then it is proposed to detect stale prefixes at the event of the respective router deletion from the default router list. If the particular prefix is not announced anymore by any active router on the default router list then the prefix (and all associated addresses) should be deprecated - see section 6.9.

#### 5.7. Link layer address of the router should be changed

Sections 6.3 and 6.4 provide an additional check also in the case of a link layer address change. Hence, additionally resolve LLA change case.

#### 5.8. Dependency between solutions and extensions

It could be useful to map, for quick reference, the dependency between the solutions listed in this section and standard's extensions as presented in section 6.

Solution discussed in		Corresponding extension
5.1.	->	6.1.
5.2.	->	6.2. & 6.7.
5.3.	->	6.3. & 6.6.
5.4.	->	6.4.
5.5.	->	6.5. & 6.6.
5.6.	->	6.8. & 6.9.
5.7.	->	6.3. & 6.4.

### 6. Extensions of the existing standards

The solution requires a number of standard extensions. They are split into separate sections for better understanding. It is better to read references from section 5. before reading this section, see section 5.8. for cross-reference.

#### 6.1. Default router choice by host

\* Section 6.3.6 (Default Router Selection) of [ND], add an initial policy to default router selection:

- 0) For the cases when a particular implementation of ND does know the source address at the time of default router selection (it means that source address was chosen first), then default routers that advertise the prefix for respective source address SHOULD be preferred over routers that do not advertise respective prefix.

## 6.2. Prefixes become dynamic

\* This document joins the request to [CPE Requirements] that has been proposed in section 11 (General Requirements for HNCP Nodes) of [HNCP]:

The requirement L-13 to deprecate prefixes is applied to all delegated prefixes in the network from which assignments have been made on the respective interface. Furthermore, the Prefix Information Options indicating deprecation MUST be included in Router Advertisements for the remainder of the prefixes' respective valid lifetime, but MAY be omitted after at least 2 hours have passed.

\* Add section 4.2 into [SLAAC]:

### 4.2 Dynamic Link Renumbering

Prefix delegation (primarily by DHCP-PD) is adopted by the industry as the primary mechanism of PA address delegation in the fixed and mobile broadband environments, including cases of small business and branches of the big enterprises.

The delegated prefix is tied to dynamic link that has a considerable probability to be disconnected, especially in a mobile environment. The delegated prefix is losing the value if the remote site is disconnected from prefix provider - this fact should be propagated to all nodes on the disconnected site, including hosts. Information Options indicating deprecation (multicast RA with zero Preferred Lifetime) MUST be sent at least one time. It SHOULD be included in Router Advertisements for the remainder of the prefixes' respective valid lifetime but MAY be omitted after 2 hours of deprecation announcements.

There is a high probability that connectivity to the provider would be restored very soon then the prefix could be announced again to all nodes on the site.



There is the probability that in a small period of time the same problem would disconnect the site again (especially for mobile uplink). Such oscillation between available and not available provider could happen frequently that would flood the remote site with [ND] updates.

Dampening mechanism MAY be implemented to suppress oscillation: if the time between a particular prefix announcement and previous deprecation was less than DampeningCheck then delay the next prefix announcement for DampeningDelay and check the need for the prefix announcement after DampeningDelay seconds.

It is recommended for protocol designers to implement a dampening mechanism for protocols (like [HNCP]) that would be used to distribute prefix delegation inside the site to relieve the majority of site routers and the protocol itself from the processing of oscillating messages.

\* Section 5.1 (Node Configuration Variables) of [SLAAC], add timers:

DampeningCheck - the time between prefix announcement and previous deprecation is checked against this value to decide about dampening need. The timer should use 16bit unsigned integer measured in seconds. The default value is 10 seconds.

DampeningDelay - the delay (penalty) for the next attempt to announce the same prefix again. The timer should use 16bit unsigned integer measured in seconds. The default value is 10 seconds.

These timers should be configurable like all other timers in [SLAAC] section 5.1.

### 6.3. Do not forget to deprecate prefixes on renumbering

\* Section 4.1 (Site renumbering) of [SLAAC], add at the end:

A network administrator SHOULD avoid the situations when renumbering is done abruptly (with the time of transition that is less than the preferred time for the respective prefix). Situations could happen when it is not possible to archive the above-mentioned goal: (1) the prefix could be withdrawn by the administrator of another domain, (2) there could be the urgent need to change the prefix for reasons not related to networking, (3) prefix could be invalidated after some network event (example: loss of uplink that was used to receive this prefix), (4) L2 connection (VLAN or VPN) could be changed

abruptly by mistake or due to not a proper design. Prefix deprecation MUST be signaled at least one time by multicast RA with Preferred Lifetime set to zero for respective PIO. It SHOULD be included in RA for the remainder of the prefixes' respective valid lifetime but MAY be omitted after 2 hours of deprecation announcements.

It is recommended for developers to check and enforce this rule in router's software: if an administrator, automated system, or other protocol would try to delete a particular prefix from the link and if that prefix has the preferred lifetime bigger than zero, then the software MUST automatically generate deprecation announcements according to the rules explained above.

System designer SHOULD make sure that in the case of abrupt change of logical connectivity at L2 (VLAN, VPN) new default router SHOULD deprecate stale prefixes inherited from the previous default router.

#### 6.4. Do not forget to deprecate prefixes on shutdown

\* Section 6.2.5 of [ND] starts from the definition of ceasing cases for the router on [ND] link. One additional reason SHOULD be added to the end of the list:

- Link layer address of the interface should be changed.

\* Section 6.2.5 (Ceasing To Be an Advertising Interface) and Section 6.2.8 (Link Local Address Change) of [ND] already discusses requirements of proper ceasing to be [ND] router advertising interface. It has requirements to announce zero for a default router lifetime. It is proposed to add at the end of both sections:

A router MUST also announce in above-mentioned announcements all previously advertised prefixes with zero Preferred LifeTime. Valid LifeTime should not be decreased from originally intended - current hosts sessions should have the possibility to be rerouted to the redundant router (if available).

#### 6.5. Store prefixes in non-volatile memory

Add the same text:

- \* [CPE Requirements], new requirement G-6 at the end of section 4.1, and
- \* [SLAAC], at the end of section 5.7:

The IPv6 router SHOULD keep in non-volatile memory all prefixes advertised on all links, including prefixes received by dynamic protocols with the reference to the respective protocol (DHCP-PD, HNCP, others).

A router could experience a non-graceful reload.

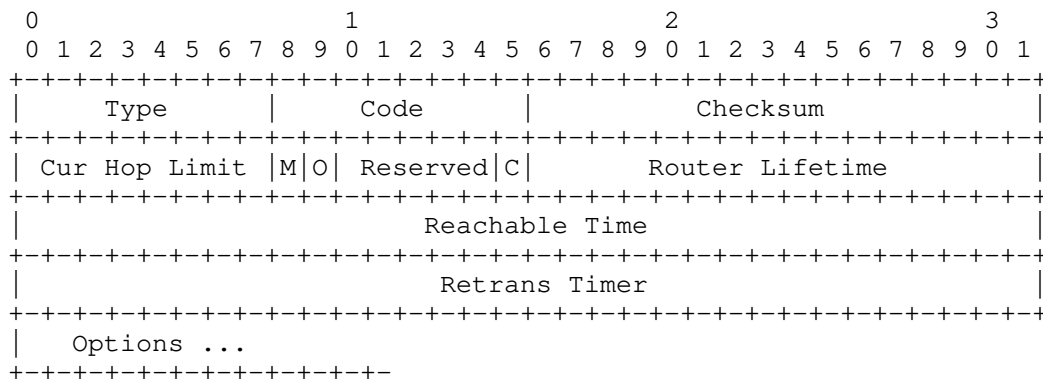
If another protocol would delegate any prefixes for router links then the router SHOULD immediately start announcing them in the normal way.

Additionally, the router should wait until the end of convergence for the respective prefix-delegation protocol. The way for how to decide that convergence is finished is the responsibility of the respective protocol design. It could be a simple timer after uplink would go to "up" or successful exchange by some protocol (like DHCP-PD).

If another protocol would not delegate prefix recorded in non-volatile memory after assumed convergence is achieved, then the old prefix MUST be announced on the link at least one time by multicast RA with the zero Preferred Lifetime. It SHOULD be included in RA for the remainder of the prefixes' respective valid lifetime but MAY be omitted after 2 hours of deprecation announcements.

#### 6.6. Find lost information by "Synchronization"

\* Section 4.2 (RA format) of [ND], introduce new flag:



- 0 1-bit "Complete configuration" flag. When set, it indicates that all configuration information has been put inside this RA. The last reserved bit has been chosen to preserve the compatibility with [Route Preferences] that already propose to use the first reserved bit.

\* Section 6.2.3 (RA content) of [ND], introduce new flag:

- In the C flag: set if it was possible to put all configuration information into this RA.

\* Section 6.2.3 (RA content) of [ND], add at the end:

It is recommended that all configuration information SHOULD be included in one RA (if MTU permits) for multicast and unicast distribution. If successful, then the "Complete" flag SHOULD be set to signal the possibility of synchronization with hosts.

\* Section 6.3.4 (RA processing) of [ND], add at the beginning:

After: "the receipt of a Router Advertisement MUST NOT invalidate all information received in a previous advertisement or from another source".

Add: "Except for the case when RA received with "Complete" flag set, then any information from the same router (same Link Local Address) missing in this RA SHOULD be deprecated. Information protected by timers SHOULD be put into the deprecated state. Other information SHOULD be returned to the original state: in compliance to information from other routers or to default configuration if other routers do not announce respected information."

\* Section 6.3.4 (RA processing) of [ND], add to the list of PIO processing options:

- If the prefix is missing in RA with the "Complete" flag set, then respective addresses should be put immediately into deprecated state up to the original valid lifetime.

[ND] section 9 mentions: "In order to ensure that future extensions properly coexist with current implementations, all nodes MUST

silently ignore any options they do not recognize in received ND packets and continue processing the packet."

There is a possibility for the gradual introduction of the "Complete" flag:

- o If the host is upgraded to the new functionality first, then the router would send this bit zero (according to the basic [ND]) that would not activate new functionality on the host.
- o If the router is upgraded to the new functionality first, then the host would not pay attention to the flag for Reserved bits.

#### 6.7. Default router announcement rules

\* This document joins [HNCP] section 11 (General Requirements for HNCP Nodes) request to [CPE Requirements]:

The generic requirements G-4 and G-5 are relaxed such that any known default router on any interface is sufficient for a router to announce itself as the default router; similarly, only the loss of all such default routers results in self-invalidation.

#### 6.8. Faster detection of the stale default router

\* Section 6.3.7 (sending Router Solicitations) of [ND].

The text: "When an interface becomes enabled, a host may be unwilling to wait for the next unsolicited Router Advertisement to locate default routers or learn prefixes. To obtain Router Advertisements quickly, a host SHOULD transmit up to MAX\_RTR\_SOLICITATIONS Router Solicitation messages, each separated by at least RTR\_SOLICITATION\_INTERVAL seconds. Router Solicitations may be sent after any of the following events:"

Should be replaced by the text: "

Interface enablement or new router arrival could be the signal of router replacement, a host may be unwilling to wait for the next unsolicited Router Advertisement to locate and invalidate default routers or learn prefixes. To obtain Router Advertisements quickly, a host SHOULD transmit up to MAX\_RTR\_SOLICITATIONS Router Solicitation messages, each separated by at least RTR\_SOLICITATION\_INTERVAL seconds. Router Solicitations may be sent after any of the following events:

- the new router is discovered from RA
- . . . <list of other reasons>
- "

\* Section 6.3.7 (sending Router Solicitations) of [ND].

After the text: "If a host sends MAX\_RTR\_SOLICITATIONS solicitations, and receives no Router Advertisements after having waited MAX\_RTR\_SOLICITATION\_DELAY seconds after sending the last solicitation, the host concludes that there are no routers on the link for the purpose of [ADDRCONF]."

Add new text: "If a host sends MAX\_RTR\_SOLICITATIONS solicitations, and receives no Router Advertisements from the router already present on the default router list after having waited MAX\_RTR\_SOLICITATION\_DELAY seconds, the host concludes that the router SHOULD be deprecated from the default router list."

#### 6.9. Clean orphaned prefixes after default router list change

\* Section 6.3.6 (Timing out Prefixes and Default Routers) of [ND] has:

"Whenever the Lifetime of an entry in the Default Router List expires, that entry is discarded. When removing a router from the Default Router list, the node MUST update the Destination Cache in such a way that all entries using the router perform next-hop determination again rather than continue sending traffic to the (deleted) router."

Add at the end:

"All prefixes announced by deprecated default router SHOULD be checked on the announcement from other default routers. If any prefix is not anymore announced from any router - it SHOULD be deprecated."

#### 7. Interoperability analysis

The primary motivation for the proposed changes originated from residential broadband requirements. [ND] extensions proposed in this document should not affect other environments (enterprise WAN,

Campus). Moreover, some precautions proposed could block mistakes originated by humans in some corner cases in all environments.

This document mostly intersects with Homenet working group documents [HomeNet Architecture], [HNCP], and [MHMP]. It was shown that it is possible to isolate [ND] in the context of Homenet to solve specific [ND] problems without any potential impact to the Homenet development and directions.

[CPE Requirements] have the assumption of managing simplified topologies by manipulating routing information injection into [ND]. It has been shown in [MHMP] and in this document that it is better to signal reachability information to [ND] (reachability information to ND sounds strange) by the deprecation of delegated prefixes. This document joins [MHMP] request to change the approach.

[Route Preferences] have been avoided as the mechanism for environments with PA address space because source address is selected first. Then next hop choice can be simplified - see section 5.1 for more details.

[Route Preferences] could still be applicable for PI (Provider-Independent) address environments because only next hops need to be chosen properly.

## 8. Applicability analysis

Two standard extensions require changes to hosts. Hence, it would take a long time to be implemented in live networks. But workaround exists for the solution to work before it would happen:

- o Absence of implementation for RA information synchronization by C flag on some hosts is not critical because router could use non-volatile memory for prefix storage.
- o Not being capable of excluding a router from the default router list (for the situation when it does not advertise respective prefix) is not critical, because it is needed only for the very advanced MHMP environment with traffic distribution by the policy between different PA providers.  
It is for the far future anyway.

## 9. Security Considerations

This document does not introduce new vulnerabilities.

## 10. IANA Considerations

This document has no any request to IANA.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [ND] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <<https://www.rfc-editor.org/info/rfc4861>>.
- [SLAAC] Thomson, S., Narten, T., and T. Jinmei, "IPv6 Stateless Address Autoconfiguration", RFC 4862, DOI 10.17487/RFC4862, September 2007, <<https://www.rfc-editor.org/info/rfc4862>>.
- [Route Preferences] R. Draves, D. Thaler, "Default Router Preferences and More-Specific Routes", RFC 4191, DOI 10.17487/RFC4191, November 2005, <<https://www.rfc-editor.org/info/rfc4191>>.
- [Multi-Homing] F. Baker, B. Carpenter, "First-Hop Router Selection by Hosts in a Multi-Prefix Network", RFC 8028, DOI 10.17487/RFC8028, November 2016, <<https://www.rfc-editor.org/info/rfc8028>>.
- [NUD improvement] E. Nordmark, I. Gashinsky, "Neighbor Unreachability Detection Is Too Impatient", RFC 7048, DOI 10.17487/RFC7048, July 2010, <<https://www.rfc-editor.org/info/rfc7048>>.
- [Default Address] D. Thaler, R. Draves, A. Matsumoto, T. Chown, "Default Address Selection for Internet Protocol Version 6 (IPv6)", RFC 6724, DOI 10.17487/RFC6724, September 2012, <<https://www.rfc-editor.org/info/rfc6724>>.



- [Node Requirements] T. Chown, J. Loughney, T. Winters, "IPv6 Node Requirements", RFC 8504, DOI 10.17487/RFC8504, January 2019, <<https://www.rfc-editor.org/info/rfc8504>>.
- [CPE Requirements] Singh, H., Beebee W., Donley, C., and B. Stark, "Basic Requirements for IPv6 Customer Edge Routers", RFC 7084, DOI 10.17487/RFC7084, November 2013, <<https://www.rfc-editor.org/info/rfc7084>>.
- [HomeNet Architecture] T. Chown, J. Arkko, A. Brandt, O. Troan, J. Weil, "IPv6 Home Networking Architecture Principles", RFC 7368, DOI 10.17487/RFC7368, October 2014, <<https://www.rfc-editor.org/info/rfc7368>>.
- [HNCP] M. Stenberg, S. Barth, P. Pfister, "Home Networking Control Protocol", RFC 7788, DOI 10.17487/RFC7788, April 2016, <<https://www.rfc-editor.org/info/rfc7788>>.
- [Policy by DHCP] A. Matsumoto, T. Fujisaki, T. Chown, "Distributing Address Selection Policy Using DHCPv6", RFC 7078 DOI 10.17487/RFC7078, January 2014, <<https://www.rfc-editor.org/info/rfc7078>>.
- [Residential practices] Palet, J., "IPv6 Deployment Survey Residential/Household Services) How IPv6 is being deployed?", UK NOF 39, January 2018, <<https://indico.uknof.org.uk/event/41/contributions/542/attachments/712/866/bcop-ipv6-prefix-v9.pdf>>.
- [SLAAC Robustness] F. Gont, J. Zorz, R. Patterson, "Improving the Robustness of Stateless Address Autoconfiguration (SLAAC) to Flash Renumbering Events", draft-ietf-6man-slaac-renum-02 (work in progress), January 2021

## 11.2. Informative References

- [RFC8200] S. Deering, R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [Flash-Renumbering] F. Gont, J. Zorz, R. Patterson, "Reaction of Stateless Address Autoconfiguration (SLAAC) to Flash-Renumbering Events", RFC 8978, March 2021.

[RA-Guard] E. Levy-Abegnoli, G. Van de Velde, C. Popoviciu, J. Mohacsi, "IPv6 Router Advertisement Guard", RFC 6105, DOI 10.17487/RFC6105, February 2011, <<https://www.rfc-editor.org/info/rfc6105>>.

[RA-Guard+] F. Gont, "Implementation Advice for IPv6 Router Advertisement Guard (RA-Guard)", RFC 7113, DOI 10.17487/RFC7113, February 2014, <<https://www.rfc-editor.org/info/rfc7113>>.

[MHMP] O. Troan, D. Miles, S. Matsushima, T. Okimoto, D. Wing, "IPv6 Multihoming without Network Address Translation", RFC 7157, DOI 10.17487/RFC7157, March 2014, <<https://www.rfc-editor.org/info/rfc7157>>.

[ULA] R. Hinden, B. Haberman, "Unique Local IPv6 Unicast Addresses", RFC 4193, DOI 10.17487/RFC4193, October 2005, <<https://www.rfc-editor.org/info/rfc4193>>.

## 12. Acknowledgments

Thanks to 6man working group for problem discussion.

## Authors' Addresses

Olorunloba Olopade  
Virgin Media  
270 & 280 Bartley Way, Bartley Wood Business Park, Hook,  
Hampshire RG27 9UP  
Email: [Loba.Olopade@virginmedia.co.uk](mailto:Loba.Olopade@virginmedia.co.uk)

Eduard Vasilenko  
Huawei Technologies  
17/4 Krylatskaya st, Moscow, Russia 121614  
Email: [vasilenko.eduard@huawei.com](mailto:vasilenko.eduard@huawei.com)

Paolo Volpato  
Huawei Technologies  
Via Lorenteggio 240, 20147 Milan, Italy  
Email: [paolo.volpato@huawei.com](mailto:paolo.volpato@huawei.com)



6MAN Working Group  
Internet-Draft  
Updates: 4884 (if approved)  
Intended status: Standards Track  
Expires: 27 October 2022

X. Min  
ZTE Corp.  
G. Mirsky  
Ericsson  
25 April 2022

ICMPv6 Echo Request/Reply for Enabled In-situ OAM Capabilities  
draft-xiao-6man-icmpv6-ioam-conf-state-01

## Abstract

This document describes the ICMPv6 IOAM Echo functionality, which uses the ICMPv6 IOAM Echo Request/Reply messages, allowing the IOAM encapsulating node to discover the enabled IOAM capabilities of each IOAM transit and decapsulating node.

This document updates RFC 4884.

## 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 <https://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 27 October 2022.

## Copyright Notice

Copyright (c) 2022 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 (<https://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 Revised BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Revised BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Conventions Used in This Document . . . . .	3
3. ICMPv6 IOAM Echo Request . . . . .	3
4. ICMPv6 IOAM Echo Reply . . . . .	5
4.1. IOAM Capabilities Objects . . . . .	6
4.2. Examples of IOAM Echo Reply . . . . .	7
5. ICMPv6 Message Processing . . . . .	10
5.1. Code Field Processing . . . . .	11
6. Updates to RFC 4884 . . . . .	12
7. IANA Considerations . . . . .	12
8. Security Considerations . . . . .	14
9. Acknowledgements . . . . .	15
10. References . . . . .	15
10.1. Normative References . . . . .	15
10.2. Informative References . . . . .	16
Authors' Addresses . . . . .	16

## 1. Introduction

IPv6 encapsulation for In-situ OAM (IOAM) data is defined in [I-D.ietf-ippm-ioam-ipv6-options], which uses IPv6 hop-by-hop options and destination option to carry IOAM data.

As specified in [I-D.ietf-ippm-ioam-conf-state], echo request/reply can be used for the IOAM encapsulating node to discover the enabled IOAM capabilities at IOAM transit and decapsulating nodes.

As specified in [RFC4443], the Internet Control Message Protocol for IPv6 (ICMPv6) is an integral part of IPv6, and the base protocol MUST be fully implemented by every IPv6 node. ICMPv6 messages include error messages and informational messages, and the latter are referred to as ICMPv6 Echo Request/Reply messages. [RFC4884] defines ICMPv6 Extension Structure by which multi-part ICMPv6 error messages are supported. [RFC8335] defines ICMPv6 Extended Echo Request/Reply messages, and the ICMPv6 Extended Echo Request contains an ICMPv6 Extension Structure customized for this message. Both [RFC4884] and [RFC8335] provide sound principles and examples on how to extend ICMPv6 error messages and echo request/reply messages.

This document describes the ICMPv6 IOAM Echo functionality, which uses the ICMPv6 IOAM Echo Request/Reply messages, allowing the IOAM encapsulating node to discover the enabled IOAM capabilities of each IOAM transit and decapsulating node.

The IOAM encapsulating node sends an ICMPv6 IOAM Echo Request message to each IOAM transit and decapsulating node, then each receiving node executes access control procedures, and if access is granted, each receiving node returns an ICMPv6 IOAM Echo Reply message which indicates the enabled IOAM capabilities of the receiving node. The ICMPv6 IOAM Echo Reply message contains an ICMPv6 Extension Structure exactly customized to this message, and the ICMPv6 Extension Structure contains one or more IOAM Capabilities Objects.

Note that before the IOAM encapsulating node sends the ICMPv6 IOAM Echo Request messages, it needs to know the IPv6 address of each node along the transport path of a data packet to which IOAM data would be added. That can be achieved by executing ICMPv6 traceroute or provisioning explicit path at the IOAM encapsulating node.

## 2. Conventions Used in This Document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

## 3. ICMPv6 IOAM Echo Request

The ICMPv6 IOAM Echo Request message is encapsulated in an IPv6 header [RFC8200], like any ICMPv6 message.

The ICMPv6 IOAM Echo Request message has the following format:

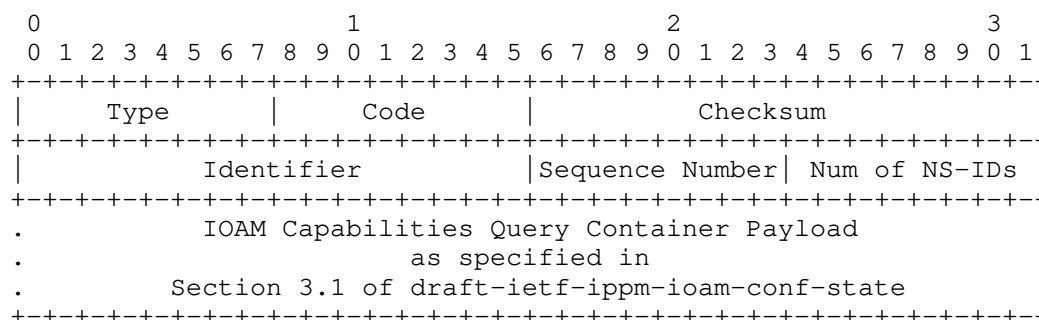


Figure 1: ICMPv6 IOAM Echo Request Message

IPv6 Header fields:

- \* Source Address: The Source Address identifies the IOAM encapsulating node. It MUST be a valid IPv6 unicast address.
- \* Destination Address: The Destination Address identifies the IOAM transit or decapsulating node. It MUST be a valid IPv6 unicast address.

ICMPv6 fields:

- \* Type: IOAM Echo Request. The value is TBD1.
- \* Code: MUST be set to 0 and MUST be ignored upon receipt.
- \* Checksum: The same as defined in [RFC4443].
- \* Identifier: An Identifier aids in matching IOAM Echo Replies to IOAM Echo Requests. It may be zeroed.
- \* Sequence Number: A Sequence Number to aid in matching IOAM Echo Replies to IOAM Echo Requests. It may be zeroed.
- \* Num of NS-IDs: Number of Namespace-IDs within the payload.
- \* Following the IOAM Echo Request header, it's a List of Namespace-IDs, which is also called IOAM Capabilities Query Container Payload in Section 3.1 of [I-D.ietf-ippm-ioam-conf-state]. If the payload would not otherwise terminate on a 4-octet boundary, it MUST be padded with zeroes.

#### 4. ICMPv6 IOAM Echo Reply

The ICMPv6 IOAM Echo Reply message is encapsulated in an IPv6 header [RFC8200], like any ICMPv6 message.

The ICMPv6 IOAM Echo Reply message has the following format:

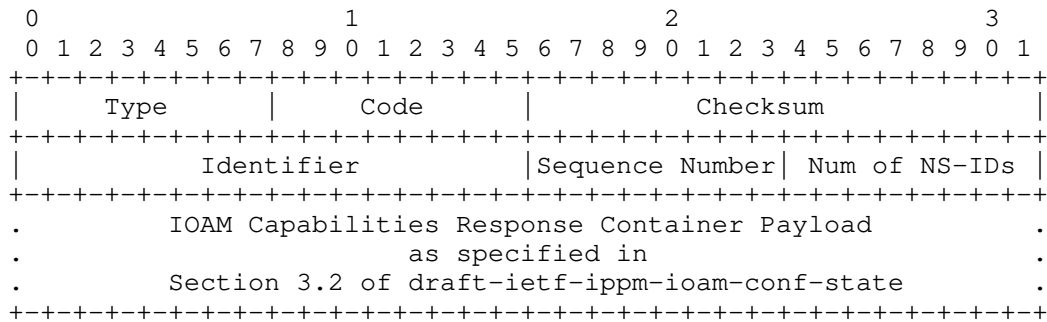


Figure 2: ICMPv6 IOAM Echo Reply Message

IPv6 Header fields:

- \* Source Address: Copied from the Destination Address field of the invoking IOAM Echo Request packet.
- \* Destination Address: Copied from the Source Address field of the invoking IOAM Echo Request packet.

ICMPv6 fields:

- \* Type: IOAM Echo Reply. The value is TBD2.
- \* Code: Values are (0) No Error, (1) Malformed Query, (2) No Matched Namespace-ID, and (3) Exceed the minimum IPv6 MTU.
- \* Checksum: The same as defined in [RFC4443].
- \* Identifier: Copied from the Identifier field of the invoking IOAM Echo Request message.
- \* Sequence Number: Copied from the Sequence Number field of the invoking IOAM Echo Request message.
- \* Num of NS-IDs: Number of different Namespace-IDs within the payload, its value MUST be no more than the Num of NS-IDs field of the invoking IOAM Echo Request message.



- \* Following the IOAM Echo Reply header, it's a List of IOAM Capabilities Objects, which is also called IOAM Capabilities Response Container Payload in Section 3.2 of [I-D.ietf-ippm-ioam-conf-state].
- \* Section 7 of [RFC4884] defines the ICMP Extension Structure. As per RFC 4884, the Extension Structure contains exactly one Extension Header followed by one or more objects. When applied to the ICMPv6 IOAM Echo Reply message, the ICMP Extension Structure MUST contain one or more IOAM Capabilities Objects.

#### 4.1. IOAM Capabilities Objects

All ICMPv6 IOAM Capabilities Objects are encapsulated in an ICMPv6 IOAM Echo Reply message.

Each ICMPv6 IOAM Capabilities Object has the following format:

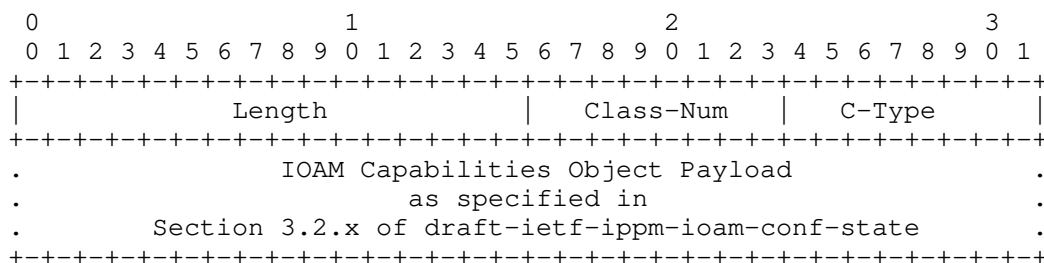


Figure 3: IOAM Capabilities Object

Object fields:

- \* Class-Num: IOAM Capabilities Objects. The values are listed as the following:

Value	Object Name
-----	-----
TBD3	IOAM Tracing Capabilities Object
TBD4	IOAM Proof-of-Transit Capabilities Object
TBD5	IOAM Edge-to-Edge Capabilities Object
TBD6	IOAM DEX Capabilities Object
TBD7	IOAM End-of-Domain Object

- \* C-Type: Values are listed as the following:

Class-Num	C-Type	C-Type Name
-----	-----	-----
TBD3	0	Reserved
	1	Pre-allocated Tracing
	2	Incremental Tracing
TBD4	0	Reserved
TBD5	0	Reserved
TBD6	0	Reserved
TBD7	0	Reserved

- \* Length: Length of the object, measured in octets, including the Object Header and Object Payload.
- \* Following the IOAM Capabilities Object Header, it's the IOAM Capabilities Object Payload, which is defined respectively in Section 3.2.1, Section 3.2.2, Section 3.2.3, Section 3.2.4, Section 3.2.5 and Section 3.2.6 of [I-D.ietf-ippm-ioam-conf-state].

#### 4.2. Examples of IOAM Echo Reply

The format of ICMPv6 IOAM Echo Reply can vary from deployment to deployment.

In a deployment where only the default Namespace-ID is used, the IOAM Pre-allocated Tracing Capabilities and IOAM Proof-of-Transit Capabilities are enabled at the IOAM transit node that received ICMPv6 IOAM Echo Request message, the ICMPv6 IOAM Echo Reply message is depicted as the following:

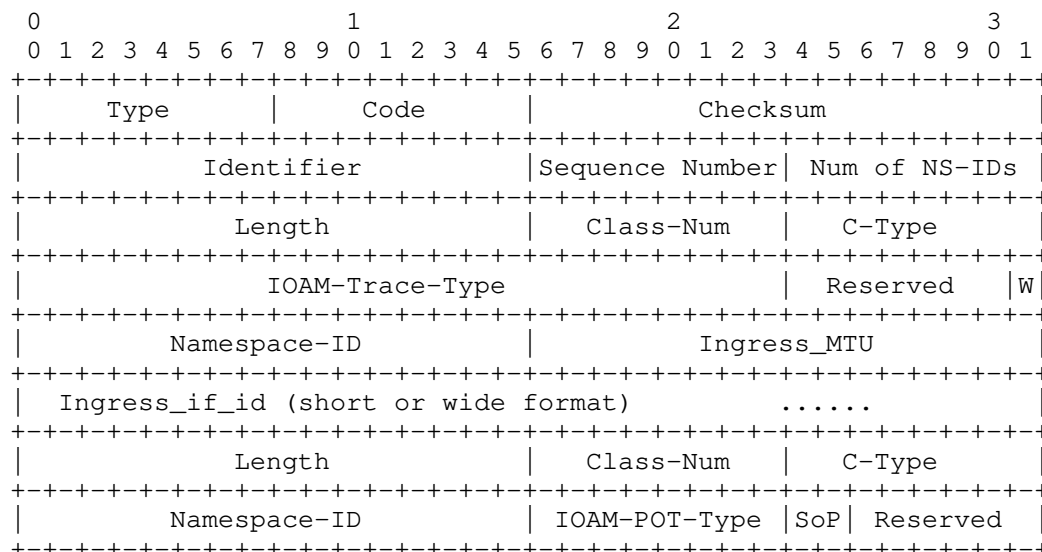


Figure 4: Example 1 of IOAM Echo Reply

In a deployment where two Namespace-IDs (Namespace-ID1 and Namespace-ID2) are used, for both Namespace-ID1 and Namespace-ID2 the IOAM Pre-allocated Tracing Capabilities and IOAM Proof-of-Transit Capabilities are enabled at the IOAM transit node that received ICMPv6 IOAM Echo Request message, the ICMPv6 IOAM Echo Reply message is depicted as the following:

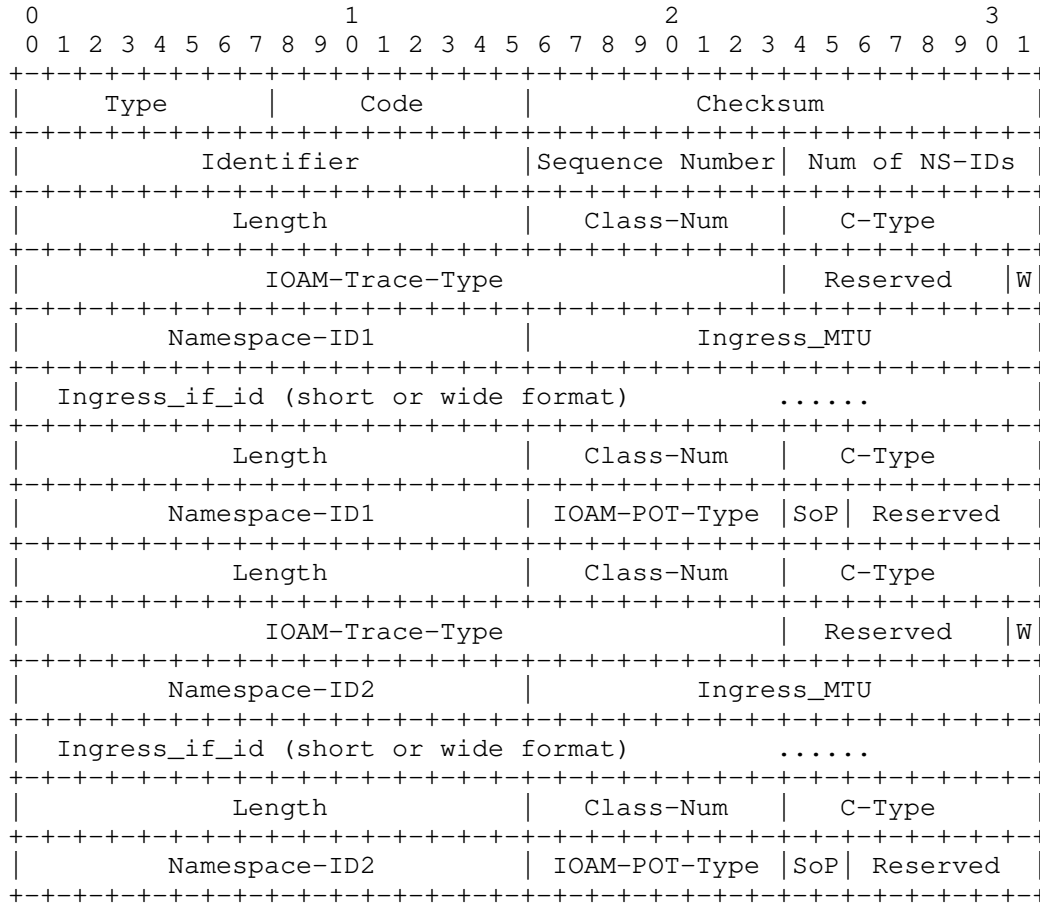


Figure 5: Example 2 of IOAM Echo Reply

In a deployment where only the default Namespace-ID is used, the IOAM Pre-allocated Tracing Capabilities, IOAM Proof-of-Transit Capabilities and IOAM Edge-to-Edge Capabilities are enabled at the IOAM decapsulating node that received ICMPv6 IOAM Echo Request message, the ICMPv6 IOAM Echo Reply message is depicted as the following:

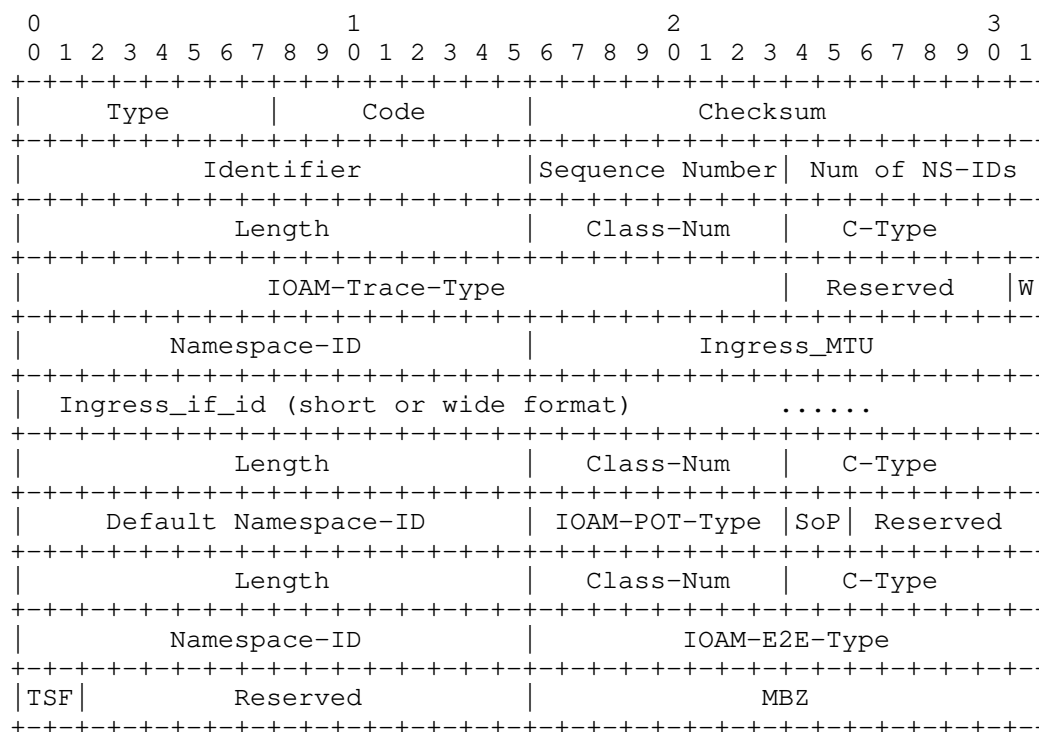


Figure 6: Example 3 of IOAM Echo Reply

Note that when an ICMPv6 IOAM Echo Request message or IOAM Echo Reply message is received, the Payload Length field of IPv6 Header [RFC8200] indicates the message length.

## 5. ICMPv6 Message Processing

When a node receives an ICMPv6 IOAM Echo Request and any of the following conditions apply, the node MUST silently discard the incoming message:

- \* The node does not recognize the ICMPv6 IOAM Echo Request message.
- \* The node has not explicitly enabled ICMPv6 IOAM Echo functionality.
- \* The incoming ICMPv6 IOAM Echo Request carries a Source Address that is not explicitly authorized.
- \* The Source Address of the incoming message is not a unicast address.

- \* The Destination Address of the incoming message is a multicast address.

Otherwise, when a node receives an ICMPv6 IOAM Echo Request, it MUST format an ICMPv6 IOAM Echo Reply as follows:

- \* Set the Hop Limit to 255.
- \* Set the DiffServ codepoint to CS0 [RFC4594].
- \* Copy the Destination Address from the IOAM Echo Request to the Source Address of the IOAM Echo Reply.
- \* Copy the Source Address from the IOAM Echo Request to the Destination Address of the IOAM Echo Reply.
- \* Set the Next Header to (58) ICMPv6.
- \* Set the ICMPv6 Type to (TBD2) IOAM Echo Reply.
- \* Copy the Identifier from the IOAM Echo Request to the IOAM Echo Reply.
- \* Copy the Sequence Number from the IOAM Echo Request to the IOAM Echo Reply.
- \* Set the Code field as described in Section 5.1.
- \* If the Code field is equal to (0) No Error, then add one or more objects as described in Section 4.1.
- \* Set the Checksum appropriately.
- \* Forward the ICMPv6 IOAM Echo Reply to its destination.

#### 5.1. Code Field Processing

The Code field MUST be set to (1) Malformed Query if any of the following conditions apply:

- \* The ICMPv6 IOAM Echo Request does not include any Namespace-ID.
- \* The value of Num of NS-IDs field does not match the contained list of Namespace-IDs.
- \* The query is otherwise malformed.

The Code field MUST be set to (2) No Matched Namespace-ID if none of the contained list of Namespace-IDs is recognized.

The Code field MUST be set to (3) Exceed the minimum IPv6 MTU if the formatted ICMPv6 IOAM Echo Reply exceeds the minimum IPv6 MTU (i.e., 1280 octets). In this case, all objects MUST be stripped before forwarding the ICMPv6 Echo Reply to its destination.

Otherwise, the Code field MUST be set to (0) No Error.

## 6. Updates to RFC 4884

Section 4.6 of [RFC4884] provides a list of extensible ICMP messages (i.e., messages that can carry the ICMP Extension Structure). This document adds the ICMPv6 IOAM Echo Request message and the ICMPv6 IOAM Echo Reply message to that list.

## 7. IANA Considerations

This document requests the following IANA actions:

- \* Add the following to the "ICMPv6 'type' Numbers" registry:
  - TBD1 IOAM Echo Request
  - As ICMPv6 distinguishes between informational and error messages, and this is an informational message, the value must be assigned from the range 128-255.
- \* Add the following to the "Type TBD1 - IOAM Echo Request" sub-registry:
  - (0) No Error
- \* Add the following to the "ICMPv6 'type' Numbers" registry:
  - TBD2 IOAM Echo Reply
  - As ICMPv6 distinguishes between informational and error messages, and this is an informational message, the value must be assigned from the range 128-255.
- \* Add the following to the "Type TBD2 - IOAM Echo Reply" sub-registry:
  - (0) No Error
  - (1) Malformed Query

- (2) No Matched Namespace-ID
- (3) Exceed the minimum IPv6 MTU
- \* Add the following to the "ICMP Extension Object Classes and Class Sub-types" registry:
  - (TBD3) IOAM Tracing Capabilities Object
- \* Add the following C-types to the "Sub-types - Class TBD3 - IOAM Tracing Capabilities Object" sub-registry:
  - (0) Reserved
  - (1) Pre-allocated Tracing
  - (2) Incremental Tracing
  - C-Type values are assigned on a First Come First Serve (FCFS) basis with a range of 0-255.
- \* Add the following to the "ICMP Extension Object Classes and Class Sub-types" registry:
  - (TBD4) IOAM Proof-of-Transit Capabilities Object
- \* Add the following C-types to the "Sub-types - Class TBD4 - IOAM Proof-of-Transit Capabilities Object" sub-registry:
  - (0) Reserved
  - C-Type values are assigned on an FCFS basis with a range of 0-255.
- \* Add the following to the "ICMP Extension Object Classes and Class Sub-types" registry:
  - (TBD5) IOAM Edge-to-Edge Capabilities Object
- \* Add the following C-types to the "Sub-types - Class TBD5 - IOAM Edge-to-Edge Capabilities Object" sub-registry:
  - (0) Reserved
  - C-Type values are assigned on an FCFS basis with a range of 0-255.



- \* Add the following to the "ICMP Extension Object Classes and Class Sub-types" registry:
  - (TBD6) IOAM DEX Capabilities Object
- \* Add the following C-types to the "Sub-types - Class TBD6 - IOAM DEX Capabilities Object" sub-registry:
  - (0) Reserved
  - C-Type values are assigned on an FCFS basis with a range of 0-255.
- \* Add the following to the "ICMP Extension Object Classes and Class Sub-types" registry:
  - (TBD7) IOAM End-of-Domain Object
- \* Add the following C-types to the "Sub-types - Class TBD7 - IOAM End-of-Domain Object" sub-registry:
  - (0) Reserved
  - C-Type values are assigned on an FCFS basis with a range of 0-255.

All codes mentioned above are assigned on an FCFS basis with a range of 0-255.

## 8. Security Considerations

Security issues discussed in [I-D.ietf-ippm-ioam-conf-state] apply to this document.

This document recommends using IP Authentication Header [RFC4302] or IP Encapsulating Security Payload Header [RFC4303] to provide integrity protection for IOAM Capabilities information.

This document recommends using IP Encapsulating Security Payload Header [RFC4303] to provide privacy protection for IOAM Capabilities information.

This document recommends that the network operators establish policies that restrict access to ICMPv6 IOAM Echo functionality. In order to enforce these policies, nodes that support ICMPv6 IOAM Echo functionality MUST support the following configuration options:

- \* Enable/disable ICMPv6 IOAM Echo functionality. By default, ICMPv6 IOAM Echo functionality is disabled.
- \* Define enabled Namespace-IDs. By default, all Namespace-IDs except the default one (i.e., Namespace-ID 0x0000) are disabled.
- \* For each enabled Namespace-ID, define the prefixes from which ICMPv6 IOAM Echo Request messages are permitted.

When a node receives an ICMPv6 IOAM Echo Request message that it is not configured to support, it MUST silently discard the message. See Section 5 for details.

In order to protect local resources, implementations SHOULD rate-limit incoming ICMPv6 IOAM Echo Request messages.

## 9. Acknowledgements

TBA.

## 10. References

### 10.1. Normative References

- [I-D.ietf-ippm-ioam-conf-state]  
Min, X., Mirsky, G., and L. Bo, "Echo Request/Reply for Enabled In-situ OAM Capabilities", Work in Progress, Internet-Draft, draft-ietf-ippm-ioam-conf-state-03, 26 January 2022, <<https://www.ietf.org/archive/id/draft-ietf-ippm-ioam-conf-state-03.txt>>.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <<https://www.rfc-editor.org/info/rfc4443>>.
- [RFC4884] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "Extended ICMP to Support Multi-Part Messages", RFC 4884, DOI 10.17487/RFC4884, April 2007, <<https://www.rfc-editor.org/info/rfc4884>>.

- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.

## 10.2. Informative References

- [I-D.ietf-ippm-ioam-ipv6-options]  
Bhandari, S. and F. Brockners, "In-situ OAM IPv6 Options", Work in Progress, Internet-Draft, draft-ietf-ippm-ioam-ipv6-options-07, 6 February 2022, <<https://www.ietf.org/archive/id/draft-ietf-ippm-ioam-ipv6-options-07.txt>>.
- [RFC4302] Kent, S., "IP Authentication Header", RFC 4302, DOI 10.17487/RFC4302, December 2005, <<https://www.rfc-editor.org/info/rfc4302>>.
- [RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, DOI 10.17487/RFC4303, December 2005, <<https://www.rfc-editor.org/info/rfc4303>>.
- [RFC4594] Babiarz, J., Chan, K., and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, DOI 10.17487/RFC4594, August 2006, <<https://www.rfc-editor.org/info/rfc4594>>.
- [RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <<https://www.rfc-editor.org/info/rfc8200>>.
- [RFC8335] Bonica, R., Thomas, R., Linkova, J., Lenart, C., and M. Boucadair, "PROBE: A Utility for Probing Interfaces", RFC 8335, DOI 10.17487/RFC8335, February 2018, <<https://www.rfc-editor.org/info/rfc8335>>.

## Authors' Addresses

Xiao Min  
ZTE Corp.  
Nanjing  
China  
Phone: +86 25 88013062  
Email: [xiao.min2@zte.com.cn](mailto:xiao.min2@zte.com.cn)

Greg Mirsky  
Ericsson  
United States of America  
Email: gregimirsky@gmail.com