intarea Working Group
Internet-Draft
Intended status: Standards Track
Expires: 3 September 2026

L. Zhang
Huawei
Z. Zhang
Sea Group
R. Sun
Y. Wang
Huawei Cloud

2 March 2026

Extending ICMP for Multi-path
draft-many-intarea-icmp-mp-01

Abstract

This document extends the ICMP message with a Multi-path Interface Information object to carry the egress interface, next hop, and the corresponding ARP or ND information of each multi-path interface of nodes along the route.

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 3 September 2026.

Copyright Notice

Table of Contents

1. Introduction

As described in [RFC2151], Traceroute is a common TCP/IP tool, which allows users to learn about the route that packets take from their local host to a remote host. It is often used by network and system managers to learn something about the ever-changing structure of the Internet.

Traceroute uses the ICMP Time Exceeded Message to collect the nodes' information along the route. The basic Traceroute can only collect the IP addresses, and host name of nodes along the route that packet forwarded.

[RFC4884] redefines some ICMP messages to support multi-part operation. It defines an extension structure which is situated at the end of the ICMP message to carry the additional information. The extension structure includes an extension header followed by one or more extension objects.

Based on that, [RFC5837] extends the ICMP messages to carry the interface information(including ifIndex, IPv4 address, IPv6 address, name and MTU) by defining a Interface Information Object.

Furthermore, [RFC8335] defines a new network diagnostic tool called PROBE. It can be used to query the status of a probed interface by sending ICMP Extended Echo Request message and receiving ICMP Extended Echo Reply message. The ICMP Extended Echo Reply message includes a "State" field to reflect the state of the ARP table or Neighbor Cache entry associated with the probed interface, which indicates whether the interface is reachable. However, the extended Echo Request message and Echo Reply message can only be used to probe the state of destination interface, cannot be used to probe the interface state of the nodes along the route.

However, when using Traceroute in a multi-path topology, the Traceroute can only get information of one of the available paths. The head end and managers even don’t know that there are multiple paths to the destination, which severely impacts the failure location of the network.

This document extends the ICMP message with a Multi-path Interface Information object to carry the egress interface, next hop, and the corresponding ARP or ND information of each multi-path interface of nodes along the route.

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.2. Terminology

The abbreviations used in this document are:

ECMP: Equal-Cost Multiple Path

ICMP: Internet Control Message Protocol

1.3. Motivation

Traceroute is a common TCP/IP tool, which allows users to learn about the route that packets take from their local host to a remote host.

However, Traceroute is typically used to collect the information of a single path, when using Traceroute in a multi-path topology (there are multiple paths from the source node to the destination node and ECMP, UCMP or other multi-path routing strategy is used.), the Traceroute can only get the information of one of the available paths, and do not know that there are multiple paths. Considering using Traceroute in a DC multi-path topology, the topology is shown in Figure 1:

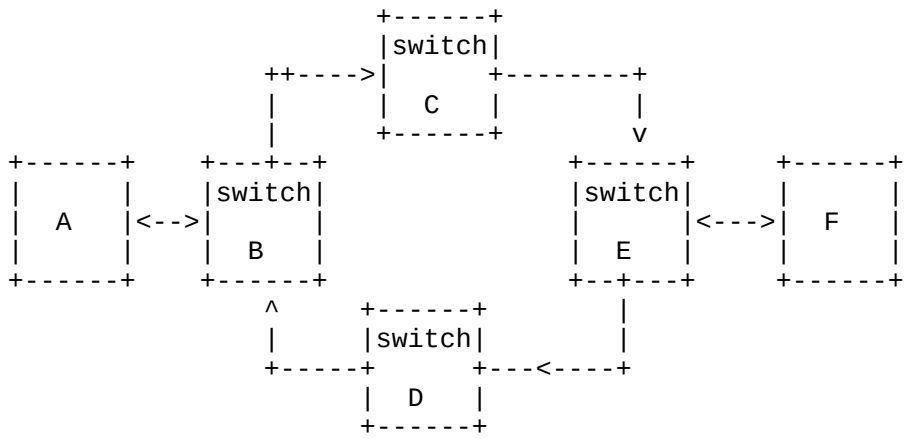

Figure 1: A multi-path topology

In Figure 1, there are four switches and two endpoints. Equal-Cost Multiple Path (ECMP) is applied at switch B. Endpoint A initiates the Traceroute procedures and the target is endpoint F. When the Traceroute request packets arrived at switch B, B have two egress interfaces that can reach endpoint F, but it can only encapsulate one of the interfaces to the reply packet. For example, the interface to switch C is encapsulated. Then the path gets by traceroute is A->B->C->E->F. However, the traffic packets are forwarded on both paths (A->B->C->E->F and A->B->D->E->F). If switch D fails and affects packet forwarding (e.g. packet loss and latency increase), then it is hard for the users to locate the failure, because switch D is not in the forwarding path detected by the traceroute.

Therefore, it is necessary to extend traceroute to enable the multi-path detection.

2. ICMP extension

This section defines the Multi-path Interface Information (MPII) Object, an ICMP extension object with a new Class-Num (Object Class Value). The format of MPII Object is shown in Figure 2.

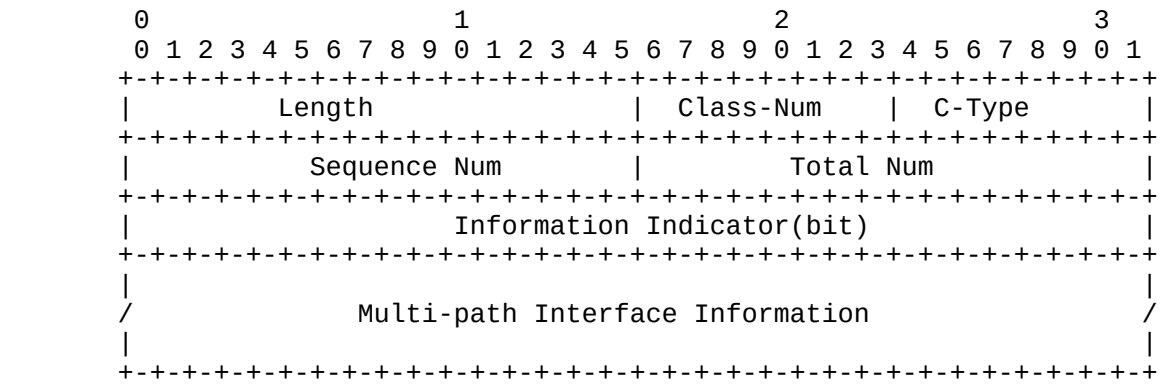

Figure 2: Format of Multi-Path Interface Information (MPII) Object

Class-Num: TBD, to be allocated by IANA.

C-Type: indicates different types of Multi-Path Interface Information, the descriptions of its values are shown as follows:

Value	Description
0	Reserved
1	IPv4 interface
2	IPv6 interface
3-255	Reserved

Table 1: Description of C-Type values

Sequence Num: 16-bit length, the sequence number of this interface in all of the multi-path interfaces.

Total Num: 16-bit length, the total number of multi-path interfaces.

Information Indicator: 32-bit length, indicates the followed multi-path interface information in the Object. The format of it is shown in Figure 3.

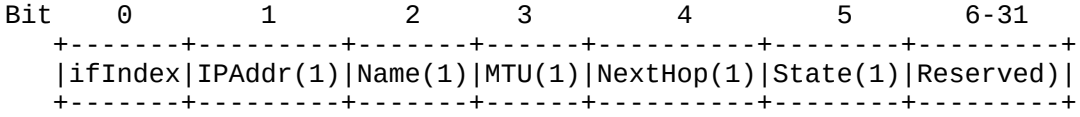

Figure 3: Information Indicator Bit Description

The following are bit-field definitions for Information Indicator:

ifIndex (bit 0) : When set, the 32-bit ifIndex of the interface is included. When clear, the ifIndex is not included.

IP Addr (bit 1) : When set, an IP Address Sub-Object is present. When clear, an IP Address Sub-Object is not present. The IP Address Sub-Object is described in Section 4.2 of [RFC5837].

Name (bit 2): When set, an Interface Name Sub-Object is included. When clear, it is not included. The Name Sub-Object is described in Section 4.3 of [RFC5837].

MTU (bit 3): When set, a 32-bit integer representing the MTU is present. When clear, this 32-bit integer is not present.

NextHop (bit 4): When set, an IP Address Sub-Object for the nexthop is present. When clear, the IP Address Sub-Object for next hop is not present. When both the IP Addr and NextHop bits are set, two IP Address Sub-Objects will be encapsulated in the reply packet. In this case, these two sub-objects MUST be placed in order(the first IP Address Sub-Object is for IP Addr, and the second is for NextHop).

State (bit 5): When set, an Interface State Sub-Object is included. When clear, it is not included. The Interface State Sub-Object is described in Section 2.1.

Multi-path Interface Information: variable, carries the detail multi-path interface information as specified in the Information Indicator.

The MPII Object can be appended to the following messages:

* ICMPv4 Time Exceeded

* ICMPv4 Destination Unreachable

* ICMPv4 Parameter Problem

* ICMPv6 Time Exceeded

* ICMPv6 Destination Unreachable

2.1. Interface State Sub-object

The Interface State Sub-object indicates the state of the ARP table or Neighbor Cache entry associated with the interface. The format of it is shown in Figure 4.

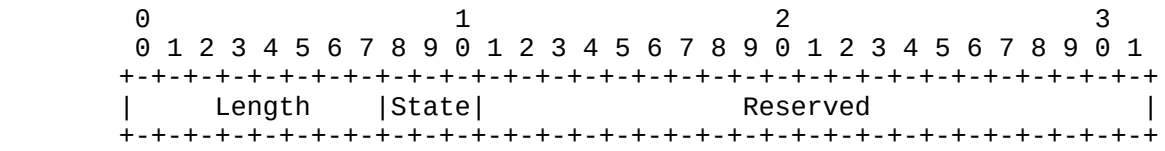

Figure 4: Format of Interface State Sub-object

Where:

Length: 8-bit length, indicates the length of this sub-object in octets, its value equals to 4.

State: the state of the ARP table or Neighbor Cache entry associated with the interface. Values are (0) Reserved, (1) Incomplete, (2) Reachable, (3) Stale, (4) Delay, (5) Probe, and (6) Failed.

Reserved: This field MUST be set to 0 and ignored upon receipt.

3. Usage

Multiple Multi-path Interface Information Object MAY be included within a single ICMP message, provided that each Multi-path Interface Information Object corresponds to a unique interface. A single ICMP message MUST NOT contain two Multi-path Interface Information Object that corresponds to the same interface.

ifIndex, IPAddr, Name, MTU, NextHop, and State information MAY be included whenever they are available. For each kind of these information, at most one instance is included in per Multi-path Interface Information Object.

The address format of IP Address Sub-Object in a Multi-path Interface Information Object depends on the C-Type:

* If the C-Type value is 1, which means it describes the information of an IPv4 interface. In this case, if an IP Address Sub-Object is included, it must specify an IPv4 address.

* If the C-Type value is 2, which means it describes the information of an IPv6 interface. In this case, if an IP Address Sub-Object is included, it must specify an IPv6 address.

An ICMP message that does not conform to these rules and contains multiple Multi-path Interface Information Object of the same interface is considered illegal; An Multi-path Interface Information Object containing more than one instance of each kind of information is considered illegal. If such an illegal ICMP message is received, it MUST be silently discarded.

4. Security Considerations

This extension makes the ICMP messages carrying excessive information, malicious parties may obtain ingress and egress interface, next-hop, the reachable of next-hop (status of ARP and ND), and detailed information about load balancing paths (number of load balancing paths, next-hop and egress interface for each load balancing path, and corresponding ARP and ND reachability) through traceroute. Based on this information, some further information can be inferred. Considering this risk, it is necessary to formulate corresponding security policies as follows:

4.1. Configuration

Network operators should have the capability to control the information carried by Traceroute reply packets.

To achieve flexible control, the following capabilities should be supported on the device:

1. Enable/disable the capability for Traceroute reply packets to carry ingress and egress interface information.

2. Enable/disable the capability for Traceroute reply packets to carry the next hop and the reachability status of the next hop (ARP/ND status). It is recommended that this capability is not enabled by default.

3. Enable/disable the capability for Traceroute reply packets to carry the number and status of load balancing entries (ARP/ND status). It is recommended that this capability is not enabled by default.

4. Control based on the source IP of the request Traceroute packet, only certain IP addresses are allowed to initiate the corresponding Traceroute function.

5. All the above capabilities can be configured separately at the global and interface levels.

4.2. Encryption:

As described in Section 3.4 of [RFC2151], Traceroute uses UDP packets to probe the forwarding path of packets through TTL expiration. For UDP packets, the payload part does not have a specific meaning for Traceroute. So, the payload of UDP packets can be extended to carry the encrypted information. In cases where the intermediate nodes do not recognize or the encrypted information does not match, the reply packets of Traceroute will only carry basic information.

5. IANA Considerations

This document requests IANA to allocate a new Object value for Multi-path Interface Information Object from the "ICMP Extension Object Classes" registry:

Value	Description	Reference
TBD	ICMP Extension Object Classes registry	This document

Table 2

6. References

6.1. 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/rfc/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/rfc/rfc8174>.

6.2. Informative References

[RFC2151] Kessler, G. and S. Shepard, "A Primer On Internet and TCP/IP Tools and Utilities", FYI 30, RFC 2151, DOI 10.17487/RFC2151, June 1997, <https://www.rfc-editor.org/rfc/rfc2151>.

[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/rfc/rfc4884>.

[RFC5837] Atlas, A., Ed., Bonica, R., Ed., Pignataro, C., Ed., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, DOI 10.17487/RFC5837, April 2010, <https://www.rfc-editor.org/rfc/rfc5837>.

[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/rfc/rfc8335>.

Acknowledgements

TBD

Contributors

Ranxiao Zhao
Huawei
China
Email: zhaoranxiao@huawei.com

Haibo Wang
Huawei
China
Email: rainsword.wang@huawei.com

Authors' Addresses

Li Zhang
Huawei
China
Email: zhangli344@huawei.com

Zizhou Zhang
Sea Group
Singapore
Email: zhangzz@sea.com

Ronghua Sun
Huawei Cloud
China
Email: sunronghua@huawei.com

Yang Wang
Huawei Cloud
China
Email: sky.wangyang@huawei.com