IPPM Working Group
Internet-Draft
Intended status: Standards Track
Expires: 1 March 2026

X. He
China Telecom
X. Min
ZTE Corp.
F. Brockners
Cisco
G. Fioccola
Huawei
C. Xie
China Telecom

28 August 2025

IOAM Trace Option Extensions for Incorporating the Alternate-Marking Method
draft-he-ippm-ioam-extensions-incorporating-am-04

Abstract

In situ Operation, Administration, and Maintenance (IOAM) is used for recording and collecting operational and telemetry information. Specifically, passport-based IOAM allows telemetry data generated by each node along the path to be pushed into data packets when they traverse the network, while postcard-based IOAM allows IOAM data generated by each node to be directly exported without being pushed into in-flight data packets. This document extends IOAM Trace Option for incorporating the Alternate-Marking Method.

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 1 March 2026.

Table of Contents

1. Introduction

IOAM [RFC9197], which defines two possible IOAM Trace Option-Types: Pre-allocated Trace and Incremental Trace, is used for monitoring traffic in the network and for incorporating IOAM data fields into in-flight data packets. IOAM Trace Option is known as the passport mode, in which each node on the path can add telemetry data to the user packets (i.e., stamps the passport). IOAM Direct Export (DEX) [RFC9326] is used as a trigger for IOAM nodes to directly export IOAM data to a receiving entity such as a collector, analyzer, or controller. IOAM DEX is also referred as the postcard mode, in which each node directly exports the telemetry data using an independent packet (i.e., sends a postcard) while the user packets are unmodified.

The disadvantage of the passport mode is that if a packet is dropped on the path, the IOAM data collected are also lost. So the passport mode such as IOAM Trace Option-Type has no ability to monitor packet drop and packet drop location.

IOAM DEX Option-Type can complement IOAM Trace Option-Type in that even if a packet is dropped on the path, the collected partial data are still available. By correlating the data from different nodes, the number of the discarded packets can be counted accurately and packet drop location can be pinpointed.

The Alternate-Marking [RFC9341] technique has been proven to work well to perform packet loss, delay, and jitter measurements on live traffic. RFC9343 describes how the Alternate-Marking Method can be used to measure performance metrics in IPv6. It defines an Extension Header Option to encode Alternate-Marking information in both the Hop-by-Hop Options Header and Destination Options Header. In order to facilitate the deployment and improve the scalability of the Alternate-Marking Method, the Flow Monitoring Identification (FlowMonID) field is introduced. The benefits of introducing FlowMonID are obvious: First, it helps to reduce the per-node configuration; Second, it simplifies the counters handling; Third, it eases the data export encapsulation and correlation for the collectors.

[draft-he-ippm-ioam-dex-extensions-incorporating-am] presents the problems statement currently faced by IOAM DEX Option in measuring performance metrics such as packet loss. In order to augment performance measurement of IOAM, it also defines the IOAM DEX Option extension to incorporate the Alternate-Marking Method into IOAM. Therefore, the extended IOAM DEX Option can be used for both IOAM trace monitoring and performance measurement such packet loss, latency and jitter, simplifing the complexity of forwarding chips.

For the same purpose, this document defines the IOAM Trace Option extensions for incorporating the Alternate-Marking Method to augment IOAM in performance measurement.

2. Conventions

2.1. 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.

2.2. Terminology

Abbreviations used in this document:

DEX: Direct Exporting

IOAM: In situ Operation, Administration, and Maintenance

MPN: Measurement Period Number

OAM: Operation, Administration, and Maintenance

SN: Sequence Number

3. The Extended IOAM Trace Format

The format of the extended DEX Option-Type is depicted in Figure 1. All fields are same as IOAM Trace Option-Type header Format defined in RFC9197 except the 8-bit Reserved field. The extended Trace Option-Type Format uses the most significant 5 bits of the Reserved field.

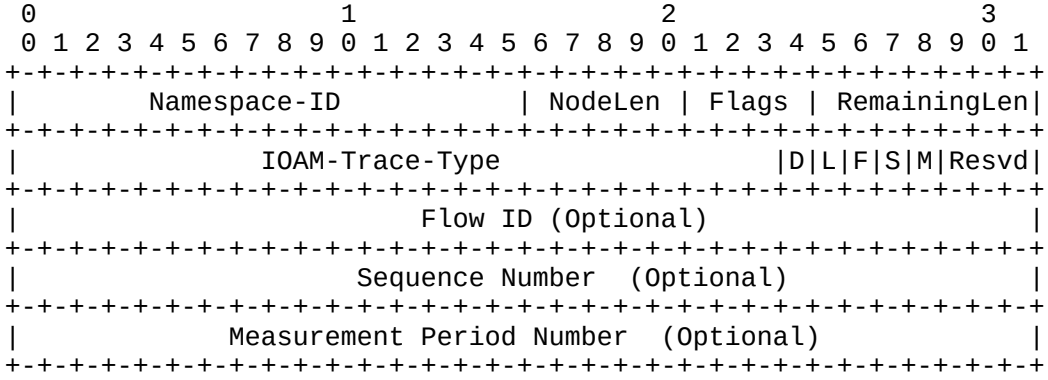

Figure 1: The Extended IOAM Trace Option header Format

Where:

Namespace-ID: 16-bit identifier of the IOAM namespace, as defined in [RFC9197].

NodeLen: 5-bit unsigned integer, as defined in [RFC9197].

Flags: 4-bit field, as defined in [RFC9197].

RemainingLen: 7-bit unsigned integer, as defined in [RFC9197].

IOAM-Trace-Type: 24-bit identifier that specifies which data types are used in the node data list, as defined in [RFC9197].

L: 1-bit Loss flag, defined in this document for Packet Loss Measurement as described in Section 4.1.

D: 1-bit Delay flag, defined in this document for Packet Delay Measurement as described in Section 4.2.

F: 1-bit Flow ID flag, defined in this document. This flag that is set to 1 indicates the existence of a corresponding optional 4-octet field.

S: 1-bit Sequence Number (SN) flag, defined in this document. This flag that is set to 1 indicates the existence of a corresponding optional 4-octet field.

M: 1-bit Measurement Period Number (MPN) flag, defined in this document. This flag that is set to 1 indicates the existence of a corresponding optional 4-octet field.

Reserved: 3-bit field, reserved for future use. These bits MUST be set to zero on transmission and ignored on receipt.

Optional fields: The optional fields, if present, reside after the Reserved field. The order of the optional fields is according to the order of the respective bits, which are enabled in the F, S and M Flags field. Each optional field is 4 octets long.

Flow ID: An optional 32-bit field representing the flow identifier. If the actual Flow ID is shorter than 32 bits, it is zero padded in its most significant bits. The field is set at the encapsulating node and exported to the receiving entity by the forwarding nodes. The Flow ID can be used to correlate the exported data of the same flow from multiple nodes and from multiple packets. Flow ID values are expected to be allocated in a way that avoids collisions. For example, random assignment of Flow ID values can be subject to collisions, while centralized allocation can avoid this problem. The specification of the Flow ID allocation method is not within the scope of this document.

Sequence Number: An optional 32-bit sequence number, starting from 0 and incremented by 1 for each packet from the same flow at the encapsulating node that includes the DEX option. The Sequence Number, when combined with the Flow ID, provides a convenient approach to correlate the exported data from the same user packet.

Measurement Period Number(MPN): An optional 32-bit field representing the measurement period number of the monitored flow, starting from 0 and incremented by 1 for the specified flow with the same Flow ID. The field is set at the encapsulating node and exported to the receiving entity by the forwarding nodes. The MPN, when combined with the Flow ID, provides a convenient approach to correlate the exported data of the same flow during the same measurement period from multiple nodes.

4. The IOAM Operation

The extended Trace Option-Type SHOULD support the three IOAM operation modes: only IOAM trace monitoring; only performance measurement; hybrid.

* Only IOAM trace monitoring: As Trace Option-Type does, an IOAM encapsulating node that supports the Extended Trace Option-Type MUST support the ability to incorporate the Extended Trace Option-Type into all packets or a selective subset of the packets that are forwarded by the IOAM encapsulating node. At the same time, it MUST set corresponding bit flag to 1 in IOAM Trace-Type field of the extended Trace Option-Type so that each node along the path needs to add the specified IOAM data to the user packets. Also, it Must set the L flag and D flag of the extended Trace Option-Type to zero on transmit and ignored by the monitoring nodes.

* Only performance measurement: To ensure the fidelity of performance measurement, an IOAM encapsulating node MUST incorporate the extended Trace Option-Type into all packets of the measured user traffic it forwards. Like the Alternate-Marking Method [RFC9343], for packet loss measurement, it Must switch the value of the L bit between 0 and 1 after a fixed number of packets or according to a fixed timer; for packet delay measurement, Single-Marking or Double-Marking methodology can be adopted by switching the value of the L bit or D bit between 0 and 1. But it Must set all 24 bits flag to 0 in IOAM Trace-Type field of the extended Trace Option-Type so that each node along the path does not need to add the IOAM data to the user packets.

* Hybrid: To perform both IOAM trace monitoring and performance measurement concurrently, an IOAM encapsulating node MUST incorporate the extended Trace Option-Type into all the traffic of interest it forwards. For performance measurement, an IOAM encapsulating node MUST mark each packet it forwards in L flag and D flag of the extended Trace Option-Type; for IOAM trace monitoring, all packets or only a subset of the packets may be selected by an IOAM encapsulating node. For every selected packet, an IOAM encapsulating node MUST set corresponding bit flag

to 1 in IOAM Trace-Type field of the extended Trace Option-Type so that each node along the path needs to add the specified IOAM data to the user packets; for all the other packets not selected, an IOAM encapsulating node MUST set all 24 bits flag to 0 in IOAM Trace-Type field of the extended Trace Option-Type, such that each node along the path does not need to add the IOAM data to the user packets.

4.1. Packet Loss Measurement

The measurement of the packet loss is detailed in [RFC9341]and [RFC9343]. The packets of the flow identified by Flow ID 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 (IOAM encapsulating 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 that occurred in any single batch between any two nodes. Each batch represents a measurable entity recognizable by all network nodes along the path, which export the counter value of this batch along with the Flow ID and the MPN (if it exists) to the receiving entity (e.g., the collector).

4.2. Packet Delay Measurement

Delay metrics MAY be calculated using the following two possibilities:

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.

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 with the D bit set to 1 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 forwarding node can store and export 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. Sequence Number can be used to identify multiple timestamps in different packets that pertain to the same measurement block in case of packets out of order. It also can be used to identify which double-marked packet is lost. 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. 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.

In summary, the approach with Double-Marking is better than the approach with Single-Marking. In the implementation, the timestamps along with Flow ID, MPN and Sequence Number (if it exists) can be sent out to the receiving entity that is responsible for the calculation.

5. Flow Identification

The Flow Identification (Flow ID) identifies the flow to be measured and is required for some general reasons, which is described in Section 5.3 of [RFC9343]. [RFC9343] uses 20-bit FlowMonID to determine a monitored flow within the measurement domain. Compared to the FlowMonID, the Flow ID in this document is a 32-bit field, which amplifies the FlowMonID space by 4096 times. Accordingly, a chance of collision is greatly reduced in a distributed way.

When the 32-bit Flow ID is used for every source node, if there are N edge nodes (source nodes) in a large-scale operator network, and each source node can generate a unique Flow ID for every measured flow independently and randomly in a distributed way. Assuming that each node randomly generates M different Flow IDs from the available K flow identification space, then the total possible sample space (PSS) is

the Nth power of C (K, M) (1)

and the total possible sample space without overlapping (PSSno)is

C1 (K, M)*C2 (K-M, M)*....*CN (k-(N-1)M, M) (2)

Theoritically, the collision probability (CP) is calculated as:

CP=1-PSSno/PSS (3)

Take K=32nd power of 2 as an example, which corresponds to 32-bit Flow ID space. When the number N and M are given different values, we can obtain the corresponding CP values shown in the following table.

	N=100	N=100	N=200	N=200	N=200	N=200
32-bit Flow ID	M=100	M=200	M=200	M=300	M=400	M=500
CP	0.0115	0.0453	0.1692	0.3410	0.5235	0.6860

It is not difficult to observe that as the number of concurrent monitored flows increases,the collision probability is rapidly increasing. As shown in the table, when generating 10000 concurrent flows, the CP is 0.0115; when generating 100000 concurrent flows, the CP rises to 0.6860. If K=20th power of 2 is taken, which corresponds to 20-bit Flow ID space, when generating 10000 concurrent flows, we can calculate the collision probability will drastically rises to

approximately 100%. In practical deployment scenarios of large-scale networks, the concurrent measurement flows could reach orders of magnitude of 100000 or even higher, thus the collision probability will rise sharply.

It is preferred that Flow ID be assigned by the central controller. Since the controller knows the network topology, it can allocate the value properly to guarantee the uniqueness of Flow ID allocation.

In some cases where the central controller is not available and the distributed way must be adopted, every source node (encapsulating node) needs to allocate Flow ID independently. In order to avoid the collision, Flow ID field may be devided into two sub-fields: NodeID and FlowMonID. NodeID is assigned uniquely in measurement domain and FlowMonID is assigned randomly and uniquely in a device. The length allocation of the two sub-fields depends on practical implementation, for example, NodeID uses 20 bits and FlowMonID uses 12 bits, or both use an average of 16 bits.

6. IANA Considerations

6.1. IOAM Type

The "IOAM Option-Type" registry is defined in Section 7.1 of [RFC9197].

IANA is requested to allocate the following code point from the "IOAM Option-Type" registry as follows:

Code Point: TBA

Name: IOAM Extended Trace Option Type

Description: IOAM Trace Option Type

Reference: This document

If possible, IANA is requested to allocate code point 6(TBA-type).

7. Performance Considerations

The extended Trace Option-Type triggers IOAM trace data to be filled in live data packets and performance measurement data to be exported to a receiving entity. The performance impact from IOAM trace data could be limited by only on subsets of the live user traffic, e.g., per interface, based on an access control list or flow specification defining a specific set of traffic, etc.

Performance measurement is implemented based on the Alternate-Marking Method. In Hop-by-Hop mode for loss measurement, every node along the path exports only a packet carrying counter value of each measurement block including a batch of packets; In End-to-End mode for loss measurement, only the IOAM encapsulating node and the IOAM decapsulating node export a packet carrying counter value of each measurement block. Similarly, in Hop-by-Hop mode for delay measurement, every node along the path exports one or multiple packets carrying the timestamps of the marked packets in each measurement block; In End-to-End mode for delay measurement, only the IOAM encapsulating node and the IOAM decapsulating node export one or multiple packets carrying the timestamps of the same marked packets in each measurement block. Because of the very small amount of exported traffic, it would not affect the network bandwidth and would not overload the receiving entity.

8. Security Considerations

The security considerations of IOAM in general are discussed in [RFC9197], and the security considerations of IOAM DEX Option-Type are discussed in [RFC9326]. There are not additional security considerations in this extended IOAM DEX Option-Type.

9. References

9.1. 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>.

[RFC9197] Brockners, F., Ed., Bhandari, S., Ed., and T. Mizrahi, Ed., "Data Fields for In Situ Operations, Administration, and Maintenance (IOAM)", RFC 9197, DOI 10.17487/RFC9197, May 2022, <https://www.rfc-editor.org/info/rfc9197>.

[RFC9326] Song, H., Gafni, B., Brockners, F., Bhandari, S., and T. Mizrahi, "In Situ Operations, Administration, and Maintenance (IOAM) Direct Exporting", RFC 9326, DOI 10.17487/RFC9326, November 2022, <https://www.rfc-editor.org/info/rfc9326>.

[RFC9341] Fioccola, G., Ed., Cociglio, M., Mirsky, G., Mizrahi, T., and T. Zhou, "Alternate-Marking Method", RFC 9341, DOI 10.17487/RFC9341, December 2022, <https://www.rfc-editor.org/info/rfc9341>.

[RFC9343] Fioccola, G., Zhou, T., Cociglio, M., Qin, F., and R. Pang, "IPv6 Application of the Alternate-Marking Method", RFC 9343, DOI 10.17487/RFC9343, December 2022, <https://www.rfc-editor.org/info/rfc9343>.

9.2. Informative References

[I-D.he-ippm-ioam-dex-extensions-incorporating-am] He, X., Brockners, F., Song, H., Fioccola, G., and A. Wang, "IOAM Direct Exporting (DEX) Option Extensions for Incorporating the Alternate-Marking Method", Work in Progress, Internet-Draft, draft-he-ippm-ioam-dex-extensions-incorporating-am-02, 28 August 2025, <https://datatracker.ietf.org/doc/html/draft-he-ippm-ioam-dex-extensions-incorporating-am-02>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC9486] Bhandari, S., Ed. and F. Brockners, Ed., "IPv6 Options for In Situ Operations, Administration, and Maintenance (IOAM)", RFC 9486, DOI 10.17487/RFC9486, September 2023, <https://www.rfc-editor.org/info/rfc9486>.

Authors' Addresses

Xiaoming He
China Telecom
Email: hexm4@chinatelecom.cn

Xiao Min
ZTE Corp.
Email: xiao.min2@zte.com.cn

Frank Brockners
Cisco
Email: fbrockne@cisco.com

Giuseppe Fioccola
Huawei
Email: giuseppe.fioccola@huawei.com

Chongfeng Xie
China Telecom
Email: xiechf@chinatelecom.cn