Workgroup: IPPM Working Group
Internet-Draft:
draft-he-ippm-integrating-am-into-ioam-01
Published: 8 January 2024
Intended Status: Standards Track
Expires: 11 July 2024
Authors: X. He
China Telecom
F. Brockners
Cisco
H. Song
Futurewei
G. Fioccola
Huawei
A. Wang
China Telecom

Integrating the Alternate-Marking Method into In Situ Operations, Administration, and Maintenance (IOAM)

Abstract

In situ Operations, 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 Direct Export (DEX) Option-Type to integrate the Alternate-Marking Method into IOAM.

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 11 July 2024.

Table of Contents

1. Introduction

IOAM [RFC9197], which defines four possible IOAM-Option-Types: Pre-allocated Trace, Incremental Trace, Proof of Transit (POT), and Edge-to-Edge, is used for monitoring traffic in the network and for incorporating IOAM data fields into in-flight data packets. IOAM [RFC9197] 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 partial data collected 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.

This document presents the problems and challenges currently faced by IOAM in measuring performance metrics such as packet loss, delay, and jitter. In order to augment performance measurement of IOAM, IOAM DEX Option-Type is extended to incorporate the Alternate-Marking Method into IOAM.

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.

3. Problems and Challenges

Although IOAM DEX Option-Type can complement IOAM Trace Option-Type for monitoring packet loss, some issues have to be considered as follows.

Issue 1: If an IOAM encapsulating node incorporates the DEX Option-Type into all the traffic of interest it forwards, it may lead to an excessive amount of exported data, which may overload the network and the receiving entity. Therefore, an IOAM encapsulating node that supports the DEX Option-Type MUST support the ability to incorporate the DEX Option-Type selectively into a subset of the packets that are forwarded by the IOAM encapsulating node.

Issue 2: In theory, if an IOAM encapsulating node incorporates the DEX Option-Type into all the traffic it forwards, the fidelity of packet loss measurement can be ensured. If the too small subset of traffic or too low traffic sampling on an encapsulating node is implemented, loss measurement results can not reflect the actual packet drop, due to the fact that the transmitting packet interval does not cover packet drop caused by instantaneous congestion such as microbursts.

Issue 3: Because the IOAM data of the same user packet is generated by every node along the path, the receiving entity needs more processing overhead to correlate these data for packet loss computation. The more user packets measured, the more processing overhead is required.

Issue 4: While using the Alternate-Marking Method, traffic flows are split into consecutive blocks: each block represents a measurable entity unambiguously recognizable by all network devices along the path. In contrast, based on IOAM DEX Option-Type, every IOAM node directly exports an IOAM data to a receiving entity when every user packet is forwarded, and the collected IOAM data are not split into independent measurement blocks. It is the responsibility of the receiving entity to determine the measurement period of performance metrics such as packet loss, delay, and jitter. It is not beneficial to uniform measurement methodology.

4. Integrate the Alternate-Marking Method into IOAM

To address the issues and challenges mentioned in Section 3, IOAM needs to be augmented to implement performance measurement. The Alternate-Marking Method has been widely employed in operators networks. By integrating the Alternate-Marking Method into IOAM, the benefits obtained include:

*While implementing performance measurement, an IOAM encapsulating node may incorporate the DEX Option-Type into all the traffic of interest it forwards; Meanwhile, an IOAM encapsulating node only needs to select a very small subset of the packets that are forwarded for IOAM trace monitoring (e.g., 1/10000 of all the traffic of interest), so the amount of exported data is significantly reduced to mitigate the network and the receiving entity. The IOAM operation is detailed in section 6.

*Using the Alternate-Marking Method, an IOAM encapsulating node could color all the traffic of interest it forwards, not a subset of the packets, thus the fidelity of performance measurement such as packet loss can be ensured.

*While using the Alternate-Marking Method, and in Hop-by-Hop mode for loss measurement, every node along the path only exports 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. It mitigates the network and the receiving entity greatly. Furthermore, compared to IOAM DEX Option-Type, the receiving entity needs much less processing overhead to correlate these counters for packet loss computation.

*While using the Alternate-Marking Method, traffic flows are split into consecutive blocks: each block represents a measurable entity unambiguously recognizable by all network devices along the path, thus the measurement period is completely determined by network devices. The receiving entity does not need to concern about determination of measurement period, but only compute the results of each measurement period. It is beneficial to uniform measurement methodology.

5. The Extended DEX Option-Type Format

The format of the extended DEX Option-Type is depicted in Figure 1. All fields are same as DEX Option-Type Format defined in RFC9326 except the Reserved field. The extended DEX Option-Type Format uses the most significant 2 bits of the Reserved field.

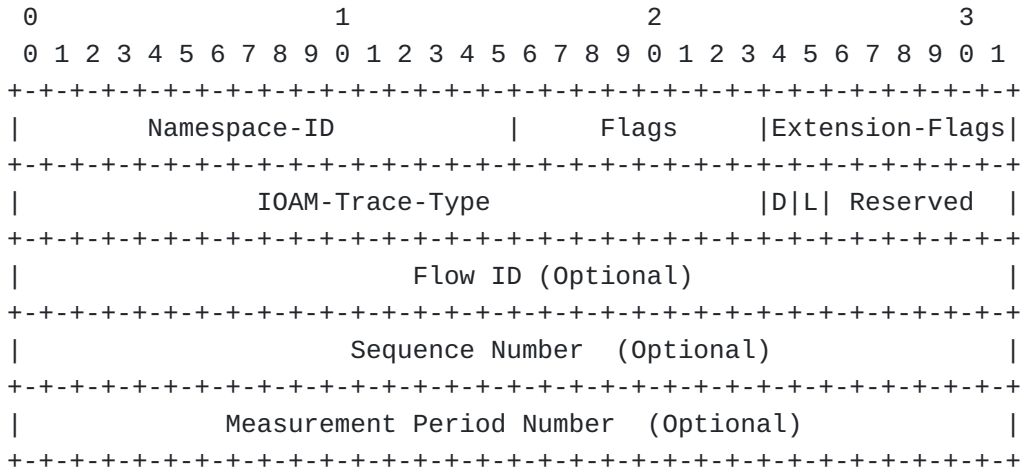

Figure 1: The Extended DEX Option-Type Format

Where:

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

Flags: 8-bit field, comprised of 8 1-bit subfields. Flags are allocated by IANA.

Extension-Flags: 8-bit field, comprised of 8 1-bit subfields. Extension-Flags are allocated by IANA. Every bit in the Extension-Flag field that is set to 1 indicates the existence of a corresponding optional 4-octet field. Bit 0 (the most significant bit) and bit 1 in the registry are allocated by [RFC9326], which are specified as Flow ID and Sequence Number of the monitored traffic, respectively. Bit 2 is specified as Measurement Period Number in this draft.An IOAM node that receives an extended DEX Option-Type with an unknown flag set to 1 MUST ignore the corresponding optional field.

IOAM-Trace-Type: 24-bit identifier that specifies which IOAM data types are used and the corresponding IOAM-Data-Fields should be exported. The format of this field is as defined in [RFC9197].

L: 1-bit Loss flag for Packet Loss Measurement as described in Section 6.1.

D: 1-bit Delay flag for Single Packet Delay Measurement as described in Section 6.2.

Reserved: 6-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, starting from the most significant bit, that are enabled in the Extension-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.

6. The IOAM Operation

The extended DEX Option-Type SHOULD support to perform both performance measurement and IOAM trace monitoring concurrently. While both performance measurement and IOAM trace monitoring are implemented concurrently, an IOAM encapsulating node MUST incorporate the extended DEX Option-Type into all the traffic of interest it forwards. For performance measurement, an IOAM encapsulating node MUST mark every packet it forwards in "L" and "D" flag of the extended DEX Option-Type; for IOAM trace monitoring, only a subset of the packets are 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 DEX Option-Type so that each node along the path needs to generate the specified IOAM data exported to the receiving entity; 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 DEX Option-Type, such that each node along the path needs not generate the IOAM data exported to the receiving entity.

6.1. 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).

6.2. 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. 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. 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. In the implementation, the timestamps along with Flow ID and Sequence Number (if it exists) can be sent out to the receiving entity that is responsible for the calculation.

6.3. 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 draft 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 pseudo-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 is

the Nth power of C (K, M)

and the total possible sample space not duplicate is

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

Theoritically, the non collision probability is calculated as the total possible sample space not duplicate divided by the total possible sample space.

Take K=32nd power of 2, N=100, M=100 as an example, and the non collision probability is 0.9885. That is to say, when generating 10000 concurrent flows, there might be 115 measured flow identifiers incurring a chance of collision. If K=20th power of 2 is taken, which corresponds to 20-bit Flow ID space, the collision probability will drastically increases to approximately 100%. In practical deployment scenarios of large-scale networks, the simultaneous 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.

7. IANA Considerations

7.1. 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:

TBD-type IOAM Extended DEX Option Type.

If possible, IANA is requested to allocate code point 5 (TBD-type).

7.2. Reserved Field

IANA is requested to allocate the following 2-bit flags for performance measurement from the 8-bit Reserved field created by IANA.

Bit 0 (the most significant bit): 1-bit Loss flag for Packet Loss Measurement.

Bit 1: 1-bit Delay flag for Packet Delay Measurement.

7.3. IOAM DEX Flags

IANA has created the "IOAM DEX Flags" registry. This registry includes 8 flag bits. Allocation is based on the "IETF Review" procedure defined in [RFC8126].

7.4. IOAM DEX Extension-Flags

IANA has created the "IOAM DEX Extension-Flags" registry. This registry includes 8 flag bits. Bit 0 (the most significant bit) and bit 1 in the registry are allocated by [RFC9326] and described in Section 5.

IANA is requested to allocate bit 2 as Measurement Period Number in the registry and described in Section 5.

Allocation of the other bits should be performed based on the "IETF Review" procedure defined in [RFC8126].

8. Performance Considerations

The extended DEX Option-Type triggers IOAM data (including IOAM trace data and performance measurement data) to be collected and/or exported packets to be exported to a receiving entity. In some cases, this may impact the receiving entity's performance.

Therefore, the performance impact of these exported packets is limited by taking two measures: at the encapsulating nodes by selective DEX encapsulation and at the transit nodes by limiting exporting rate, which are detailed in [RFC9326]. These two measures ensure that direct exporting is used at a rate that does not significantly affect the network bandwidth and does not overload the receiving entity.

When performance measurement is implemented based on the Alternate-Marking Method, and in Hop-by-Hop mode for loss measurement, every node along the path only exports 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. Meanwhile, an IOAM encapsulating node only needs to select a very small subset of the packets that are forwarded for IOAM trace monitoring (e.g., 1/10000 of all the traffic), so the amount of exported data is significantly reduced to mitigate the network and the receiving entity. In addition, compared with IOAM DEX Option-Type for packet loss calculation, due to a significant reduction in the number of exported packets, the receiving entity needs much less processing overhead to correlate these counters for packet loss computation.

9. 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.

10. References

10.1. 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>.

10.2. Informative References

[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

Frank Brockners
Cisco

Email: fbrockne@cisco.com

Haoyu Song
Futurewei

Email: haoyu.song@futurewei.com

Giuseppe Fioccola
Huawei

Email: giuseppe.fioccola@huawei.com

Aijun Wang
China Telecom

Email: wangaj3@chinatelecom.cn