Workgroup: IPPM
Internet-Draft:
draft-song-ippm-postcard-based-telemetry-13
Published: 16 August 2022
Intended Status: Informational
Expires: 17 February 2023
A H. Song G. Mirsky
u Futurewei Technologies Ericsson
t
h
o
r
s
:
C. Filsfils A. Abdelsalam T. Zhou
Cisco Systems, Inc. Cisco Systems, Inc. Huawei
Z. Li T. Graf G. Mishra J. Shin K. Lee
Huawei Swisscom Verizon Inc. SK Telecom LG U+

Marking-based Direct Export for On-path Telemetry

Abstract

The document describes a packet-marking variation of the IOAM DEX option, referred to as PBT-M (i.e., Postcard-Based Telemetry by Marking). Similar to IOAM DEX, PBT-M does not carry the telemetry data in user packets but send the telemetry data through a dedicated packet. Unlike IOAM DEX, PBT-M does not require an extra instruction header. However, PBT-M raises some unique issues that need to be considered. This document formally describes the high level scheme and cover the common requirements and issues when applying PBT-M in different networks. PBT-M is complementary to the other on-path telemetry schemes such as IOAM trace and E2E 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 17 February 2023.

Copyright Notice

Table of Contents

1. Motivation

To gain detailed data plane visibility to support effective network OAM, it is essential to be able to examine the trace of user packets along their forwarding paths. Such on-path flow data reflect the state and status of each user packet's real-time experience and provide valuable information for network monitoring, measurement, and diagnosis.

The telemetry data include but not limited to the detailed forwarding path, the timestamp/latency at each network node, and, in case of packet drop, the drop location, and the reason. The emerging programmable data plane devices allow user-defined data collection or conditional data collection based on trigger events. Such on-path flow data are from and about the live user traffic, which complements the data acquired through other passive and active OAM mechanisms such as IPFIX [RFC7011] and ICMP [RFC2925].

On-path telemetry was developed to cater to the need of collecting on-path flow data. There are two basic modes for on-path telemetry: the passport mode and the postcard mode. In the passport mode which is represented by IOAM trace option [I-D.ietf-ippm-ioam-data], each node on the path adds the telemetry data to the user packets (i.e., stamp the passport). The accumulated data-trace carried by user packets are exported at a configured end node. In the postcard mode which is represented by IOAM direct export option (DEX) [I-D.ietf-ippm-ioam-direct-export], each node directly exports the telemetry

data using an independent packet (i.e., send a postcard) to avoid carrying the data with user packets. The postcard mode is complementary to the passport mode.

IOAM DEX uses an instruction header to explicitly instruct the telemetry data to be collected. This document describes another variation of the postcard mode on-path telemetry, PBT-M. Unlike IOAM DEX, PBT-M does not require a telemetry instruction header. However, PBT-M has unique issues that need to be considered. This document discusses the challenges and their solutions which are common to the high-level scheme of PBT-M.

2. PBT-M: Marking-based Direct Export for On-path Telemetry

As the name suggests, PBT-M only needs a marking-bit in the existing headers of user packets to trigger the telemetry data collection and export. The sketch of PBT-M is as follows. If on-path data need to be collected, the user packet is marked at the path head node. At each PBT-M-aware node, if the mark is detected, a postcard (i.e., the dedicated OAM packet triggered by a marked user packet) is generated and sent to a collector. The postcard contains the data requested by the management plane. The requested data are configured by the management plane. Once the collector receives all the postcards for a single user packet, it can infer the packet's forwarding path and analyze the data set. The path end node is configured to un-mark the packets to its original format if necessary.

The overall architecture of PBT-M is depicted in Figure 1.

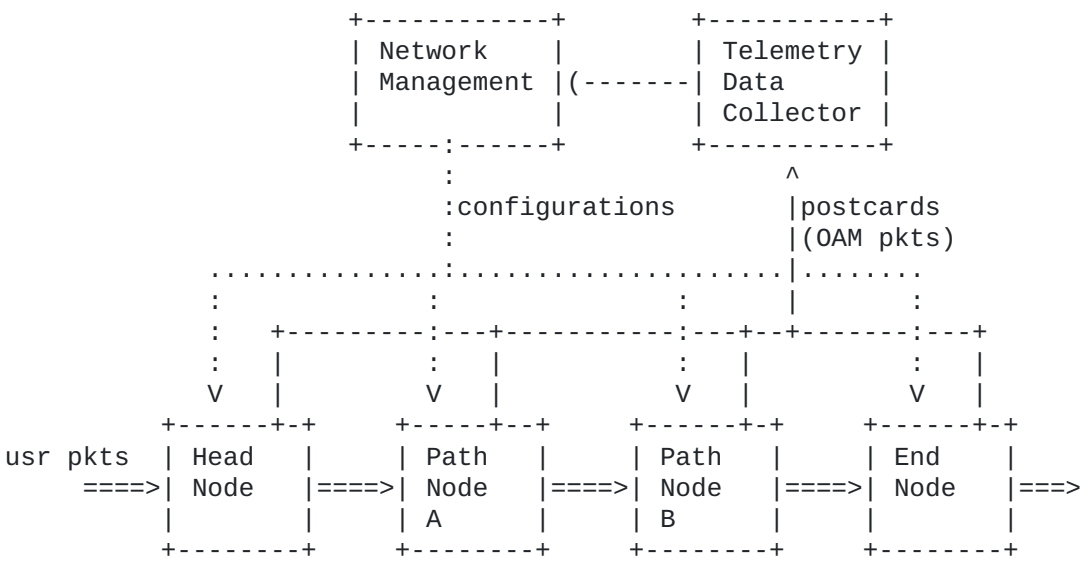

Figure 1: Architecture of PBT-M

The advantages of PBT-M are summarized as follows.

*1: PBT-M avoids augmenting user packets with new headers and the signaling for telemetry data collection remains in the data plane.

*2: PBT-M is extensible for collecting arbitrary new data to support possible future use cases. The data set to be collected can be configured through the management plane or control plane.

*3: PBT-M can avoid interfering with the normal forwarding. The collected data are free to be transported independently through in-band or out-of-band channels. The data collecting, processing, assembly, encapsulation, and transport are, therefore, decoupled from the forwarding of the corresponding user packets and can be performed in data-plane slow-path if necessary.

*4: For PBT-M, the types of data collected from each node can vary depending on application requirements and node capability.

*5: PBT-M makes it easy to secure the collected data without exposing it to unnecessary entities. For example, both the configuration and the telemetry data can be encrypted and/or authenticated before being transported, so passive eavesdropping and a man-in-the-middle attack can both be deterred.

*6: Even if a user packet under inspection is dropped at some node in the network, the postcards collected from the preceding nodes are still valid and can be used to diagnose the packet drop location and reason.

*7: Raw data can be processed or aggregated in data plane to reduce the exporting traffic load.

3. New Challenges

Although PBT-M has some unique features compared to the passport mode telemetry and the instruction-based IOAM DEX, it introduces a few new challenges.

*Challenge 1 (Packet Marking): A user packet needs to be marked to trigger the path-associated data collection. Since PBT-M does not augment user packets with any new header fields, it needs to reserve or reuse bits from the existing header fields. This raises a similar issue as in the Alternate Marking Scheme [RFC8321]

*Challenge 2 (Configuration): Since the packet header will not carry telemetry instructions anymore, the data plane devices need to be configured to know what data to collect. However, in general, the forwarding path of a flow packet (due to ECMP or dynamic routing) is unknown beforehand (note that there are some notable exceptions, such as segment routing). If the per-flow customized data collection is required, configuring the data set for each flow at all data plane devices might be expensive in terms of configuration load and data plane resources.

* Challenge 3 (Data Correlation): Due to the variable transport latency, the dedicated postcard packets for a single packet may arrive at the collector out of order or be dropped in networks for some reason. In order to infer the packet forwarding path, the collector needs some information from the postcard packets to identify the user packet affiliation and the order of path node traversal.

* Challenge 4 (Load Overhead): Since each postcard packet has its header, the overall network bandwidth overhead of PBT-M can be high. A large number of postcards could add processing pressure on data collecting servers. That can be used as an attack vector for DoS.

4. Design Considerations

To address the above challenges, we propose several design details of PBT-M.

4.1. Packet Marking

To trigger the path-associated data collection, usually, a single bit from some header field is sufficient. While no such bit is available, other packet-marking techniques are needed. We discuss several possible application scenarios.

* IPv4. Alternate Marking (AM) [RFC8321] is an IP flow performance measurement framework that also requires a single bit for packet coloring. The difference is that AM does in-network measurement while PBT-M only collects and exports data at network nodes (i.e., the data analysis is done at the collector rather than in the network nodes). AM suggests to use some reserved bit of the Flag field or some unused bit of the TOS field. Actually, AM can be considered a sub-case of PBT-M, so that the same bit can be used for IOAM Marking. The management plane is responsible for configuring the actual operation mode.

* SFC NSH. The OAM bit in the NSH header can be used to trigger the on-path data collection [RFC8300]. PBT-M does not add any other metadata to NSH.

* MPLS. Instead of choosing a header bit, we take advantage of the synonymous flow label [I-D.bryant-mpls-synonymous-flow-labels] approach to mark the packets. A synonymous flow label indicates the on-path data should be collected and forwarded through a postcard.

* SRv6: A flag bit in SRH can be reserved to trigger the on-path data collection [I-D.song-6man-srv6-pbt]. SRv6 OAM [I-D.ietf-6man-spring-srv6-oam] has adopted the O-bit in SRH flags as the marking bit to trigger the telemetry.

4.2. Flow Path Discovery

In case the path that a flow traverses is unknown in advance, all PBT-M-aware nodes should be configured to react to the marked packets by exporting some basic data, such as node ID and TTL before

a data set template for that flow is configured. This way, the management plane can learn the flow path dynamically.

If the management plane wants to collect the on-path data for some flow, it configures the head node(s) with a probability or time interval for the flow packet marking. When the first marked packet is forwarded in the network, the PBT-M-aware nodes will export the basic data set to the collector. Hence, the flow path is identified. If other data types need to be collected, the management plane can further configure the data set's template to the target nodes on the flow's path. The PBT-M-aware nodes collect and export data accordingly if the packet is marked and a data set template is present.

If the flow path is changed for any reason, the new path can be quickly learned by the collector. Consequently, the management plane controller can be directed to configure the nodes on the new path. The outdated configuration can be automatically timed out or explicitly revoked by the management plane controller.

4.3. Packet Identity for Export Data Correlation

The collector needs to correlate all the postcard packets for a single user packet. Once this is done, the TTL (or the timestamp, if the network time is synchronized) can be used to infer the flow forwarding path. The key issue here is to correlate all the postcards for the same user packet.

The first possible approach includes the flow ID plus the user packet ID in the OAM packets. For example, the flow ID can be the 5-tuple IP header of the user traffic, and the user packet ID can be some unique information pertaining to a user packet (e.g., the sequence number of a TCP packet).

If the packet marking interval is large enough, the flow ID is enough to identify a user packet. As a result, it can be assumed that all the exported postcard packets for the same flow during a short time interval belong to the same user packet.

Alternatively, if the network is synchronized, then the flow ID plus the timestamp at each node can also infer the postcard affiliation. However, some errors may occur under some circumstances. For example, two consecutive user packets from the same flows are marked, but one exported postcard from a node is lost. It is difficult for the collector to decide to which user packet the remaining postcard is related. In many cases, such a rare error has no catastrophic consequence. Therefore it is tolerable.

4.4. Control the Load

PBT-M should not be applied to all the packets all the time. It is better to be used in an interactive environment where the network telemetry applications dynamically decide which subset of traffic is under scrutiny. The network devices can limit the packet marking rate through sampling and metering. The postcard packets can be distributed to different servers to balance the processing load.

It is important to understand that the total amount of data exported by PBT-M is identical to that of IOAM trace option. The only extra overhead is the packet header of the postcards. In the case of IOAM trace option, it carries the data from each node throughout the path to the end node before exporting the aggregated data. On the other hand, PBT-M directly exports local data. The overall network bandwidth impact depends on the network topology and scale, and in some cases PBT-M could be more bandwidth efficient.

5. Implementation Recommendation

5.1. Configuration

Access lists with an optional sampler, [RFC5476], should be configured and attached at the ingress of the IOAM encapsulation node's to select the intended flows for IOAM.

Based on the PBT-M, the flow data should be exported at each transit node and at the end edge node with IPFIX [RFC7011].

5.2. Data Export

The data decomposition can be achieved on the PBT-M-aware node exporting the data or on the IPFIX data collection. [I-D.spiegel-ippm-ioam-rawexport] describes how data is being exported when decomposed at IPFIX data collection. When being decomposed on the PBT-M-aware node the data can be aggregated according to section 5 of [RFC7015]. The following IPFIX entities are of interest to describe the relationship to the forwarding topology and the control-plane.

* node id and egressInterface(IE14) describes on which node which logical egress interfaces have been used to forward the packet.

* Node id and egressPhysicalInterface(253) describes on which node which physical egress interfaces have been used to forward the packet.

* Node id and ipNextHopIPv4Address(IE15) or ipNextHopIPv6Address(IE62), describes the forwarding path to which next-hop IP address.

* Node id and mplsTopLabelIPv4Address(IE47) or srhActiveSegmentIPv6 from [I-D.tgraf-opsawg-ipfix-srv6-srh] describes the forwarding path to which MPLS top label IPv4 address or SRv6 active segment.

* BGP communities are often used for setting a path priority or service selection. bgpDestinationExtendedCommunityList(488) or bgpDestinationCommunityList(485) or bgpDestinationLargeCommunityList(491) describes which group of prefixes have been used to forward the packet.

* Node id and destinationIPv4Address(13), destinationTransportPort(11), protocolIdentifier (4) and sourceIPv4Address(IE8) describes the forwarding path on each node from each IPv4 source address to a specific application in the network.

*In order to distinguish wherever the packet has been export due to the packet marking or not, in case of SRv6, srhFlagsIPv6 as described in section 4.1 of [I-D.tgraf-opsawg-ipfix-srv6-srh] can be added to the data export.

6. Use Cases

The MPLS Design Team has been investigating extensibility options for the MPLS data plane.

The challenge has been to continue to support existing MPLS architecture, backwards compatibility as well as not excessively increase the depth of the MPLS label stack with a variety of functional SPL labels and NAI indicators similar in concept to the MPLS Entropy label ELI, EL added to the label stack, as well as the MPLS extension headers being in Stack or post stack.

Reference Augmented Forwarding (RAF) [I-D.raszuk-mpls-raf-fwk] utilizes In Stack Data (ISD) with parity to Entropy Label stack {TL,RFI,RFV,AL} and control plane extension to distribute special network actions and forwarding behaviors.

Reference Augmented Forwarding (RAF) keeps the ISD and PSD stack depth in check by using an alternative means of carrying the IOAM data using IGP control plane extension TLV to carry the data to provide In-Situ IOAM on path telemetry using the postcard based telemetry.

The MPLS Design Team may come up with other alternatives to carry IOAM data such as the IGP extension mentioned and maybe other solutions, which will heavily rely on the the postcard based solution.

With Segment Routing SR-MPLS and SRv6 as Maximum SID Depth(MSD) as well as PMTU in SR Policy are critical issues for SR path instantiation by a controller, postcard based telemetry will become a critical solution to ensure that IOAM telemetry can be viable for operators by eliminating IOAM data from being carried in-situ in the SR-TE policy path.

This draft provides a critical optimization that fills the gaps with IOAM DEX related to packet marking triggers using existing mechanisms as well as flow path discovery mechanisms to avoid configuration of on path data plane node complexity and helps mitigate SR MSD and PMTU issues.

7. Security Considerations

Several security issues need to be considered.

*Eavesdrop and tamper: the postcards can be encrypted and authenticated to avoid such security threats.

*DoS attack: PBT-M can be limited to a single administrative domain. The mark must be removed at the egress domain edge. The node can rate-limit the extra traffic incurred by postcards.

8. IANA Considerations

No requirement for IANA is identified.

9. Contributors

TBD.

10. Acknowledgments

We thank Robert Raszuk, Alfred Morton who provided valuable suggestions and comments helping improve this draft.

11. Informative References

[I-D.bryant-mpls-synonymous-flow-labels]
Bryant, S., Swallow, G., Sivabalan, S., Mirsky, G., Chen, M., and Z. Li, "RFC6374 Synonymous Flow Labels", Work in Progress, Internet-Draft, draft-bryant-mpls-synonymous-flow-labels-01, 4 July 2015, <https://www.ietf.org/archive/id/draft-bryant-mpls-synonymous-flow-labels-01.txt>.

[I-D.ietf-6man-spring-srv6-oam] Ali, Z., Filsfils, C., Matsushima, S., Voyer, D., and M. Chen, "Operations, Administration, and Maintenance (OAM) in Segment Routing Networks with IPv6 Data plane (SRv6)", Work in Progress, Internet-Draft, draft-ietf-6man-spring-srv6-oam-13, 23 January 2022, <https://www.ietf.org/archive/id/draft-ietf-6man-spring-srv6-oam-13.txt>.

[I-D.ietf-ippm-ioam-data] Brockners, F., Bhandari, S., and T. Mizrahi, "Data Fields for In-situ OAM", Work in Progress, Internet-Draft, draft-ietf-ippm-ioam-data-17, 13 December 2021, <https://www.ietf.org/archive/id/draft-ietf-ippm-ioam-data-17.txt>.

[I-D.ietf-ippm-ioam-direct-export]
Song, H., Gafni, B., Brockners, F., Bhandari, S., and T. Mizrahi, "In-situ OAM Direct Exporting", Work in Progress, Internet-Draft, draft-ietf-ippm-ioam-direct-export-09, 15 June 2022, <https://www.ietf.org/archive/id/draft-ietf-ippm-ioam-direct-export-09.txt>.

[I-D.raszuk-mpls-raf-fwk]
Raszuk, R., "Framework of MPLS Reference Augmented Forwarding", Work in Progress, Internet-Draft, draft-raszuk-mpls-raf-fwk-00, 25 April 2022, <https://www.ietf.org/archive/id/draft-raszuk-mpls-raf-fwk-00.txt>.

[I-D.song-6man-srv6-pbt]
Song, H., "Support Postcard-Based Telemetry for SRv6 OAM", Work in Progress, Internet-Draft, draft-song-6man-

srv6-pbt-01, 14 October 2019, <https://www.ietf.org/archive/id/draft-song-6man-srv6-pbt-01.txt>.

[I-D.spiegel-ippm-ioam-rawexport] Spiegel, M., Brockners, F., Bhandari, S., and R. Sivakolundu, "In-situ OAM raw data export with IPFIX", Work in Progress, Internet-Draft, draft-spiegel-ippm-ioam-rawexport-06, 21 February 2022, <https://www.ietf.org/archive/id/draft-spiegel-ippm-ioam-rawexport-06.txt>.

[I-D.tgraf-opsawg-ipfix-srv6-srh] Graf, T., Claise, B., and P. Francois, "Export of Segment Routing IPv6 Information in IP Flow Information Export (IPFIX)", Work in Progress, Internet-Draft, draft-tgraf-opsawg-ipfix-srv6-srh-05, 24 July 2022, <https://www.ietf.org/archive/id/draft-tgraf-opsawg-ipfix-srv6-srh-05.txt>.

[RFC2925] White, K., "Definitions of Managed Objects for Remote Ping, Traceroute, and Lookup Operations", RFC 2925, DOI 10.17487/RFC2925, September 2000, <https://www.rfc-editor.org/info/rfc2925>.

[RFC5476] Claise, B., Ed., Johnson, A., and J. Quittek, "Packet Sampling (PSAMP) Protocol Specifications", RFC 5476, DOI 10.17487/RFC5476, March 2009, <https://www.rfc-editor.org/info/rfc5476>.

[RFC7011] Claise, B., Ed., Trammell, B., Ed., and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information", STD 77, RFC 7011, DOI 10.17487/RFC7011, September 2013, <https://www.rfc-editor.org/info/rfc7011>.

[RFC7015] Trammell, B., Wagner, A., and B. Claise, "Flow Aggregation for the IP Flow Information Export (IPFIX) Protocol", RFC 7015, DOI 10.17487/RFC7015, September 2013, <https://www.rfc-editor.org/info/rfc7015>.

[RFC8300] Quinn, P., Ed., Elzur, U., Ed., and C. Pignataro, Ed., "Network Service Header (NSH)", RFC 8300, DOI 10.17487/RFC8300, January 2018, <https://www.rfc-editor.org/info/rfc8300>.

[RFC8321] Fioccola, G., Ed., Capello, A., Cociglio, M., Castaldelli, L., Chen, M., Zheng, L., Mirsky, G., and T. Mizrahi, "Alternate-Marking Method for Passive and Hybrid Performance Monitoring", RFC 8321, DOI 10.17487/RFC8321, January 2018, <https://www.rfc-editor.org/info/rfc8321>.

Authors' Addresses

Haoyu Song
Futurewei Technologies
2330 Central Expressway
Santa Clara, 95050,
United States of America

Email: hsong@futurewei.com

Greg Mirsky
Ericsson

Email: gregimirsky@gmail.com

Clarence Filsfils
Cisco Systems, Inc.
Belgium

Email: cfilsfil@cisco.com

Ahmed Abdelsalam
Cisco Systems, Inc.
Italy

Email: ahabdels@cisco.com

Tianran Zhou
Huawei
156 Beiqing Road
Beijing, 100095
P.R. China

Email: zhoutianran@huawei.com

Zhenbin Li
Huawei
156 Beiqing Road
Beijing, 100095
P.R. China

Email: lizhenbin@huawei.com

Thomas Graf
Swisscom
Switzerland

Email: thomas.graf@swisscom.com

Gyan Mishra
Verizon Inc.

Email: hayabusagsm@gmail.com

Jongyoon Shin
SK Telecom
South Korea

Email: jongyoon.shin@sk.com

Kyungtae Lee
LG U+
South Korea

Email: coolee@lguplus.co.kr