Workgroup: IPPM
Internet-Draft:
draft-song-ippm-postcard-based-telemetry-16
Published: 30 May 2023
Intended Status: Informational
Expires: 1 December 2023
Authors: H. Song G. Mirsky T. Zhou
Futurewei Technologies Ericsson Huawei
Z. Li T. Graf G. Mishra J. Shin K. Lee
Huawei Swisscom Verizon Inc. SK Telecom LG U+

On-Path Telemetry using Packet Marking to Trigger Dedicated OAM Packets

Abstract

The document describes an on-path telemetry method using packet-marking, referred to as PBT-M. Similar to IOAM DEX, PBT-M does not carry the telemetry data in user packets but sends the telemetry data through a dedicated packet. However, PBT-M does not require an extra instruction header but claims a bit in existing header fields or uses some other equivalent means as a trigger for telemetry data processing and collection. Due to this feature, PBT-M raises some unique issues that need to be considered for its application in different networks. This document describes the high level scheme, summarizes the common requirements and issues, and provides recommendations for solutions. PBT-M is complementary to the other on-path telemetry schemes.

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.

Status of This Memo

This Internet-Draft will expire on 1 December 2023.

Table of Contents

1. Introduction

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 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 [RFC4560].

This document describes PBT-M, a new on-path telemetry technique which complements IOAM Trace [RFC9197] and IOAM DEX [RFC9326]. PBT-M does not require a telemetry instruction header but a trigger bit in some existing header fields or some equivalent means. Due to this feature, the seemingly simple scheme raises some unique issues that need to be considered for successful application. This document serves as a central location to archive the challenges common to PBT-M and provides solution recommendations, aiming to eliminate duplicated efforts when applying PBT-M in different network scenarios.

2. PBT-M

As the name suggests, PBT-M only needs a marking-bit in the existing headers of user packets (or some equivalent means) to trigger the telemetry data collection and export. The sketch of PBT-M is as follows. If some on-path data need to be collected for a user packet, the user packet is marked at the path head node. At each PBT-M-aware node on the path, if the mark is detected, a telemetry data packet (i.e., the dedicated OAM packet triggered by the marked user packet) is generated and sent to a collector. Meanwhile, the original user packet is forwarded without delay and alteration. The telemetry data packet contains the data requested by the management plane. The requested data are configured by the management plane. Once the collector receives the postcards for a single user packet from different path nodes, 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.

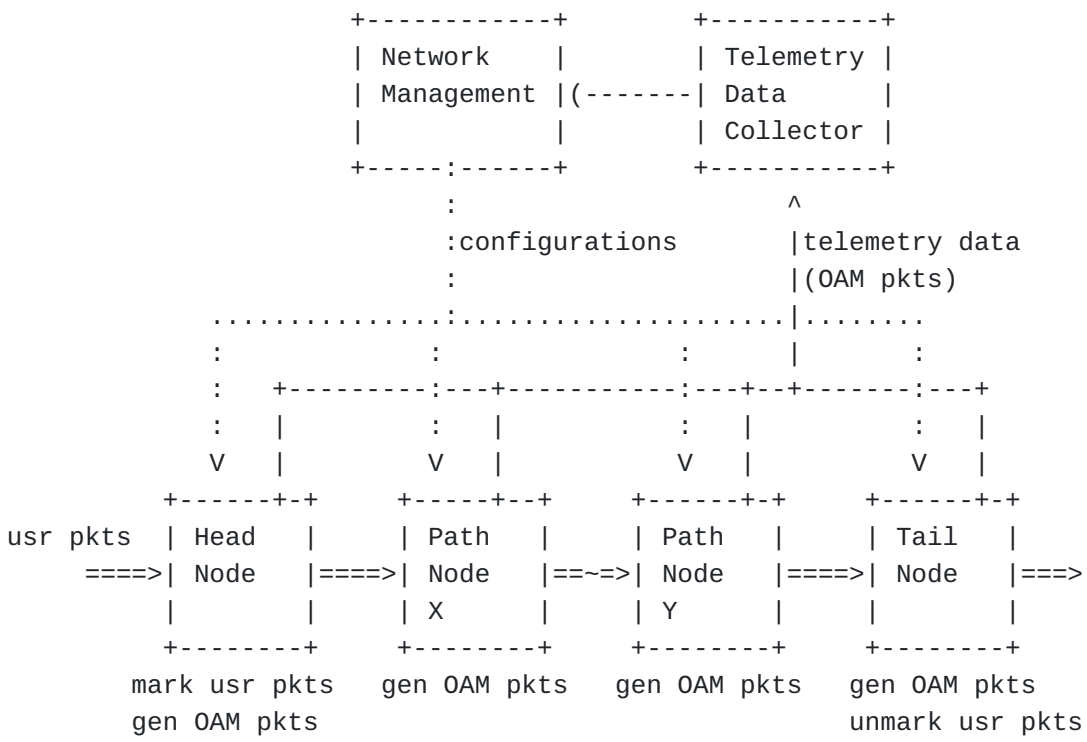

Figure 1: PBT-M Architecture

The advantages of PBT-M are summarized as follows.

*1: PBT-M avoids the need to augment user packets with new headers while the signaling for telemetry data collection remains in the data plane.

*2: PBT-M is extensible for collecting arbitrary new data types 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 is not intrusive to the normal forwarding of user traffic. 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 even be performed in data-plane slow-path if necessary.

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

*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 incomplete set of OAM packets collected from the preceding nodes are still valid and can be used to diagnose the packet drop location and reason.

*7: Since the OAM packets are generated and exported separately, raw data can be processed or aggregated in data plane to reduce the exporting traffic load and post-processing burden.

3. Requirements and Challenges

Although PBT-M is simple and has many advantages, it also introduces a few new requirements and challenges due to its unique feature.

OAM Packet Trigger: A user packet needs to be marked to trigger the on-path data collection. Since PBT-M aims to avoid the need to augment user packets with new headers, it needs to reserve or reuse a single bit from the existing header fields, or engage with some other equivalent means. This raises a similar issue as in the Alternate Marking Scheme [RFC9341]

Data Plane Configuration: Since the packet header will not carry explicit telemetry instructions anymore, the data plane needs to be configured to know where and 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 desired, configuring the data set for each flow at all data plane devices can be expensive in terms of configuration load and data plane resources.

Data Export: A standard and extensible OAM packet encoding and export protocol is needed, applicable to any application scenarios and in any networks. This can also simplify the data consumption and post processing.

Data Correlation: Due to the variable transport latency, the dedicated OAM 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 OAM packets to identify the user packet affiliation and the order of path node traversal. Data

correlation is especially challenging for PBT-M due to the lack of facilitating metadata.

Security: Last but not the least, security issues need to be considered for PBT-M. PBT-M makes it easier to trigger data collection and more nodes participate in data exporting, so a potential attack is easier to launch and more vulnerable points are involved for PBT-M than for the other OPT techniques. For example, since each OAM packet has its header, the overall network bandwidth overhead of PBT-M is higher. A large number of OAM packets could add data collecting pressure on network devices and data processing pressure on data collecting servers, leading to performance concerns and a potential attack vector for DoS. While many measures can be taken to optimize the performance, we defer the further security considerations in Section 6.

4. Design Considerations and Recommendations

To address the above requirements and challenges, we propose the considerations and recommendations for implementing and applying PBT-M.

4.1. Packet Marking

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

* IPv4. Alternate Marking (AM) [RFC9341] is an IP flow performance measurement framework that also requires a single bit for packet coloring. The difference is that AM conducts in-network measurements such as latency and packet loss rate based on the bit alternating patterns while PBT-M only collects and exports data at each network nodes when the trigger bit is set. AM suggests to use some reserved bit of the Flag field or some unused bit of the TOS field. PBT-M can share the same bit with AM, and rely on the management plane to configure 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. The ongoing MPLS Network Action (MNA) work [I-D.andersson-mpls-mna-fwk] may provide new in-stack headers for

MNAs. A bit can be claimed for PBT-M as proposed in [I-D.song-mpls-flag-based-opt].

*SRv6: A flag bit in SRH can be reserved to trigger the on-path data collection [I-D.song-6man-srv6-pbt]. SRv6 OAM [RFC9259] has adopted the O-bit in SRH flags as the marking bit to trigger the telemetry.

The marking method for other protocols (e.g., IPv6) is subject to further study and is out of scope of this document.

4.2. Flow Path Discovery

In case the path that a flow traverses is unknown in advance, all PBT-M-aware nodes in an application domain should by default 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 from the postcard packets and delay the decision on collecting more comprehensive data by configuring only the relevant nodes.

If the management plane wants to collect the on-path data for some flow, in order to reduce the data redundancy, workload for network devices and data collectors, and network bandwidth consumption, it is unnecessary to mark every flow packet. Instead, it is recommended to configure the head node(s) with a sampling 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 OAM Packet Correlation

For a marked user packet, each PBT-M-aware node will send an independent OAM packet. The collector needs to correlate all the OAM packets corresponding to the 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. Due to the lack of some explicit

identifiers as in IOAM DEX, the OAM packet correlation needs to take different measures.

The first possible approach is to require that the exported data in the OAM packets must include the flow ID plus the user packet ID extracted for the marked user packet. 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). Alternatively, a hashing digest of the invariant part of the packet during the forwarding (e.g., excluding the TTL and checksum fields of an IPv4 header) can serve as both the flow ID and the packet ID. The possibility of hash collision is negligible since the set of correlated OAM packets can only appear in a very short time frame.

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

The second approach requires the network to be synchronized. In this case, the flow ID plus the timestamp at each node can also infer the OAM packet affiliation. For the OAM packets from the same flow, the collector only needs to sort them based on the timestamp. However, some errors may occur under some circumstances. For example, two consecutive user packets from the same flows are marked, but one exported OAM packet from a node is lost. It is difficult for the collector to decide to which user packet the remaining OAM packet is related. In many cases, such a rare error has no catastrophic consequence. Therefore it is tolerable. Again, a larger sampling gap can mitigate this problem.

4.4. Load Control

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 OAM packets can be distributed to different servers to balance the processing load.

Because PBT-M sends telemetry data by dedicated OAM packets, it allows data aggregation and compression. Each node can process the generated raw data according to the configured local data-export policies. Such policies may specify how raw data is used to calculate performance metrics, e.g., max, min, mean, percentile, etc.

It is also possible to customize the data collection on each node to reduce the data exporting load. For example, if only end-to-end latency rather than the per-hop delay is of interest to the application, then only the head and tail nodes need to be configured to export the timestamps while the other on-path nodes are just configured to collect the other routine data.

Combining the above recommendations, PBT-M can be made flexible and efficient.

4.5. Incremental Deployment

Given that even an incomplete set of OAM packets for a user packet are useful for network monitoring and measurement, PBT-M is ideal for incremental deployment. A node which is node updated to support PBT-M SHOULD ignore the trigger and continue to forward any marked packet normally.

It is also possible for a node to not export certain data items for various reasons (e.g., node busy or data unavailable).

4.6. Node Configuration

Access lists with an optional sampler, [RFC5476], should be configured and attached at the ingress of the PBT-M encapsulation node's to select the intended flows for PTB-M. A flow packet sampling policy meeting the application requirement should also be configured.

A telemetry data template pertaining to a flow or a node should be configured to define the type and format of the data to be collected.

The OAM packet format should also be configured. Particularly, the flow data should be exported at each participating node using IPFIX [RFC7011].

4.7. 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.

5. Use Cases

PBT-M has been used for SRv6 OAM [RFC9259]. Currently, the MPLS Open Design Team is investigating network action support on the MPLS data plane [I-D.andersson-mpls-mna-fwk]. 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 special purpose labels and network action 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.

The MPLS Design Team may come up with other alternatives to carry network actions and PBT-M can be supported as a use case.

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, PBT-M can become a critical solution to ensure that OPT can be viable for operators by eliminating telemetry 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 data plane complexity and helps mitigate SR MSD and PMTU issues.

6. Security Considerations

Several security issues need to be considered.

*Eavesdrop and tamper: the OAM packets can be encrypted and authenticated to avoid such security threats. Since the telemetry data are not required to be attached to the user packet in real time, PBT-M has more time and freedom to process the collected data. If necessary, the device slow-path can be used.

*DoS attack: PBT-M can be limited to a single administrative domain. The mark must be removed at the egress domain edge. The telemetry data can be aggregated and accumulated. The node can rate-limit the extra traffic incurred by OAM packets. In the worst case, the node can ignore the data collecting request from the marked packets.

7. IANA Considerations

No requirement for IANA is identified.

8. Contributors

9. Acknowledgments

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

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>.

10.2. Informative References

[I-D.andersson-mpls-mna-fwk] Andersson, L., Bryant, S., Bocci, M., and T. Li, "MPLS Network Actions Framework", Work in Progress, Internet-Draft, draft-andersson-mpls-mna-fwk-04, 27 June 2022, <https://datatracker.ietf.org/doc/html/draft-andersson-mpls-mna-fwk-04>.

[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://datatracker.ietf.org/doc/html/draft-bryant-mpls-synonymous-flow-labels-01>.

[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://datatracker.ietf.org/doc/html/draft-raszuk-mpls-raf-fwk-00>.

[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://datatracker.ietf.org/doc/html/draft-song-6man-srv6-pbt-01>.

[I-D.song-mpls-flag-based-opt] Song, H., Fioccola, G., and R. Gandhi, "Flag-based MPLS On Path Telemetry Network Actions", Work in Progress, Internet-Draft, draft-song-mpls-flag-based-opt-01, 9 March 2023, <https://datatracker.ietf.org/doc/html/draft-song-mpls-flag-based-opt-01>.

[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://datatracker.ietf.org/doc/html/draft-spiegel-ippm-ioam-rawexport-06>.

[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://datatracker.ietf.org/doc/html/draft-tgraf-opsawg-ipfix-srv6-srh-05>.

[RFC4560] Quittek, J., Ed. and K. White, Ed., "Definitions of Managed Objects for Remote Ping, Traceroute, and Lookup Operations", RFC 4560, DOI 10.17487/RFC4560, June 2006, <https://www.rfc-editor.org/info/rfc4560>.

[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>.

[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>.

[RFC9259] Ali, Z., Filsfils, C., Matsushima, S., Voyer, D., and M. Chen, "Operations, Administration, and Maintenance (OAM) in Segment Routing over IPv6 (SRv6)", RFC 9259, DOI 10.17487/RFC9259, June 2022, <https://www.rfc-editor.org/info/rfc9259>.

[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>.

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

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