IPPM Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 10, 2020

G. Mirsky
ZTE Corp.
W. Lingqiang
G. Zhui
ZTE Corporation

October 8, 2019

Hybrid Two-Step Performance Measurement Method
draft-mirsky-ippm-hybrid-two-step-04

Abstract

Development of, and advancements in, automation of network operations brought new requirements for measurement methodology. Among them is the ability to collect instant network state as the packet being processed by the networking elements along its path through the domain. This document introduces a new hybrid measurement method, referred to as hybrid two-step, as it separates the act of measuring and/or calculating the performance metric from the act of collecting and transporting network state.

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 April 10, 2020.

Table of Contents

1. Introduction

Successful resolution of challenges of automated network operation, as part of, for example, overall service orchestration or data center operation, relies on a timely collection of accurate information that reflects the state of network elements on an unprecedented scale. Because performing the analysis and act upon the collected information requires considerable computing and storage resources, the network state information is unlikely to be processed by the network elements themselves but will be relayed into the data storage facilities, e.g., data lakes. The process of producing, collecting network state information also referred to in this document as network telemetry, and transporting it for post-processing should work equally well with data flows or injected in the network test packets. RFC 7799 [RFC7799] describes a combination of elements of passive and active measurement as a hybrid measurement.

Several technical methods have been proposed to enable collection of network state information instantaneous to the packet processing, among them [P4.INT] and [I-D.ietf-ippm-ioam-data].

This document introduces Hybrid Two-Step (HTS) as a new hybrid measurement method that separates measuring or calculating the performance metric from the collecting and transporting this information. The Hybrid Two-Step method extends the two-step mode of Residence Time Measurement (RTM) defined in [RFC8169] to on-path network state collection and transport.

2. Conventions used in this document

2.1. Terminology

RTM Residence Time Measurement

ECMP Equal Cost Multipath

MTU Maximum Transmission Unit

HTS Hybrid Two-Step

Network telemetry - the process of collecting and reporting of network state

2.2. 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. Problem Overview

Performance measurements are meant to provide data that characterize conditions experienced by traffic flows in the network and possibly trigger operational changes (e.g., re-route of flows, or changes in resource allocations). Modifications to a network are determined based on the performance metric information available at the time that a change is to be made. The correctness of this determination is based on the quality of the collected metrics data. The quality of collected measurement data is defined by:

- the resolution and accuracy of each measurement;
- predictability of both the time at which each measurement is made and the timeliness of measurement collection data delivery for use.

Consider the case of delay measurement that relies on collecting time of packet arrival at the ingress interface and time of the packet transmission at the egress interface. The method includes recording a local clock value on receiving the first octet of an affected message at the device ingress, and again recording the clock value on transmitting the first byte of the same message at the device egress. In this ideal case, the difference between the two recorded clock times corresponds to the time that the message spent in traversing the device. In practice, the time that has been recorded can differ from the ideal case by any fixed amount and a correction can be applied to compute the same time difference taking into account the known fixed time associated with the actual measurement. In this way, the resulting time difference reflects any variable delay associated with queuing.

Depending on the implementation, it may be a challenge to compute the difference between message arrival and departure times and - on the fly - add the necessary residence time information to the same message. And that task may become even more challenging if the packet is encrypted. Implementations SHOULD NOT record a message departure time that may be significantly inaccurate in the same message, as the result of estimating the departure time that includes the variable time component (such as that associated with buffering and queuing of the message). A similar problem may cause a lower quality of, for example, information that characterizes utilization of the egress interface. If unable to obtain the data consistently, without variable delays for additional processing, information may not accurately reflect the state at the egress interface. To mitigate this problem [RFC8169] defined an RTM two-step mode.

Another challenge associated with methods that collect network state information into the actual data packet is the risk to exceed the Maximum Transmission Unit (MTU) size, especially if the packet traverses overlay domains or VPNs. Since the fragmentation is not available at the transport network, operators may have to reduce MTU size advertised to client layer or risk missing network state data for the part, most probably the latter part, of the path.

4. Theory of Operation

The HTS method consists of the two phases:

- performing a measurement or obtaining network state information, one or more than one type, on a node;
- collecting and transporting the measurement.

HTS uses HTS Trigger carried in a data packet or a specially constructed test packet. Nature of the HTS Trigger is transport network layer specific, and its description is outside the scope of this document. The packet that includes the HTS Trigger in this document also referred to as the trigger packet.

The HTS method uses the HTS Follow-up packet, in this document also referred to as the follow-up packet, to collect measurement and network state data from the nodes. The node that creates the HTS Trigger also generates the HTS Follow-up packet. The follow-up packet contains characteristic information, copied from the trigger packet, sufficient for participating HTS nodes to associate it with the original packet. The exact composition of the characteristic information is specific for each transport network, and its definition is outside the scope of this document. The follow-up packet also uses the same encapsulation as the data packet. If not payload but only network information used to load-balance flows in equal cost multipath (ECMP), use of the network encapsulation identical to the trigger packet should guarantee that the follow-up packet remains in-band, i.e., traverses the same set of network elements, with the original data packet with the HTS Trigger. Only one outstanding follow-up packet MUST be on the node for the given path. That means that if the node receives an HTS Trigger for the flow on which it still waits for the follow-up packet to the previous HTS Trigger, the node will originate the follow-up packet to transport the former set of the network state data and transmit it before it sends the follow-up packet with the latest collection of network state information.

4.1. Operation of the HTS Ingress Node

A node that originates the HTS Trigger is referred to as HTS ingress node. As stated, the ingress node originates the follow-up packet. The follow-up packet has the transport network encapsulation identical with the trigger packet followed by the HTS shim and one or more telemetry information elements encoded as Type-Length-Value {TLV}. Figure 1 displays the example of the follow-up packet format.

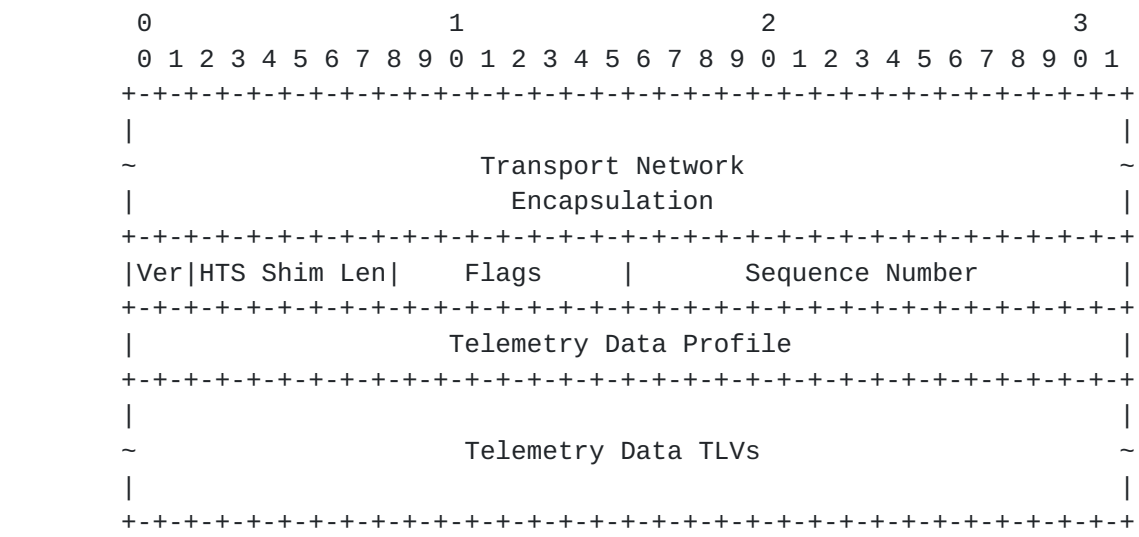

Figure 1: Follow-up Packet Format

Fields of the HTS shim are as follows:

Version (Ver) is the two-bits long field. It specifies the version of the HTS shim format. This document defines the format for the 0b00 value of the field.

HTS Shim Length is the six bits-long field. It defines the length of the HTS shim in bytes. The minimal value of the field is four bytes.

Flags is eight-bits long field. The format of the Flags field displayed in Figure 2.

Full (F) flag MUST be set to zero by the node originating the HTS follow-up packet and MUST be set to one by the node that does not add its telemetry data to avoid exceeding MTU size.

The node originating the follow-up packet MUST zero the Reserved field and ignore it on the receipt.

Sequence Number is 16 bits-long field. The value of the field reflects the number of the HTS follow-up packet in the sequence of the HTS follow-up packets originated in response to the same HTS trigger. The ingress node MUST set the value of the field to zero.

Telemetry Data Profile is the optional variable length field of bit-size flags. Each flag indicates requested type of telemetry data to be collected at the each HTS node. The increment of the field is four bytes with a minimum length of zero.

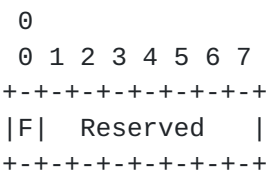

Figure 2: Flags Field Format

4.2. Operation of the HTS Transient Node

Upon receiving the trigger packet the HTS transient node MUST:

- copy the transport information;
- start the HTS Follow-up Timer for the obtained flow.

Upon receiving the follow-up packet the HTS transient node MUST:

- verify that the matching transport information exists and the Full flag is cleared, then stop the associated HTS Follow-up timer;
- collect telemetry data requested in the Telemetry Data Profile field or defined by the local HTS policy;
- if adding the collected telemetry would not exceed MTU, then append data into Telemetry Data TLVs field and transmit the follow-up packet;
- otherwise, set the value of the Full flag to one and transmit the received a follow-up packet;
- originate the new follow-up packet using the same transport information. The value of the Sequence Number field in the HTS shim MUST be set to the value of the field in the received follow-up packet incremented by one. Copy collected telemetry data and transmit the packet.

If the follow-up timer expires the transient node MUST:

- originate the follow-up packet using transport information associated with the expired timer;
- initialize the HTS shim by setting Version field to 0b00 and Sequence Number field to 0. Values of HTS Shim Length and Telemetry Data Profile fields MAY be set according to the local policy.

- copy telemetry information into Telemetry Data TLVs field and transmit the packet.

4.3. Operation of the HTS Egress Node

Upon receiving the trigger packet the HTS egress node MUST:

- copy the transport information;
- start the HTS Collection timer for the obtained flow.

When the egress node receives the follow-up packet for the known flow, i.e., the flow to which the Collection timer is running, the node MUST:

- copy telemetry information;
- restart the corresponding Collection timer.

When the Collection timer expires the egress relays the collected telemetry information for processing and analysis to a local or remote agent.

4.4. Considerations for HTS Timers

This specification defines two timers - HTS Follow-up and HTS Collection. Because for the particular flow there MUST be not more than one HTS Trigger, values of HTS timers bounded by the rate of the trigger generation for that flow.

4.5. Deploying HTS in a Multicast Network

Previous sections discussed the operation of HTS in a unicast network. Multicast services are important, and the ability to collect telemetry information is an invaluable component in delivering a high quality of experience. While the replication of data packets is necessary, replication of HTS follow-up packets is not. Replication of multicast data packets down a multicast tree may be set based on multicast routing information or explicit information included in the special header, as, for example, in Bit-Indexed Explicit Replication [RFC8296]. A replicating node processes HTS packet as defined below:

- the first transmitted multicast packet MUST be followed by the received corresponding HTS packet as described in Section 4.2;
- each consecutively transmitted copy of the original multicast packet MUST be followed by the new HTS packet originated by the

replicating node that acts as a transient HTS node when the Follow-up timer expired.

As a result, there are no duplicate copies of Telemetry Data TLV for the same pair of ingress and egress interfaces. At the same time, all ingress/egress pairs traversed by the given multicast packet reflected in their respective Telemetry Data TLV. Consequently, a centralized controller would be able to reconstruct and analyze the state of the particular multicast distribution tree based on HTS packets collected from egress nodes.

5. IANA Considerations

TBD

6. Security Considerations

Nodes that practice HTS method are presumed to share a trust model that depends on the existence of a trusted relationship among nodes. This is necessary as these nodes are expected to correctly modify the specific content of the data in the follow-up packet, and the degree to which HTS measurement is useful for network operation depends on this ability. In practice, this means either confidentiality or integrity protection cannot cover those portions of messages that contain the network state data. Though there are methods that make it possible in theory to provide either or both such protections and still allow for intermediate nodes to make detectable yet authenticated modifications, such methods do not seem practical at present, particularly for protocols that used to measure latency and/or jitter.

The ability to potentially authenticate and/or encrypt the network state data for scenarios both with and without the participation of intermediate nodes that participate in HTS measurement is left for further study.

While it is possible for a supposed compromised node to intercept and modify the network state information in the follow-up packet, this is an issue that exists for nodes in general - for all data that to be carried over the particular networking technology - and is therefore the basis for an additional presumed trust model associated with an existing network.

7. Acknowledgments

Authors express their gratitude and appreciation to Joel Halpern for the most helpful and insightful discussion on the applicability of HTS in a Service Function Chaining domain.

8. References

8.1. 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>.

8.2. Informative References

[I-D.ietf-ippm-ioam-data] Brockners, F., Bhandari, S., Pignataro, C., Gredler, H., Leddy, J., Youell, S., Mizrahi, T., Mozes, D., Lapukhov, P., Chang, R., daniel.bernier@bell.ca, d., and J. Lemon, "Data Fields for In-situ OAM", draft-ietf-ippm-ioam-data-07 (work in progress), September 2019.

[P4.INT] "In-band Network Telemetry (INT)", P4.org Specification, October 2017.

[RFC7799] Morton, A., "Active and Passive Metrics and Methods (with Hybrid Types In-Between)", RFC 7799, DOI 10.17487/RFC7799, May 2016, <https://www.rfc-editor.org/info/rfc7799>.

[RFC8169] Mirsky, G., Ruffini, S., Gray, E., Drake, J., Bryant, S., and A. Vainshtein, "Residence Time Measurement in MPLS Networks", RFC 8169, DOI 10.17487/RFC8169, May 2017, <https://www.rfc-editor.org/info/rfc8169>.

[RFC8296] Wijnands, IJ., Ed., Rosen, E., Ed., Dolganow, A., Tantsura, J., Aldrin, S., and I. Meilik, "Encapsulation for Bit Index Explicit Replication (BIER) in MPLS and Non-MPLS Networks", RFC 8296, DOI 10.17487/RFC8296, January 2018, <https://www.rfc-editor.org/info/rfc8296>.

Authors' Addresses

Greg Mirsky
ZTE Corp.

Email: gregimirsky@gmail.com

Wang Lingqiang
ZTE Corporation
No 19 ,East Huayuan Road
Beijing 100191
P.R.China

Phone: +86 10 82963945
Email: wang.lingqiang@zte.com.cn

Guo Zhui
ZTE Corporation
No 19 ,East Huayuan Road
Beijing 100191
P.R.China

Phone: +86 10 82963945
Email: guo.zhui@zte.com.cn