IPPM Working Group
Internet-Draft
Intended status: Standards Track
Expires: 26 February 2026

G. Mirsky
Ericsson
W. Lingqiang
G. Zhui
ZTE Corporation
H. Song
Futurewei Technologies
P. Thubert
Independent
25 August 2025

Hybrid Two-Step Performance Measurement Method

draft-ietf-ippm-hybrid-two-step-05

Abstract

The development and advancements in network operation automation have brought new measurement methodology requirements. 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. That task can be solved using on-path telemetry, also called hybrid measurement. An on-path telemetry method allows the collection of essential information that reflects the operational state and network performance experienced by the packet. This document introduces a method complementary to on-path telemetry that causes the generation of Network ~~t~~Telemetry information. This method, referred to as Hybrid Two-Step (HTS), separates the act of measuring and/or calculating the performance metric from collecting and transporting network state. The HTS packet traverses the same set of nodes and links as the trigger packet, thus simplifying the correlation of informational elements originating on nodes traversed by the trigger packet.

Commented [TG1]: Please define and refer the term „Network Telemetry“ in the terminology section and use throughout the document.

I suggest to refer to https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology for „Network Telemetry“ or alternatively to https://datatracker.ietf.org/doc/html/rfc9232.

This Internet-Draft will expire on 26 February 2026.

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~~exported to big data systems for processing and storing. The process of ~~producing~~exporting, 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. [RFC7799] describes a combination of elements of passive and active measurement as a hybrid measurement.

Several technical methods have been proposed to enable the collection of network state information instantaneous to the packet processing, among them [P4.INT] and [RFC9197]. The instantaneous, i.e., in the data packet itself, collection of telemetry information simplifies the process of attribution of telemetry information to the particular monitored flow. On the other hand, this collection method impacts the data packets, potentially changing their treatment by the network~~ing~~ nodes. Also, the amount of information the instantaneous method collects might be incomplete because of the limited space it can be allotted. Other proposals defined methods to collect telemetry information in a separate packet from each network node traversed by the monitored data flow. Examples of this approach to collecting telemetry information are [RFC9326] and [I-D.song-ippm-postcard-based-telemetry]. These methods allow data collection from any arbitrary path and avoid directly impacting data packets. On the other hand, the correlation of data and the monitored flow requires that each packet with telemetry information also includes characteristic information about the monitored flow.

Commented [TG2]: I suggest to use the term „big data“ as a general term to describe data processing and storing at high scale.

I suggest to use the term „export“ and „collect“ in context of „Network Telemetry“. https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology. There are different Network Telemetry protocols at IETF which use different terms such as „publish“ in context of YANG.

Commented [TG3]: The meaning of „network state“ is also defined in https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology-23#section-3.2. I suggest to define and refer in terminology section.

Worth mentioning could be that the „network relevant state“ is computed at Big Data.

Commented [TG4]: The term „node“ has several meanings. In terms of YANG it refers to a schema element. I suggest to use the term „network node“ throughout the document in context of network but not in context of the HTS processing.

Commented [TG5]: Is also defined in https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology-23#section-3.2. I would appreciate to define and refer in the terminology section.

This document introduces Hybrid Two-Step (HTS) as a new method of telemetry data collection that improves accuracy of a measurement by separating the act of measuring or calculating the performance metric from the collecting and transporting this information while minimizing the overhead of the generated load in a network. HTS method extends the two-step mode of Residence Time Measurement (RTM) defined in [RFC8169] to on-path network state collection and transport. HTS allows the collection of telemetry information from any arbitrary path. HTS instruments data packets of the monitored flow or specially constructed test packets that are already equipped with a shim of on-path telemetry protocol to use as an HTS trigger packet, making the process of attribution of telemetry to the data flow simple.

2. Conventions used in this document

2.1. Acronyms and Terminology

RTM Residence Time Measurement

ECMP Equal Cost Multipath

MTU Maximum Transmission Unit

HTS Hybrid Two-Step

HMAC Hashed Message Authentication Code

TLV Type-Length-Value

RTT Round-Trip Time

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

Commented [TG6]: Please list RFC 7799 and https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology-23 defined terms used in this document and refer to those document for their terminology specification. draft-ietf-nmop-terminology is at RFC editor state.

Please define also the HTS node types such as, HTS Ingress Node, HTS Intermediate Node, HTS Egress node with a small sentence to describe their meaning and reference to the document section where their function is described in detail.

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 when 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 recorded can differ from the ideal case by any fixed amount. A correction can be applied to compute the same time difference ~~taking into account~~considering 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~~challenging 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. Recording the departure of a packet time in the same packet may be decremental to the accuracy of the measurement because the departure time includes the variable time component (such as that associated with buffering and queuing of the packet). A similar problem may lower the 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 egress interface state. 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 on the path, 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 the client layer or risk missing network state data for the part, most probably the latter part, of the path.

Performance measurement methods that instrument data flows inherently collect one-way performance metrics at the egress of the measurement domain. In some networks, for example, wireless that are in the scope of [RFC9450], it is beneficial to collect the telemetry, including the calculated performance metrics, that reflects conditions experienced by the monitored flow at a network node other than the egress network node. For example, a head-end can optimize path selection based on the compounded information that reflects network conditions and resource utilization. This mode is referred to as the upstream collection and the other - downstream collection to differentiate between two modes of telemetry collection.

4. Theory of Operation

The HTS method consists of two phases:

* Performing a measurement and/or obtaining network state information on a network node. HTS Trigger is a data or test packet instrumented to trigger the collection of telemetry information on a network node.
* Collecting and transporting the measurement and/or the telemetry information. HTS Follow-up is a packet constructed to transport telemetry information that includes operational state and performance measurements originated on the nodes along the path traversed by the HTS Trigger.

4.1. HTS Packets

4.1.1. HTS Trigger in In-Situ OAM

```
        0                   1                   2                   3
        0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |        Namespace-ID           |     Flags     |Extension-Flags|
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |               IOAM-Trace-Type                 |   Reserved    |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                         Flow ID (Optional)                    |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                     Sequence Number  (Optional)               |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Hybrid Two-Step Trace IOAM Header

An HTS Trigger may be carried in a data packet or a specially constructed test packet. For example, an HTS Trigger could be a packet that has IOAM Option-Type set to the "IOAM Hybrid Two-Step Option-Type" value (TBA1) allocated by IANA (see Section 7.1). The HTS Trigger includes HTS IOAM Header (shown in Figure 1) consists of:

* IOAM Namespace-ID - as defined in Section 5.3 [RFC9197];
* Flags - as defined in Section 3.2 [RFC9326];
* Extension-Flags - as defined in Section 3.2 [RFC9326];
* IOAM-Trace-Type - as defined in Section 5.4 [RFC9197];
* optional Flow ID - as defined in Section 3.2 [RFC9326];
* optional Sequence Number - as defined in Section 3.2 [RFC9326].

4.1.2. HTS Trigger in the Alternate Marking Method

A packet in the flow to which the Alternate-Marking method, defined in [RFC9341] and [RFC9342], is applied can be used as an HTS Trigger. The nature of the HTS Trigger is a transport network layer-specific, and its description is outside the scope of this document. The packet that includes the HTS Trigger in this document is also referred to as the trigger packet.

4.1.3. HTS Follow-up Packet

The HTS method uses the HTS Follow-up packet, 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. In some use cases, e.g., when HTS is used to collect the telemetry, including performance metrics, calculated based on a series of measurements, an HTS follow-up packet can be originated without using the HTS Trigger. The follow-up packet contains characteristic information sufficient for participating HTS nodes to associate it with the monitored data flow. The characteristic information can be obtained using the information of the trigger packet or constructed by a node that originates the follow-up packet. As the follow-up packet is expected to traverse the same sequence of nodes, one element of the characteristic information is the information that determines the path in the data plane. For example, in a segment routing domain [RFC8402], a list of segment identifiers of the trigger packet is applied to the follow-up packet. And in the case of the service function chain based on the Network Service Header [RFC8300], the Base Header and Service Path Header of the trigger packet will be applied to the follow-up packet. Also, when HTS is used to collect the telemetry information in an IOAM domain, the IOAM trace option header [RFC9197] of the trigger packet is applied in the follow-up packet. The follow-up packet also uses the same network information used to load-balance flows in equal-cost multipath (ECMP) as the trigger packet, e.g., IPv6 Flow Label [RFC6437] or an entropy label [RFC6790]. The exact composition of the characteristic information is specific for each transport network, and its definition is outside the scope of this document. 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.

The following sections describe the operation of HTS nodes in the downstream mode of collecting the telemetry information. In the upstream mode, the behavior of HTS nodes, in general, identical with the exception that the HTS Trigger packet does not precede the HTS Follow-up packet.

Commented [TG7]: I understand the debate on https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-oam-characterization. Exactly in this context I would cheer to see https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-oam-characterization-10#section-3.2 and https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-oam-characterization-10#section-3.3 related context.

Commented [TG8]: See remark on „node“ term. For the reader, the implementor of the protocol, it would be much easier and straight forward if ambiguity around „node“ could be resolved by introducing HTS node types in terminology section.

4.2. Operation of the HTS Ingress Node

A node that originates the HTS Trigger is referred to as the 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 2 displays an example of the follow-up packet format.

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                      Transport Network                        ~
    |                        Encapsulation                          |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |Ver|HTS Shim L |     Flags     |Sequence Number|   Reserved    |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                         HTS Max Length                        |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    ~                     Telemetry Data Profile                    ~
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |                                                               |
    ~                      Telemetry Data TLVs                      ~
    |                                                               |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: 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 octets. The minimal value of the field is eight octets.

```
      0
      0 1 2 3 4 5 6 7
     +-+-+-+-+-+-+-+-+
     |F|  Reserved   |
     +-+-+-+-+-+-+-+-+
```

Figure 3: Flags Field Format

Flags is eight-bits long. The format of the Flags field displayed in Figure 3.

- 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 one octet-long field. The zero-based value of the field reflects the place of the HTS follow-up packet in the sequence of the HTS follow-up packets that originated in response to the same HTS trigger. The ingress node MUST set the value of the field to zero.

Reserved is one octet-long field. It MUST be zeroed on transmission and ignored on receipt.

HTS Max Length is four octet-long field. The value of the HTS Max Length field indicates the maximum length of the HTS Follow-up packet in octets. An operator MUST be able to configure the HTS Max Length field's value. The value SHOULD be set equal to the path MTU.

Telemetry Data Profile is the optional variable-length field of bit-size flags. Each flag indicates the requested type of telemetry data to be collected at each HTS node. The increment of the field is four bytes with a minimum length of zero. For example, IOAM-Trace-Type information defined in [RFC9197], Sequence Number and/or Flow ID (Figure 1) can be used in the Telemetry Data Profile field.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Reserved   |           Length              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                            Value                              ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Telemetry Data TLV Format

Telemetry Data TLV is a variable-length field. Multiple TLVs MAY be placed in an HTS packet. Additional TLVs may be enclosed within a given TLV, subject to the semantics of the (outer) TLV in question. Figure 4 presents the format of a Telemetry Data TLV, where fields are defined as the following:

- Type - a one-octet-long field that characterizes the interpretation of the Value field.
- Reserved - one-octet-long field.
- Length - two-octet-long field equal to the length of the Value field in octets.
- Value - a variable-length field. The value of the Type field determines its interpretation and encoding. IOAM data fields, defined in [RFC9197], MAY be carried in the Value field.

All multibyte fields defined in this specification are in network byte order.

4.3. Operation of the HTS Intermediate Node

Commented [TG9]: Please define in terminology section.

Upon receiving the trigger packet, the HTS intermediate node MUST:

* copy the transport information;
* start the HTS Follow-up Timer for the obtained flow;
* transmit the trigger packet.

Upon receiving the follow-up packet, the HTS intermediate node MUST:

1. verify that the matching transport information exists and the Full flag is cleared, then stop the associated HTS Follow-up Timer;
2. otherwise, transmit the received packet. Proceed to Step 8;
3. collect telemetry data requested in the Telemetry Data Profile field or defined by the local HTS policy;
4. if adding the collected telemetry would not exceed HTS Max Length field's value, then append data as a new Telemetry Data TLV and transmit the follow-up packet. Proceed to Step 8;
5. otherwise, set the value of the Full flag to one, copy the transport information from the received follow-up packet and transmit it accordingly;

6. originate the new follow-up packet using the transport information copied from the received follow-up packet. 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;
7. copy collected telemetry data into the first Telemetry Data TLV's Value field and then transmit the packet;
8. processing completed.

If the HTS Follow-up Timer expires, the intermediate node MUST:

* originate the follow-up packet using transport information associated with the expired timer;
* initialize the HTS shim by setting the Version field's value 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 TLV's Value field and transmit the packet.

If the intermediate node receives a "late" follow-up packet, i.e., a packet to which the node has no associated HTS Follow-up timer, the node MUST forward the "late" packet.

4.4. Operation of the HTS Egress Node

Commented [TG10]: Please define in terminology section.

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 for each of Telemetry Data TLVs MUST:

* if HTS is used in the authenticated mode, verify the authentication of the Telemetry Data TLV using the Authentication sub-TLV (see Section 6);
* copy telemetry information from the Value field;
* 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.

5. Operational Considerations

Correctly attributing information originated by the particular trigger packet to the proper HTS Follow-up packet is essential for the HTS protocol. That can be achieved using characteristic information that uniquely identifies the trigger packet within a given HTS domain. For example, a combination of the flow identifier and packet's sequence number within that flow, as Flow ID and Sequence Number in IOAM Direct Export [RFC9326], can be used to correlate between stored telemetry information and the appropriate HTS Follow-up packet. In case the trigger packet doesn't include data that distinguish it from other trigger packets in the HTS domain, then for the particular flow, there MUST be no more than one HTS Trigger, values of HTS timers bounded by the rate of the trigger generation for that flow. In practice, the minimal interval between HTS Trigger packets SHOULD be selected from the range determined by the round-trip time (RTT) between HTS Ingress and HTS Egress nodes as [RTT/2, RTT].

Commented [TG11]: Please define in terminology section.

5.1. 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 invaluable 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 the HTS packet as defined below:

* the first transmitted multicast packet MUST be followed by the received corresponding HTS packet as described in Section 4.3;
* 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 an intermediate HTS node when the HTS 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 reconstruct and analyze the state of the particular multicast distribution tree based on HTS packets collected from egress nodes.

6. Authentication in HTS

Telemetry information may be used to drive network operation, closing the control loop for self-driving, self-healing networks. ~~Thus~~Thus, it is critical to provide a mechanism to protect the telemetry information collected using the HTS method. This document defines an optional authentication of a Telemetry Data TLV that protects the collected information's integrity.

The format of the Authentication sub-TLV is displayed in Figure 5.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Authentic. Type|   HMAC Type   |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                            Digest                             |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: HMAC sub-TLV

where fields are defined as follows:

* Authentication Type - is a one-octet-long field, value 1 is allocated by IANA Section 7.2.
* Length - two-octet-long field, set equal to the length of the Digest field in octets.
* HMAC Type - is a one-octet-long field that identifies the type of the HMAC and the length of the digest and the length of the digest according to the HTS HMAC Type sub-registry (see Section 7.4).
* Digest - is a variable-length field that carries HMAC digest of the text that includes the encompassing TLV.

This specification defines the use of HMAC-SHA-256 truncated to 128 bits ([RFC4868]) in HTS. Future specifications may define the use in HTS of more advanced cryptographic algorithms or the use of digest of a different length. HMAC is calculated as defined in [RFC2104] over text as the concatenation of the Sequence Number field of the follow-up packet (see Figure 2) and the preceding data collected in the Telemetry Data TLV. The digest then MUST be truncated to 128 bits and written into the Digest field. Distribution and management of shared keys are outside the scope of this document. In the HTS authenticated mode, the Authentication sub-TLV MUST be present in each Telemetry Data TLV. HMAC MUST be verified before using any data in the included Telemetry Data TLV. If HMAC verification fails, the system MUST stop processing corresponding Telemetry Data TLV and notify an operator. Specification of the notification mechanism is outside the scope of this document.

7. IANA Considerations

7.1. IOAM Option-Type for HTS

The IOAM Option-Type registry is requested in [RFC9197]. IANA is requested to allocate a new code point as listed in Table 1.

Value	Name	Description	Reference
TBA1	IOAM Hybrid Two-Step (HTS) Option-Type	HTS Exporting	This document

Table 1: IOAM Option-Type for HTS

7.2. HTS TLV Registry

IANA is requested to create "Hybrid Two-Step" registry group. IANA is requested to create the HTS TLV Type registry in "Hybrid Two-Step" registry group. All code points in the range 1 through 175 in this registry shall be allocated according to the "IETF Review" procedure specified in [RFC8126]. Code points in the range 176 through 239 in this registry shall be allocated according to the "First Come First Served" procedure specified in [RFC8126]. The remaining code points are allocated according to Table 2:

Value	Description	Reference
0	Reserved	This document
1- 175	Unassigned	This document
176 - 239	Unassigned	This document
240 - 251	Experimental	This document
252 - 254	Private Use	This document
255	Reserved	This document

Table 2: HTS TLV Type Registry

7.3. HTS Sub-TLV Type Sub-registry

IANA is requested to create the HTS sub-TLV Type sub-registry as part of the HTS TLV Type registry. All code points in the range 1 through 175 in this registry shall be allocated according to the "IETF Review" procedure specified in [RFC8126]. Code points in the range 176 through 239 in this registry shall be allocated according to the "First Come First Served" procedure specified in [RFC8126]. The remaining code points are allocated according to Table 3:

Value	Description	TLV Used	Reference
0	Reserved	None	This document
1	HMAC	Any	This document
2 - 175	Unassigned		This document
176 - 239	Unassigned		This document
240 - 251	Experimental		This document
252 - 254	Private Use		This document
255	Reserved	None	This document

Table 3: HTS Sub-TLV Type Sub-registry

7.4. HMAC Type Sub-registry

IANA is requested to create the HMAC Type sub-registry as part of the HTS TLV Type registry. All code points in the range 1 through 127 in this registry shall be allocated according to the "IETF Review" procedure specified in [RFC8126]. Code points in the range 128 through 239 in this registry shall be allocated according to the "First Come First Served" procedure specified in [RFC8126]. The remaining code points are allocated according to Table 4:

Value	Description	Reference
0	Reserved	This document
1	HMAC-SHA-256 16 octets long	This document
2 - 127	Unassigned	This document
128 - 239	Unassigned	This document
240 - 249	Experimental	This document
250 - 254	Private Use	This document
255	Reserved	This document

Table 4: HMAC Type Sub-registry

8. Security Considerations

Nodes that practice the 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.

This document defines the use of authentication (Section 6) to protect the integrity of the telemetry information collected using the HTS method. Privacy protection can be achieved by, for example, sharing the IPsec tunnel with a data flow that generates information that is collected using HTS.

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.

9. 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. Also, the authors thank Bjørn Ivar Teigen for the discussion about ensuring proper correlation between generated telemetry information and an HTS Follow-up packet. And a special thank you to Xiao Min for thorough review and thoughtful suggestions that helped in improving the document.

10. References

10.1. Normative References

[RFC2104] Krawczyk, H., Bellare, M., and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, DOI 10.17487/RFC2104, February 1997, <https://www.rfc-editor.org/info/rfc2104>.

[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>.

[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>.

[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>.

10.2. Informative References

[I-D.song-ippm-postcard-based-telemetry] Song, H., Mirsky, G., Zhou, T., Li, Z., Graf, T., Mishra, G. S., Shin, J., and K. Lee, "On-Path Telemetry using Packet Marking to Trigger Dedicated OAM Packets", Work in Progress, Internet-Draft, draft-song-ippm-postcard-based-telemetry-16, 2 June 2023, <https://datatracker.ietf.org/doc/html/draft-song-ippm-postcard-based-telemetry-16>.

[P4.INT] "In-band Network Telemetry (INT)", P4.org Specification, November 2020.

[RFC4868] Kelly, S. and S. Frankel, "Using HMAC-SHA-256, HMAC-SHA-384, and HMAC-SHA-512 with IPsec", RFC 4868, DOI 10.17487/RFC4868, May 2007, <https://www.rfc-editor.org/info/rfc4868>.

[RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, DOI 10.17487/RFC6437, November 2011, <https://www.rfc-editor.org/info/rfc6437>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[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>.

[RFC9342] Fioccola, G., Ed., Cociglio, M., Sapio, A., Sisto, R., and T. Zhou, "Clustered Alternate-Marking Method", RFC 9342, DOI 10.17487/RFC9342, December 2022, <https://www.rfc-editor.org/info/rfc9342>.

[RFC9450] Bernardos, CJ., Ed., Papadopoulos, G., Thubert, P., and F. Theoleyre, "Reliable and Available Wireless (RAW) Use Cases", RFC 9450, DOI 10.17487/RFC9450, August 2023, <https://www.rfc-editor.org/info/rfc9450>.

Authors' Addresses

Greg Mirsky
Ericsson
Email: gregimirsky@gmail.com

Wang Lingqiang
ZTE Corporation
No 19 ,East Huayuan Road
Beijing
100191
China
Phone: +86 10 82963945
Email: wang.lingqiang@zte.com.cn

Formatted: English (United States)

Guo Zhui
ZTE Corporation
No 19 ,East Huayuan Road
Beijing
100191
China
Phone: +86 10 82963945
Email: guo.zhui@zte.com.cn
Haoyu Song
Futurewei Technologies
2330 Central Expressway
Santa Clara,
United States of America
Email: hsong@futurewei.com
Pascal Thubert
Independent
06330 Roquefort-les-Pins
France
Email: pascal.thubert@gmail.com

Hybrid Two-Step Performance Measurement Method
draft-ietf-ippm-hybrid-two-step-05

Abstract

The development and advancements in network operation automation have brought new measurement methodology requirements. 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. That task can be solved using on-path telemetry, also called hybrid measurement. An on-path telemetry method allows the collection of essential information that reflects the operational state and network performance experienced by the packet. This document introduces a method complementary to on-path telemetry that causes the generation of Network ~~t~~Telemetry information. This method, referred to as Hybrid Two-Step (HTS), separates the act of measuring and/or calculating the performance metric from collecting and transporting network state. The HTS packet traverses the same set of nodes and links as the trigger packet, thus simplifying the correlation of informational elements originating on nodes traversed by the trigger packet.

Commented [TG12]: Please define and refer the term „Network Telemetry“ in the terminology section and use throughout the document.

I suggest to refer to https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology for „Network Telemetry“ or alternatively to https://datatracker.ietf.org/doc/html/rfc9232.

Status of This Memo

This Internet-Draft will expire on 26 February 2026.

Copyright Notice

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~~exported to big data systems for processing and storing. The process of ~~producing~~exporting, 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. [RFC7799] describes a combination of elements of passive and active measurement as a hybrid measurement.

Several technical methods have been proposed to enable the collection of network state information instantaneous to the packet processing, among them [P4.INT] and [RFC9197]. The instantaneous, i.e., in the data packet itself, collection of telemetry information simplifies the process of attribution of telemetry information to the particular monitored flow. On the other hand, this collection method impacts the data packets, potentially changing their treatment by the network~~ing~~ nodes. Also, the amount of information the instantaneous method collects might be incomplete because of the limited space it can be allotted. Other proposals defined methods to collect telemetry information in a separate packet from each network node traversed by the monitored data flow. Examples of this approach to collecting telemetry information are [RFC9326] and [I-D.song-ippm-postcard-based-telemetry]. These methods allow data collection from any arbitrary path and avoid directly impacting data packets. On the other hand, the correlation of data and the monitored flow requires that each packet with telemetry information also includes characteristic information about the monitored flow.

Commented [TG13]: I suggest to use the term „big data“ as a general term to describe data processing and storing at high scale.

I suggest to use the term „export“ and „collect“ in context of „Network Telemetry“. https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology. There are different Network Telemetry protocols at IETF which use different terms such as „publish“ in context of YANG.

Commented [TG14]: The meaning of „network state“ is also defined in https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology-23#section-3.2. I suggest to define and refer in terminology section.

Worth mentioning could be that the „network relevant state“ is computed at Big Data.

Commented [TG15]: The term „node“ has several meanings. In terms of YANG it refers to a schema element. I suggest to use the term „network node“ throughout the document in context of network but not in context of the HTS processing.

Commented [TG16]: Is also defined in https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology-23#section-3.2. I would appreciate to define and refer in the terminology section.

This document introduces Hybrid Two-Step (HTS) as a new method of telemetry data collection that improves accuracy of a measurement by separating the act of measuring or calculating the performance metric from the collecting and transporting this information while minimizing the overhead of the generated load in a network. HTS method extends the two-step mode of Residence Time Measurement (RTM) defined in [RFC8169] to on-path network state collection and transport. HTS allows the collection of telemetry information from any arbitrary path. HTS instruments data packets of the monitored flow or specially constructed test packets that are already equipped with a shim of on-path telemetry protocol to use as an HTS trigger packet, making the process of attribution of telemetry to the data flow simple.

2. Conventions used in this document

2.1. Acronyms and Terminology

RTM Residence Time Measurement

ECMP Equal Cost Multipath

MTU Maximum Transmission Unit

HTS Hybrid Two-Step

HMAC Hashed Message Authentication Code

TLV Type-Length-Value

RTT Round-Trip Time

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

Commented [TG17]: Please list RFC 7799 and https://datatracker.ietf.org/doc/html/draft-ietf-nmop-terminology-23 defined terms used in this document and refer to those document for their terminology specification. draft-ietf-nmop-terminology is at RFC editor state.

Please define also the HTS node types such as, HTS Ingress Node, HTS Intermediate Node, HTS Egress node with a small sentence to describe their meaning and reference to the document section where their function is described in detail.

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 when 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 recorded can differ from the ideal case by any fixed amount. A correction can be applied to compute the same time difference ~~taking into account~~considering 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~~challenging 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. Recording the departure of a packet time in the same packet may be decremental to the accuracy of the measurement because the departure time includes the variable time component (such as that associated with buffering and queuing of the packet). A similar problem may lower the 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 egress interface state. 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 on the path, 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 the client layer or risk missing network state data for the part, most probably the latter part, of the path.

Performance measurement methods that instrument data flows inherently collect one-way performance metrics at the egress of the measurement domain. In some networks, for example, wireless that are in the scope of [RFC9450], it is beneficial to collect the telemetry, including the calculated performance metrics, that reflects conditions experienced by the monitored flow at a network node other than the egress network node. For example, a head-end can optimize path selection based on the compounded information that reflects network conditions and resource utilization. This mode is referred to as the upstream collection and the other - downstream collection to differentiate between two modes of telemetry collection.

4. Theory of Operation

The HTS method consists of two phases:

* Performing a measurement and/or obtaining network state information on a network node. HTS Trigger is a data or test packet instrumented to trigger the collection of telemetry information on a network node.

* Collecting and transporting the measurement and/or the telemetry information. HTS Follow-up is a packet constructed to transport telemetry information that includes operational state and performance measurements originated on the nodes along the path traversed by the HTS Trigger.

4.1. HTS Packets

4.1.1. HTS Trigger in In-Situ OAM

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Namespace-ID           |     Flags     |Extension-Flags|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|               IOAM-Trace-Type                 |   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Flow ID (Optional)                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Sequence Number  (Optional)               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1: Hybrid Two-Step Trace IOAM Header

An HTS Trigger may be carried in a data packet or a specially constructed test packet. For example, an HTS Trigger could be a packet that has IOAM Option-Type set to the "IOAM Hybrid Two-Step Option-Type" value (TBA1) allocated by IANA (see Section 7.1). The HTS Trigger includes HTS IOAM Header (shown in Figure 1) consists of:

* IOAM Namespace-ID - as defined in Section 5.3 [RFC9197];
* Flags - as defined in Section 3.2 [RFC9326];
* Extension-Flags - as defined in Section 3.2 [RFC9326];
* IOAM-Trace-Type - as defined in Section 5.4 [RFC9197];
* optional Flow ID - as defined in Section 3.2 [RFC9326];
* optional Sequence Number - as defined in Section 3.2 [RFC9326].

4.1.2. HTS Trigger in the Alternate Marking Method

A packet in the flow to which the Alternate-Marking method, defined in [RFC9341] and [RFC9342], is applied can be used as an HTS Trigger. The nature of the HTS Trigger is a transport network layer-specific, and its description is outside the scope of this document. The packet that includes the HTS Trigger in this document is also referred to as the trigger packet.

4.1.3. HTS Follow-up Packet

The HTS method uses the HTS Follow-up packet, 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. In some use cases, e.g., when HTS is used to collect the telemetry, including performance metrics, calculated based on a series of measurements, an HTS follow-up packet can be originated without using the HTS Trigger. The follow-up packet contains characteristic information sufficient for participating HTS nodes to associate it with the monitored data flow. The characteristic information can be obtained using the information of the trigger packet or constructed by a node that originates the follow-up packet. As the follow-up packet is expected to traverse the same sequence of nodes, one element of the characteristic information is the information that determines the path in the data plane. For example, in a segment routing domain [RFC8402], a list of segment identifiers of the trigger packet is applied to the follow-up packet. And in the case of the service function chain based on the Network Service Header [RFC8300], the Base Header and Service Path Header of the trigger packet will be applied to the follow-up packet. Also, when HTS is used to collect the telemetry information in an IOAM domain, the IOAM trace option header [RFC9197] of the trigger packet is applied in the follow-up packet. The follow-up packet also uses the same network information used to load-balance flows in equal-cost multipath (ECMP) as the trigger packet, e.g., IPv6 Flow Label [RFC6437] or an entropy label [RFC6790]. The exact composition of the characteristic information is specific for each transport network, and its definition is outside the scope of this document.

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.

The following sections describe the operation of HTS nodes in the downstream mode of collecting the telemetry information. In the upstream mode, the behavior of HTS nodes, in general, identical with the exception that the HTS Trigger packet does not precede the HTS Follow-up packet.

Commented [TG18]: I understand the debate on https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-oam-characterization. Exactly in this context I would cheer to see https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-oam-characterization-10#section-3.2 and https://datatracker.ietf.org/doc/html/draft-ietf-opsawg-oam-characterization-10#section-3.3 related context.

Commented [TG19]: See remark on „node“ term. For the reader, the implementor of the protocol, it would be much easier and straight forward if ambiguity around „node“ could be resolved by introducing HTS node types in terminology section.

4.2. Operation of the HTS Ingress Node

A node that originates the HTS Trigger is referred to as the 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 2 displays an example of the follow-up packet format.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                      Transport Network                        ~
|                        Encapsulation                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Ver|HTS Shim L |     Flags     |Sequence Number|   Reserved    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        HTS Max Length                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                     Telemetry Data Profile                    ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                      Telemetry Data TLVs                      ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2: 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 octets. The minimal value of the field is eight octets.

```
 0
 0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|F|  Reserved   |
+-+-+-+-+-+-+-+-+
```

Figure 3: Flags Field Format

Flags is eight-bits long. The format of the Flags field displayed in Figure 3.

- 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 one octet-long field. The zero-based value of the field reflects the place of the HTS follow-up packet in the sequence of the HTS follow-up packets that originated in response to the same HTS trigger. The ingress node MUST set the value of the field to zero.

Reserved is one octet-long field. It MUST be zeroed on transmission and ignored on receipt.

HTS Max Length is four octet-long field. The value of the HTS Max Length field indicates the maximum length of the HTS Follow-up packet in octets. An operator MUST be able to configure the HTS Max Length field's value. The value SHOULD be set equal to the path MTU.

Telemetry Data Profile is the optional variable-length field of bit-size flags. Each flag indicates the requested type of telemetry data to be collected at each HTS node. The increment of the field is four bytes with a minimum length of zero. For example, IOAM-Trace-Type information defined in [RFC9197], Sequence Number and/or Flow ID (Figure 1) can be used in the Telemetry Data Profile field.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|      Type     |    Reserved   |           Length              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
~                            Value                              ~
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 4: Telemetry Data TLV Format

Telemetry Data TLV is a variable-length field. Multiple TLVs MAY be placed in an HTS packet. Additional TLVs may be enclosed within a given TLV, subject to the semantics of the (outer) TLV in question. Figure 4 presents the format of a Telemetry Data TLV, where fields are defined as the following:

- Type - a one-octet-long field that characterizes the interpretation of the Value field.

- Reserved - one-octet-long field.

- Length - two-octet-long field equal to the length of the Value field in octets.

- Value - a variable-length field. The value of the Type field determines its interpretation and encoding. IOAM data fields, defined in [RFC9197], MAY be carried in the Value field.

All multibyte fields defined in this specification are in network byte order.

4.3. Operation of the HTS Intermediate Node

Commented [TG20]: Please define in terminology section.

Upon receiving the trigger packet, the HTS intermediate node MUST:

* copy the transport information;

* start the HTS Follow-up Timer for the obtained flow;

* transmit the trigger packet.

Upon receiving the follow-up packet, the HTS intermediate node MUST:

1. verify that the matching transport information exists and the Full flag is cleared, then stop the associated HTS Follow-up Timer;

2. otherwise, transmit the received packet. Proceed to Step 8;

3. collect telemetry data requested in the Telemetry Data Profile field or defined by the local HTS policy;

4. if adding the collected telemetry would not exceed HTS Max Length field's value, then append data as a new Telemetry Data TLV and transmit the follow-up packet. Proceed to Step 8;

5. otherwise, set the value of the Full flag to one, copy the transport information from the received follow-up packet and transmit it accordingly;

6. originate the new follow-up packet using the transport information copied from the received follow-up packet. 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;

7. copy collected telemetry data into the first Telemetry Data TLV's Value field and then transmit the packet;

8. processing completed.

If the HTS Follow-up Timer expires, the intermediate node MUST:

* originate the follow-up packet using transport information associated with the expired timer;

* initialize the HTS shim by setting the Version field's value 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 TLV's Value field and transmit the packet.

If the intermediate node receives a "late" follow-up packet, i.e., a packet to which the node has no associated HTS Follow-up timer, the node MUST forward the "late" packet.

4.4. Operation of the HTS Egress Node

Commented [TG21]: Please define in terminology section.

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 for each of Telemetry Data TLVs MUST:

* if HTS is used in the authenticated mode, verify the authentication of the Telemetry Data TLV using the Authentication sub-TLV (see Section 6);

* copy telemetry information from the Value field;

* 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.

5. Operational Considerations

Correctly attributing information originated by the particular trigger packet to the proper HTS Follow-up packet is essential for the HTS protocol. That can be achieved using characteristic information that uniquely identifies the trigger packet within a given HTS domain. For example, a combination of the flow identifier and packet's sequence number within that flow, as Flow ID and Sequence Number in IOAM Direct Export [RFC9326], can be used to correlate between stored telemetry information and the appropriate HTS Follow-up packet. In case the trigger packet doesn't include data that distinguish it from other trigger packets in the HTS domain, then for the particular flow, there MUST be no more than one HTS Trigger, values of HTS timers bounded by the rate of the trigger generation for that flow. In practice, the minimal interval between HTS Trigger packets SHOULD be selected from the range determined by the round-trip time (RTT) between HTS Ingress and HTS Egress nodes as [RTT/2, RTT].

Commented [TG22]: Please define in terminology section.

5.1. 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 invaluable 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 the HTS packet as defined below:

* the first transmitted multicast packet MUST be followed by the received corresponding HTS packet as described in Section 4.3;

* 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 an intermediate HTS node when the HTS 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 reconstruct and analyze the state of the particular multicast distribution tree based on HTS packets collected from egress nodes.

6. Authentication in HTS

Telemetry information may be used to drive network operation, closing the control loop for self-driving, self-healing networks. ~~Thus~~Thus, it is critical to provide a mechanism to protect the telemetry information collected using the HTS method. This document defines an optional authentication of a Telemetry Data TLV that protects the collected information's integrity.

The format of the Authentication sub-TLV is displayed in Figure 5.

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Authentic. Type|   HMAC Type   |             Length            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
|                            Digest                             |
|                                                               |
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5: HMAC sub-TLV

where fields are defined as follows:

* Authentication Type - is a one-octet-long field, value 1 is allocated by IANA Section 7.2.

* Length - two-octet-long field, set equal to the length of the Digest field in octets.

* HMAC Type - is a one-octet-long field that identifies the type of the HMAC and the length of the digest and the length of the digest according to the HTS HMAC Type sub-registry (see Section 7.4).

* Digest - is a variable-length field that carries HMAC digest of the text that includes the encompassing TLV.

This specification defines the use of HMAC-SHA-256 truncated to 128 bits ([RFC4868]) in HTS. Future specifications may define the use in HTS of more advanced cryptographic algorithms or the use of digest of a different length. HMAC is calculated as defined in [RFC2104] over text as the concatenation of the Sequence Number field of the follow-up packet (see Figure 2) and the preceding data collected in the Telemetry Data TLV. The digest then MUST be truncated to 128 bits and written into the Digest field. Distribution and management of shared keys are outside the scope of this document. In the HTS authenticated mode, the Authentication sub-TLV MUST be present in each Telemetry Data TLV. HMAC MUST be verified before using any data in the included Telemetry Data TLV. If HMAC verification fails, the system MUST stop processing corresponding Telemetry Data TLV and notify an operator. Specification of the notification mechanism is outside the scope of this document.

7. IANA Considerations

7.1. IOAM Option-Type for HTS

The IOAM Option-Type registry is requested in [RFC9197]. IANA is requested to allocate a new code point as listed in Table 1.

Value	Name	Description	Reference
TBA1	IOAM Hybrid Two-Step (HTS) Option-Type	HTS Exporting	This document

Table 1: IOAM Option-Type for HTS

7.2. HTS TLV Registry

IANA is requested to create "Hybrid Two-Step" registry group. IANA is requested to create the HTS TLV Type registry in "Hybrid Two-Step" registry group. All code points in the range 1 through 175 in this registry shall be allocated according to the "IETF Review" procedure specified in [RFC8126]. Code points in the range 176 through 239 in this registry shall be allocated according to the "First Come First Served" procedure specified in [RFC8126]. The remaining code points are allocated according to Table 2:

Value	Description	Reference
0	Reserved	This document
1- 175	Unassigned	This document
176 - 239	Unassigned	This document
240 - 251	Experimental	This document
252 - 254	Private Use	This document
255	Reserved	This document

Table 2: HTS TLV Type Registry

7.3. HTS Sub-TLV Type Sub-registry

IANA is requested to create the HTS sub-TLV Type sub-registry as part of the HTS TLV Type registry. All code points in the range 1 through 175 in this registry shall be allocated according to the "IETF Review" procedure specified in [RFC8126]. Code points in the range 176 through 239 in this registry shall be allocated according to the "First Come First Served" procedure specified in [RFC8126]. The remaining code points are allocated according to Table 3:

Value	Description	TLV Used	Reference
0	Reserved	None	This document
1	HMAC	Any	This document
2 - 175	Unassigned		This document
176 - 239	Unassigned		This document
240 - 251	Experimental		This document
252 - 254	Private Use		This document
255	Reserved	None	This document

Table 3: HTS Sub-TLV Type Sub-registry

7.4. HMAC Type Sub-registry

IANA is requested to create the HMAC Type sub-registry as part of the HTS TLV Type registry. All code points in the range 1 through 127 in this registry shall be allocated according to the "IETF Review" procedure specified in [RFC8126]. Code points in the range 128 through 239 in this registry shall be allocated according to the "First Come First Served" procedure specified in [RFC8126]. The remaining code points are allocated according to Table 4:

Value	Description	Reference
0	Reserved	This document
1	HMAC-SHA-256 16 octets long	This document
2 - 127	Unassigned	This document
128 - 239	Unassigned	This document
240 - 249	Experimental	This document
250 - 254	Private Use	This document
255	Reserved	This document

Table 4: HMAC Type Sub-registry

8. Security Considerations

Nodes that practice the 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.

This document defines the use of authentication (Section 6) to protect the integrity of the telemetry information collected using the HTS method. Privacy protection can be achieved by, for example, sharing the IPsec tunnel with a data flow that generates information that is collected using HTS.

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.

9. 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. Also, the authors thank Bjørn Ivar Teigen for the discussion about ensuring proper correlation between generated telemetry information and an HTS Follow-up packet. And a special thank you to Xiao Min for thorough review and thoughtful suggestions that helped in improving the document.

10. References

10.1. Normative References

[RFC2104] Krawczyk, H., Bellare, M., and R. Canetti, "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, DOI 10.17487/RFC2104, February 1997, <https://www.rfc-editor.org/info/rfc2104>.

[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>.

[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>.

[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>.

10.2. Informative References

[I-D.song-ippm-postcard-based-telemetry] Song, H., Mirsky, G., Zhou, T., Li, Z., Graf, T., Mishra, G. S., Shin, J., and K. Lee, "On-Path Telemetry using Packet Marking to Trigger Dedicated OAM Packets", Work in Progress, Internet-Draft, draft-song-ippm-postcard-based-telemetry-16, 2 June 2023, <https://datatracker.ietf.org/doc/html/draft-song-ippm-postcard-based-telemetry-16>.

[P4.INT] "In-band Network Telemetry (INT)", P4.org Specification, November 2020.

[RFC4868] Kelly, S. and S. Frankel, "Using HMAC-SHA-256, HMAC-SHA-384, and HMAC-SHA-512 with IPsec", RFC 4868, DOI 10.17487/RFC4868, May 2007, <https://www.rfc-editor.org/info/rfc4868>.

[RFC6437] Amante, S., Carpenter, B., Jiang, S., and J. Rajahalme, "IPv6 Flow Label Specification", RFC 6437, DOI 10.17487/RFC6437, November 2011, <https://www.rfc-editor.org/info/rfc6437>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[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>.

Commented [TG23]: Should be normative since the document uses terms from RFC 7799. See https://datatracker.ietf.org/doc/statement-iesg-iesg-statement-normative-and-informative-references-20060419/

[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>.

[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>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

[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>.

[RFC9342] Fioccola, G., Ed., Cociglio, M., Sapio, A., Sisto, R., and T. Zhou, "Clustered Alternate-Marking Method", RFC 9342, DOI 10.17487/RFC9342, December 2022, <https://www.rfc-editor.org/info/rfc9342>.

[RFC9450] Bernardos, CJ., Ed., Papadopoulos, G., Thubert, P., and F. Theoleyre, "Reliable and Available Wireless (RAW) Use Cases", RFC 9450, DOI 10.17487/RFC9450, August 2023, <https://www.rfc-editor.org/info/rfc9450>.

Authors' Addresses

Greg Mirsky
Ericsson
Email: gregimirsky@gmail.com

Wang Lingqiang
ZTE Corporation
No 19 ,East Huayuan Road
Beijing
100191
China
Phone: +86 10 82963945
Email: wang.lingqiang@zte.com.cn

Guo Zhui
ZTE Corporation
No 19 ,East Huayuan Road
Beijing
100191
China
Phone: +86 10 82963945
Email: guo.zhui@zte.com.cn

Haoyu Song
Futurewei Technologies
2330 Central Expressway
Santa Clara,
United States of America
Email: hsong@futurewei.com

Pascal Thubert
Independent
06330 Roquefort-les-Pins
France
Email: pascal.thubert@gmail.com