In-band Telemetry for a Proactive SLA Monitoring Framework

draft-krishnan-opsawg-in-band-pro-sla-03

Abstract

The goal of in-band telemetry is to drive per packet, per hop real-time monitoring for the infrastructure towards achieving a programmable proactive SLA monitoring framework. Some of the key aspects from a switch/NIC perspective are - ingress/egress timestamp (latency), queue depth, bandwidth etc. Some of the key aspects from a server perspective are - cache/memory statistics etc. This document summarizes the current work in the industry in this area and identifies key requirements for a comprehensive solution. Towards addressing the requirements, this document describes uses cases and defines reusable monitoring packet formats across all layers in the OAM hierarchy.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.
The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on April, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC 2119].

Table of Contents

1. Introduction ....................................................... 4
   1.1. Acronyms .................................................... 5
2. In-band Telemetry for IPSEC tunnel packets .......................... 5
   2.1. Packet Format 1 - Geneve .................................. 6
   2.2. Packet Format 2 - VXLAN GPE ............................... 7
   2.3. Packet Format 3 - IP options ............................... 8
3. In-band Telemetry for Service Chaining ............................ 8
   3.1. NSH for service chaining Packet Format ....................... 10
   3.2. VXLAN-GPE for overlay and NSH for service chaining Packet Format .................................................. 10
   3.3. VXLAN-GPE for overlay and NSH for service chaining Packet Format .................................................. 11
4. Pre-construction/Minimizing of In-band Telemetry Header ........... 12
   4.1. Pre-construction of In-band Telemetry Header ............... 12
   4.2. Minimizing of In-band Telemetry Header for latency measurement .................................................. 12
5. Summarizing information in Telemetry header ........................ 13
6. IANA Considerations ............................................. 14
7. Security Considerations .......................................... 14
8. Acknowledgements ................................................ 14
9. References ...................................................... 15
9.1. Normative References.................................15
9.2. Informative References...............................15
1. Introduction

Proactive SLA monitoring is key for enabling DevOps in a converged infrastructure. As new services are continuously enabled using DevOps methodologies, it is critical to make sure that the users are delivered the promised SLAs through proactive SLA monitoring; in the case where SLAs are violated, the system should be able to automatically fix the issue or revert back to the old configuration as needed.

Standards-based monitoring schemes [ietf-twamp] are coarse grained – first, based on injected packets and not on customer data packets and next, lack of per hop visibility while monitoring end-to-end and last, lack of coverage for network functions.

New proposed monitoring schemes focus on switches/routers end-to-end in the DC – in-band network telemetry [p4-in-band] is to enable per packet, per hop monitoring for timestamp (latency), queue depth, bandwidth etc., Data-plane probe for in-band telemetry collection [ietf-in-band-dpp] is to enable the above per injected packet, [ietf-sfc-monitor] describes one-way latency monitoring for service chaining nodes using timestamps.

Given the above landscape, the key requirements for a comprehensive proactive SLA monitoring framework are as follows

- Ability to monitor selective flows, e.g. monitor only low latency traffic
- Ability to mirror selective flows which are monitored, e.g. mirror only low latency traffic (mirroring all flows may not scale)
- Ability to convey summarized information to the central management entity, e.g. alert the central management system only when a programmable percentile (for example 99.9th) latency exceeds a high threshold for a flow since mirroring entire flow may not scale
- Ability to strip monitoring information in the network edge since the application network stacks may not be able to process the additional monitoring information
- Ability to handle encrypted packets, e.g. enterprise cloud VPN across WAN, secure IaaS tunnels within a DC
Ability to monitor individual network function paths, e.g. VNF service chaining where several VNFs/VMs are sharing the same physical server.

Ability to address each layer in the OAM hierarchy in a generic way using a common monitoring format. Within a DC, the various OAM layers could be Service Function, Overlay and Underlay.

Ability to pre-construct the space for monitoring headers (described in Section 4) to guarantee deterministic performance especially for virtual network functions which are subject to a cache hierarchy in an industry standard server.

Ability to programmably select the hops being monitored to make sure the monitoring header size is bounded.

Towards addressing the key requirements, this document describes use cases and packet formats for handling encrypted data packets (e.g. IPSEC for IaaS deployment) and service chaining and also describes options for maintaining deterministic application performance while performing elaborate monitoring.

1.1. Acronyms

DPI: Deep Packet Inspection

MPLS: Multiprotocol Label Switching

NVGRE: Network Virtualization using Generic Routing Encapsulation

OAM: Operations, Administration, and Maintenance

SF: Service Function

SFC: Service Function Chain

SFP: Service Function Path

VXLAN: Virtual Extensible LAN

2. In-band Telemetry for IPSEC tunnel packets

The following describes in-band telemetry for IPSEC tunnels which is the most popular WAN tunneling protocol for secure communication.

Use Cases:
Cloud VPN: IPSEC tunnel between Enterprise branch and Enterprise/Cloud DC

- Primary use case for IPSEC is inter-domain, for example, enterprise branch office to PoP could be one network domain (operator A) and PoP to Enterprise/Cloud DC could be another network domain (Operator B), e.g. Google Cloud Interconnect

- Value proposition:
  - Real-time visibility/Service assurance for high priority tunnels carrying applications such as real-time voice/video
  - Minimal WAN switch/router buffer overprovisioning for all classes of traffic and maximizing WAN link utilization

Intra-DC: IPSEC tunnel between overlay end points for a private multi-tenant environment in a converged infrastructure (vlan, VXLAN provide isolation but not privacy)

- Value proposition:
  - Real-time visibility/Service assurance for high priority tunnels carrying applications such as transactional storage, real-time big data
  - Minimal DC switch/router buffer overprovisioning for all classes of traffic

There are several possible packet formats for achieving the above use cases. They are described below.

2.1. Packet Format 1 - Geneve

- Outer MAC Header
- Outer IP Header
  - IP protocol - UDP
  - Destination IP, Source IP, other fields
- Outer UDP Header
  - Destination UDP port - Geneve (6081)
. Outer Geneve Header

  oProtocol type - 0x6558 (RFC 1701- trans ethernet bridging)

  oOption Length - greater than zero

  oOption "INT"

    . Option Class (16 bits) - INT

    . Option Class needs to sync up with [ietf-geneve]

  oOption "Next Protocol" - new option (total length including
data is 8 bytes)

    . Option class (16 bits) - Next Protocol

      . Overrides protocol type in base Geneve header

      . Type (8 bits) - Critical bit is set, Lower 8 bit byte in
        4 bytes of data is protocol

      . Reserved (3 bits)

      . Length (5 bits) - set to 0x1 (4 bytes of data)

      . Data (32 bits) - for IPSEC - set to 0x0000032 (ESP) or
        0x00000033 (AH)

    . Encrypted or Authenticated payload

2.2. Packet Format 2 - VXLAN GPE

  . Outer MAC Header

  . Outer IP Header

    oIP protocol - UDP

    oDestination IP, Source IP, other fields

  . Outer UDP Header

    oDestination UDP port - VXLAN GPE (4790)

  . Outer VXLAN GPE Header
oNext Protocol - 0x5 - INT
  
  . Outer INT Header(s)

  oNext Protocol - ESP (0x7) or AH (0x8)
    
    . Need to create two new next protocols, aligning with [ietf-nsh] and [p4-in-band]

    . Encrypted or Authenticated payload

2.3. Packet Format 3 - IP options

Just like Geneve option format, IP options could be leveraged for in-band telemetry data.

  . Outer MAC Header

  . Outer IP Header

    oIP protocol - ESP (0x7) or AH (0x8)

    oDestination IP, Source IP, other fields

    oIP Header length > 5 (indicate presence of IP options)

  . Outer IP options Header

    oOption-type

      . Copied Flag - 1

      . Option Class - 2

      . Option Number

        . 10 - In-band Telemetry (new)

      . Option-Length - variable

      . Option-Data - in-band telemetry data

    . Encrypted or Authenticated payload

3. In-band Telemetry for Service Chaining

Use cases:
No. 1) Monitoring of the networking interconnect. This would typically involve monitoring the overlay/underlay across the individual service chain nodes and the service chaining header ([ietf-nsh] etc.) across the entire service chain at the entry and exit points.

No. 2) Monitoring of the individual network functions comprising a service chain using the service chaining header ([ietf-nsh] etc.). The network functions could be virtual (VMs, Containers etc.) or physical.

No. 1) and 2) Combination of above two use cases involving simultaneous monitoring of networking interconnect and individual network functions.

Value Proposition:

- Monitoring of the networking interconnect provides the usual benefits of underlay monitoring

- Monitoring of individual network functions through vSwitch/NIC helps rapid identification of any server side issues especially for virtual network functions.

  For example, a programmable percentile such as 99.9th latency/queue depth/queue drops exceeding a high threshold might mean collision in a scarce shared resource such as L1/L2/L3 cache in an industry standard server.

  The above information could be conveyed to a central management right away, which could remedy the situation by migrating the virtual network function to another appropriate industry standard server with no collision in a scarce shared resource such as L1/L2/L3 cache.

Typical elements involved in service chain monitoring are vSwitches/NIC/ToR. For each individual network functions comprising a service chain, vSwitch/NIC/ToR will monitor ingress traffic to the network function for one or more of the INT [p4-in-band] parameters such as timestamp, queue depth, bandwidth and egress traffic to the vSwitch/NIC/ToR for one or more of the aforementioned INT parameters. Monitoring of the entire service chain at the entry point involves monitoring traffic sent to the first network function from vSwitch/NIC/ToR and exit point involves monitoring traffic from the last network function to the vSwitch/NIC/ToR for one of the aforementioned INT parameters. For highly accurate monitoring, it is
recommended to use NIC/ToR vs a software based vSwitch. For example, HW implementations can measure timestamps to a nanosecond accuracy and can synchronize accurately with the master clock using protocols like IEEE 1588 PTP. A useful reference is [odl-nsh] which describes NSH service chaining operations from a ToR perspective.

Typical elements involved in underlay monitoring are ToR, Aggregation and Core switches/routers.

There are several possible packet formats for achieving the above use cases. Some are described here. More packet formats are work in progress.

3.1. NSH for service chaining Packet Format

- NSH Header
  - oNext Protocol - 0x5 - INT
    - Needs to sync with next protocol in [ietf-nsh]
  - NSH INT Header(s) (processed in vSwitch/NIC/ToR at each configured service chaining hop besides entry and exit points)
    - oNext Protocol - 0x3 - Ethernet
  - Inner Ethernet payload

3.2. VXLAN-GPE for overlay and NSH for service chaining Packet Format

- Outer MAC Header
- Outer IP Header
  - oIP protocol - UDP
    - oDestination IP, Source IP, other fields
- Outer UDP Header
  - oDestination UDP port - VXLAN GPE (4790)
- Outer VXLAN GPE Header
  - oNext Protocol - 0x5 - INT
- Outer INT Header(s)
.  NSH Header

.  NSH INT Header(s) (processed in vSwitch/NIC/ToR at each configured service chaining hop besides entry and exit points)

.  NSH INT Header(s) (processed in vSwitch/NIC/ToR at each configured service chaining hop besides entry and exit points)

.  Outer Ethernet payload

3.3. VXLAN-GPE for overlay and NSH for service chaining Packet Format

.  Outer MAC Header

.  Outer IP Header

.  Outer UDP Header

.  Outer VXLAN GPE Header

.  Outer INT Header(s)

.  NSH Header

oneeds to sync with next protocol in [ietf-nsh]

oneeds to sync with next protocol in [ietf-nsh]

.  NSH INT Header(s) (processed in vSwitch/NIC/ToR at each configured service chaining hop besides entry and exit points)
4. Pre-construction/Minimizing of In-band Telemetry Header

The following describes methods for pre-constructing and minimizing the In-band Telemetry header(s) and the corresponding benefits.

4.1. Pre-construction of In-band Telemetry Header

Method:

- Pre-construct timestamp header for all hops; Use timestamp append model in all switches/routers - virtual switch/router in server, switch/router in NIC, switch/router in TOR etc.

- Mirror packet with the entire timestamp information in the last hop; Examine the mirrored packet offline to determine any latency violations in near-real-time

Benefits:

- Last hop does not have to support latency measurement
- Easy to implement in SW/HW switches/routers with minimal performance impact with programmable data path
- No MTU change as packet traverses multiple hops leading to deterministic performance especially for VNF workloads
- Applicable to switches/routers as well as network functions

Besides timestamp/latency measurement, the aforementioned scheme is applicable to other parameters such as queue depth, bandwidth, packet drops etc. with the key benefit of no MTU change as the packet traverses multiple hops.

4.2. Minimizing of In-band Telemetry Header for latency measurement

Method:

- Only 1 additional field in the packet for latency measurement
  - Cumulative timestamp - 4 bytes
- Computes hop-by-hop latency information using the last cumulative timestamp
Hop-by-hop latency = current timestamp - last cumulative timestamp

Update the cumulative timestamp field with the current timestamp

Mirror (replicate) the entire (or truncated) packet with the received cumulative timestamp and append hop-by-hop latency

Examine the mirrored information from packets of the same flow in an offline collector to determine any latency violations in near-real-time

Benefits:

1. Only *one* additional fields needed in the packet for cumulative timestamp
2. No MTU change as packet traverses multiple hops leading to deterministic performance
3. Applicable to switches/routers as well as network functions

5. Summarizing information in Telemetry header

Delivering every packet of the flow with detailed per hop telemetry information, for example through flow mirroring, to the central management system may not scale for certain use cases. Hence, the ability to convey summarized monitoring information to the central management entity, e.g. alert the central management system only when a programmable percentile (for example 99.9th) queue depth exceeds a high threshold for a flow is needed. A possible element for performing this function could be vSwitch; this function could be potentially implemented in the NIC too or this could be implemented as a separate process in Linux-based server. This needs a common policy definition across the monitoring summarizer element and the central management entity.

The following describes fields needed in the packet header which is conveyed from the summarizing entity to the central management entity; this example focuses on queue depth, but can be easily extended to other monitoring parameters such as latency etc.

Telemetry packet structure from monitoring summarizer element to central management entity for conveying queue depth information.
- Monitoring summarizer L2/L3/L4 header (TCP/UDP packet; SIP is monitoring summarizer node, DIP is central management entity

- L2/L3/L4 Flow Header - 256 bytes - from original packet, covers IPv6 also

- Monitoring summarizer node id - 6 bytes (unique mac address)

- Monitoring summarizer policy id - 4 bytes - common policy definition across monitoring summarizer element and the central management entity

  oExample policy: Programmable percentile (for example 99.9th) queue depth exceeds a pre-programmed high threshold

- No. of nodes where this policy was violated - 3 bits - sufficient to cover 3 Tier DC topology

- Contiguous List of nodes where this policy was violated - each entry has this structure

  oViolating node id - 6 bytes (unique mac address)

  oIngress packet interface id in violating node id - 2 bytes

  oEgress packet interface id in violating node id - 2 bytes

6. IANA Considerations

This draft does not have any IANA considerations.

7. Security Considerations

Flexibility must be provided to preserve/strip the in-band telemetry information across multiple operator domains to address privacy concerns.

8. Acknowledgements

The authors would like to thank Anoop Ghanwani, Jack Harwood from Dell EMC and Mukesh Hira, Sumit Verdi from VMware for all the discussions.
9. References

9.1. Normative References

9.2. Informative References


[ietf-twamp] "A Two-Way Active Measurement Protocol (TWAMP)," RFC 5357


Authors’ Addresses

Ram (Ramki) Krishnan
Support Vectors