Network Service Header Timestamping

draft-browne-sfc-nsh-timestamp-00.txt

Abstract

This draft describes a method of timestamping Network Service Header (NSH) encapsulated packets or frames on service chains in order to measure accurately hop-by-hop performance delays of application flows carried within the chain. This method may be used to monitor performance and highlight problems with virtual links (vlinks), Virtual Network Functions (VNFs) or Physical Network Functions (PNFs) on the Rendered Service Path (RSP).

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 June 4, 2016.
1. Introduction

Network Service Header (NSH), as defined by [NSH], defines a method to insert a service-aware header in between payload and transport headers. This allows a great deal of flexibility and programmability in the forwarding plane allowing user flows to be programmed on-the-fly for the appropriate Service Functions (SFs).
Whilst NSH promises a compelling vista of operational agility for Service Providers, many service providers are concerned about losing service visibility in the transition from physical appliance SFs to virtualized SFs running in the Network Function Virtualization (NFV) domain. This concern increases when we consider that many service providers wish to run their networks seamlessly in 'hybrid' mode, whereby they wish to mix physical and virtual SFs and run services seamlessly between the two domains.

This draft describes a generic method to monitor and debug service chains and application performance of the flows within a service chain. This method is compliant with hybrid architectures in which VNFs and PNFs are freely mixed in the service chain. This method also is flexible to monitor the performance of an entire chain or part thereof as desired. Please refer to [NSH] as background architecture for the method described in this document.

The method described in this draft is not an OAM protocol like [Y.1731] or [Y.1564] for example. As such it does not define new OAM packet types or operation. Rather it monitors the service chain performance for subscriber payloads and indicates subscriber QoE rather than out-of-band infrastructure metrics.

2. Terminology

2.1. Requirement Language

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 [RFC2119].

2.2. Definition of Terms

Classification: Locally instantiated policy and customer/network/service profile matching of traffic flows for identification of appropriate outbound forwarding actions.

First TS Node (FTSN): Must mark packet correctly. Must understand 5 tuple information in order to match TS Controller flow table.

Last TS Node (LTSN): must read all MD & export to system performance statistics agent or repository. Should also send NSH header - the Service Index (SI) will indicate if a PNF(s) was at the end of the chain. The LTSN changes the SPI in order that the underlay routes the metadata back directly to the TSDB.
Network Node/Element: Device that forwards packets or frames based on outer header information. In most cases is not aware of the presence of NSH.

Network Overlay: Logical network built on top of existing network (the underlay). Packets are encapsulated or tunneled to create the overlay network topology.

Network Service Header: Data plane header added to frames/packets. The header contains information required for service chaining, as well as metadata added and consumed by network nodes and service elements.

NSH Proxy: Acts as a gateway: removes and inserts SH on behalf of a service function that is not NSH aware.

Service Classifier: Function that performs classification and imposes an NSH. Creates a service path. Non-initial (i.e. subsequent) classification can occur as needed and can alter, or create a new service path.

Service Function (SF): A function that is responsible for specific treatment of received packets. A service function can act at the network layer or other OSI layers. A service function can be virtual instance or be embedded in a physical network element. One of multiple service functions can be embedded in the same network element. Multiple instances of the service function can be enabled in the same administrative domain.

Service Function Chain (SFC): A service function chain defines an ordered set of service functions that must be applied to packets and/or frames selected as a result of classification. The implied order may not be a linear progression as the architecture allows for nodes that copy to more than one branch. The term service chain is often used as shorthand for service function chain.

Service Function Path (SFP): The instantiation of a SFC in the network. Packets follow a service function path from a classifier through the requisite service functions.

TS Controller: The TS Controller may be part of the service chaining application, SDN controller, NFVO or any MANO entity. For clarity we define the TS Controller separately here as the central logic that decides what packets to timestamp and how. The TS Controller instructs the classifier on how to mark the NSH header.
Timestamp Control Plane (TSCP): the control plane between the FTSN and the TS Controller.

Timestamp Database (TSDB): external storage of Metadata for reporting, trend analysis etc.

2.3. Abbreviations

FTSN     First Timestamp Node
LTSN     Last Timestamp Node
MD       Metadata
NFV      Network Function Virtualization
NFVI-PoP NFV Infrastructure Point of Presence
NIC      Network Interface Card
NSH      Network Service Header
OAM      Operations, Administration, and Maintenance
PNF      Physical Network Function
PNFN     Physical Network Function Node
QoE      Quality of Experience
RSP      Rendered Service Path
SCL      Service Classifier
SI       Service Index
SF       Service Function
SFC      Service Function Chain
SFN      Service Function Node
SFP      Service Function Path
TS       Timestamp
TSCP     Timestamp Control Plane
3. NSH Timestamping

As a generic architecture, please refer to Figure 1 below.

```
+----------------+     +----------------+     +----------------+     +----------------+
| TS Controller  |     | TSCP Interface |     | TSDB            |
|                |     |                |     |                 |
| ,---.           |     | ,---.          |     | ,---.           |
| /               |     | /             |     | /              |
| ( SCL )--------->( SF1 )--------->( SF2 )--------->( SFN ) |
| \ FTSN/        |     | \__/_          |     | \ LTSN/         |
| '---'          |     | '---'          |     | '---'           |
```

Figure 1 Logical roles in NSH Timestamping

The TS Controller will most probably be part of the SFC controller but is explained separately in this document for clarity. The TS Controller is responsible for initiating start/stop timestamp requests to the SCL or FTSN, and also for distributing timestamp NSH policy into the service chain via the Timestamping Control Plane (TSCP) interface.

The First Timestamp Node (FTSN) will typically be part of the SCL but again is called out as separate logical entity for clarity. The FTSN is responsible for marking NSH MD Type 0x2 fields for the correct flow with the appropriate NSH fields. This tells all upstream nodes how to behave in terms of timestamping at VNF ingress, egress or both, or ignoring the timestamp NSH metadata completely. The FTSN also writes the Reference Time value, a (possibly inaccurate) estimate of the current time-of-day, into the header, allowing the (chain,flow) performance to be compared to previous samples for offline analysis. The FTSN should return an error to the TS Controller if not synchronized to the current time-of-day and forward the packet along the service-chain unchanged.

SF1, SF2 timestamp the packets as dictated by the FTSN and process the payload as per normal.
Note 1: The exact location of the timestamp creation may not be in the VNF itself, as referenced in Section 3.3.

Note 2: Special cases exist where some of the SFs (PNFs or VNFs) are NSH-unaware. This is covered in Section 5.

The Last Timestamp Node (LTSN) should strip the entire header and forward the packet to the IP next hop. The LTSN also exports NSH timestamp information to the Timestamp Database (TSDB) for offline analysis; the LTSN may either export the timestamping information of all packets, or a subset based on packet sampling. In fully virtualized environments the LTSN will be co-located with the VNF that decrements the NSH Service Index to zero. Corner cases exist whereby this is not the case and is covered in section 5.

3.1. Prerequisites

In order to guarantee metadata accuracy, all servers hosting VNFs should be synchronized from a centralized stable clock. As PNFs do not timestamp there is no need for them to synchronize. There are two possible levels of synchronization:

Level A: Low accuracy time-of-day synchronization, based on NTP [RFC5905].

Level B: High accuracy synchronization (typically on the order of microseconds), based on [IEEE1588].

Each platform SHOULD have a level A synchronization, and MAY have a level B synchronization.

Level A requires each platform (including the TS Controller) to synchronize its system real-time-clock to an NTP server. This is used to mark the metadata in the chain, using the <Reference Time> field in the NSH timestamp header (Section 4.). This timestamp is written to the NSH header by the first SF in the chain. NTP accuracy can vary by several milliseconds between locations. This is not an issue as the Reference Time is merely being used as a reference inserted into the TSDB for performance monitoring.

Level B synchronization requires each platform to be synchronized to a Primary Reference Clock (PRC) using the Precision Time Protocol [IEEE1588]. A platform MAY also use Synchronous Ethernet ([G.8261], [G.8262], [G.8264]), allowing more accurate frequency synchronization.
If a SF is not synchronized at the moment of timestamping, it should indicate synch status in the NSH header. This is described in more detail in section 4.

By synchronizing the network in this way, the timestamping operation is independent of the current RSP, whether the entire chain is served by one NFVI-PoP or by multiple. Indeed the timestamp MD can indicate where a chain has been moved due to a resource starvation event as indicated in Figure 2 below, between VNF 3 and VNF 4 at time B.

```
Delay

v

v

x

x = reference time A

xv

v = reference time B

xv

VNF1 VNF2 VNF3 VNF4 VNF5
```

Figure 2 Flow performance in a service chain

3.2. Operation

Section 3.5 of [NSH] defines NSH metadata type 2 encapsulation as per the figure below. Please refer to the draft for a detailed explanation. Timestamped flows will use this format.

```
+---------------+---------------+---------------+---------------+---------------+
| Ver | O | C | R | R | R | R | R | R | Length |  MD-type=0x2 | Next Protocol |
+---------------+---------------+---------------+---------------+---------------+
|            |            | Service Path ID |            | Service Index |
+---------------+---------------+---------------+---------------+---------------+
| TLV Class     | Type         | R | R | R | Len |
+---------------+---------------+---------------+---------------+---------------+
| Variable Metadata |
+---------------+---------------+---------------+---------------+---------------+
```

Figure 3 NSH MD type 2 Encapsulation
The TS Controller should maintain a list of flows within each service chain to be monitored. This flow table should be in the format SPI:5 tuple ID. The TS Controller should map these pairs to unique Flow IDs per service chain within the extended NSH header specified in this draft. The TS Controller should instruct the FTSN to initiate timestamping on flow table match. The TS Controller may also tell the classifier the duration of the timestamping operation, either by a number of packets in the flow or by a time duration.

In this way the system can monitor the performance of the all en-route traffic, or an individual subscriber in a chain, or just a specific application the subscriber is running.

The TS Controller should write the list of monitored flows into the TSDB for correlation of performance data. Thus, when the TSDB receives data from the LTSN it understands to which flow the data pertains.

The association of source IP to subscriber identity is outside the scope of this draft and will vary by network application. For example, the method of association of a source IP to IMSI in mobile cores will be different to how a CPE with NAT function may be chained in an enterprise NFV application.

TSCP Interface

A new timestamp control plane (TSCP) interface is required between the TS Controller and the FTSN or classifier. This interface:

- Communicates which chains and flows to timestamp. This can be a specific {chain,flow} combination or include wildcards for monitoring subscribers across multiple chains or multiple flows within one chain.

- How the timestamp should be applied (ingress, egress, both or specific).

- When to stop timestamping.

Exact specification of TSCP is for further study.

3.3. Performance Considerations

This draft does not mandate a specific timestamping implementation method, and thus NSH timestamping can either be performed by hardware mechanisms, or by software. If software-based timestamping is used, applying and operating on the timestamps themselves incur an
additional small delay in the service chain. However, it can be assumed that these additional delays are all relative for the flow in question. Thus, whilst the absolute timestamps may not be fully accurate for normal non-timestamped traffic they can be assumed to be relative.

It is assumed that the monitoring method described in this document would only operate on a small percentage of user flows. The service provider may choose a flexible policy in the TS Controller to timestamp a selection of user-plane every minute for example to highlight any performance issues. Alternatively, the LTSN may selectively export a subset of the timestamps it receives, based on a predefined sampling method. Of course the TS Controller can stress test an individual flow or chain should a deeper analysis be required. We can expect that this type of deep analysis has an impact on the performance of the chain itself whilst under investigation. The impact will be dependent on vendor implementation and outside the scope of this document.

The timestamp may be applied at various parts of the NFV architecture. The VNF, hypervisor (assuming no SRIOV pass-through), vSwitch or NIC are all potential locations that can append the packet with the requested timestamp. Whilst it is desirable to timestamp as close as possible to the VNF for performance accuracy, the exact location of the timestamp application is outside the scope of this document, but should be consistent across the individual TS Controller domain.

4. NSH Timestamping Encapsulation

The NSH timestamping encapsulation is shown below in figure 4:
Figure 4 NSH Timestamp Encapsulation
Relevant fields in header that the FTSN must implement:

- The O bit should not be set as we are operating on subscriber packets
- The C bit should be set indicating critical metadata exists
- The MD type must be set to 0x2
- The TLV Class must be set to 0x10 (General KPI Monitoring) as requested in Section 9. The timestamp type is defined to be 0x01:
  - Type = 0x00 Reserved.
  - Type = 0x01 Timestamp.
- The MSB of the Type field must be set to zero. Thus if a receiver along the path does not understand the timestamping protocol it will pass the packet transparently and not drop. This scheme allows for extensibility to the mechanism described in this document to other KPI collections and operations.

The FTSN timestamp metadata contains the Timestamp Service Index (TSI) field which must be set to one of the following values:

- 0x0 Timestamp mode, no Service index specified in the TS Service Index field.
- 0x1 Timestamp Hybrid mode is selected, Timestamp Service Index contains LTSN Service index. This is used when PNFs or NSH-unaware SFs are used at the tail of the chain. If TSI=0x1, then the value in the type field informs the chain which SF should act as the LTSN.
- 0x2 Timestamp Specific mode is selected, Timestamp Service Index contains the targeted Service Index. In this case E&I bits are ignored and the Timestamp Service Index field indicates which SF is to be timestamped. Both ingress and egress timestamps are performed when the SI=TSSI on the chain. In this mode the FTSN will also apply the Reference Time and Ingress Timestamp. This will indicate the delay along the service chain to the targeted SF.
o 0x3 Timestamp E2E. In this case E&I bits are again ignores, FTSN writes ingress timestamp and reference time. This method may also be used as a light implementation to monitor end-to-end service chain performance whereby the targeted SF is the LTSN. The advantage here is very low overhead in the header and quick notification if there is a chain problem. This could then instigate a deeper examination of the chain performance.

The Flow ID is a unique 16 bit identifier written into the header by the classifier. This allow 65536 flows to be concurrently timestamped on any given NSH service chain (SPI). Flow IDs are not written by subsequent SFs in the chain. The FTSN exports monitored flow IDs to the TSDB for correlation.

The E bit should be set if Egress timestamp is requested.

The I bit should be set if Ingress timestamp is requested.

Reference Time is the wall clock of the FTSN, and may be used for historical comparison of SC performance. If the FTSN is not Level A synchronized (see Section 3.1.) it should inform the TS controller over the TSCP interface. The Reference Time is represented in 64-bit NTP format [RFC5905].

The Syn bits are an indication of the synchronization status of the node performing the timestamp and must be set to one of the following values:

- In Synch: 0x00
- In holdover: 0x01 (SF sees synch source and is in process of synching)
- Out of Synch: 0x02 (SF sees synch source but cannot synch to it)
- In free run: 0x03 (SF does not see acceptable external synch source)

If the network node is out of synch or in free run no timestamp is applied by the node (but other timestamp MD is applied) and the packet is processed normally.

If FTSN is out of synch or in free run timestamp request rejected and not propagated though the chain. The FTSN should inform the TS controller in such an event over the TSCP interface.
The outer service index value is copied into the timestamp metadata to help cater for hybrid chains that’s are a mix of VNFs and PNFs or through SFs that do not understand NSH. Thus if a flow transits through a PNF or an NSH-unaware node the delta in the inner service index between timestamps will indicate this.

The Ingress Timestamp and Egress Timestamp are represented in 64-bit NTP format [RFC5905]. The corresponding bits (I and E) reported in the timestamp metadata header.

The 64-bit timestamp format [RFC5905] is presented below:

```
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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Seconds                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            Fraction                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 5 NTP [RFC5905] 64-bit Timestamp Format

5. Hybrid Models

A hybrid chain may be defined as a chain whereby there is a mix of NSH-aware and NSH-unaware SFs. This may be the case if some PNFs are used in the chain or if VNFs are used that do not support NSH.

Example 1. PNF in the middle

```
TS
Controller
| TSCP Interface | TSDB

( SCL )-------->( SF1 )-------->( SF2 )-------->( SFN )
\ FTSN/          \     /          \ PNF1/          \ LTSN/
   '---'              '---'              '---'              '---'
```

Figure 6 Hybrid chain with PNF in middle

In this example the FTSN begins operation and sets the SI to 3, SF1 decrements this to 2 and passes the flow to an SFC proxy (not shown).
The proxy strips the NSH header and passes to the PNF. On receipt back from the PNF the Proxy decrements the SI and passes the packet onto the LTSN with a SI=1.

After the LTSN processes the traffic it knows it is the last node on the chain from the SI value and exports the entire NSH header and all metadata to the TSDB. The payload is forwarded to the next hop on the underlay minus the NSH header. The TS information packet is given a new SPI which acts as a homing tag to transport the timestamp data back to the TSDB.

Example 2. PNF at the end

```
| TS                  |
| Controller          |
| TSCP Interface      |
|                    | TSDS                 |
|                    | /                     |
|                    | /                     |
| ( SCL )-------->( SF1 )-------->( SF2 )-------->( PNF ) |
| \___/              | \___/                |
| \ LTSN/            | \ LTNS/              |
| \___/              | \___/                |
```

**Figure 7 Hybrid Chain with PNF at end**

In this example the FTSN begins operation and sets the SI to 3, the TSI field set to 0x1, and the type to 1. Thus when SF2 receives the packet with SI=1, it understands that it is expected to take on the role of the LTSN as it is the last NSH-aware node in the chain.

5.1. Targeted VNF Timestamp

For the majority of flows within the service chain, timestamps (ingress, egress or both) will be carried out at each hop until the SI decrements to zero and the NSH header and TS MD is exported to the TSDB. There may exist however the need to just test a particular VNF (perhaps after a scale out operation or software upgrade for example). In this case the FTSN should mark the NSH header as follows:

TSI field is set to 0x2. Type is set to the expected SI at the SF in question. When outer SI is equal to the TSSI, timestamps are applied at SF ingress and egress, and the NSH header and MD are exported to the TSDB.
6. Fragmentation Considerations

The method described in this draft does not support fragmentation. The TS Controller should return an error should a timestamping request from an external system exceed MTU limits and require fragmentation.

Depending on the length of the payload and the type of timestamp and chain length, this will vary for each packet.

In most service provider architectures we would expect a SI << 10, and that may include some PNFs in the chain which do not add overhead. Thus for typical IMIX packet sizes we expect to able to perform timestamping on the vast majority of flows without fragmenting.

7. Security Considerations

The security considerations of NSH in general are discussed in [NSH].

The use of in-band timestamping, as defined in this document, can be used as a means for network reconnaissance. By passively eavesdropping to timestamped traffic, an attacker can gather information about network delays and performance bottlenecks.

The NSH timestamp is intended to be used by various applications to monitor the network performance and to detect anomalies. Thus, a man-in-the-middle attacker can maliciously modify timestamps in order to attack applications that use the timestamp values. For example, an attacker could manipulate the SFC classifier operation, such that it forwards traffic through ‘better’ behaving chains. Furthermore, if timestamping is performed on a fraction of the traffic, an attacker can selectively induce synthetic delay only to timestamped packets, causing systematic error in the measurements.

An attacker that gains access to the TSCP can enable timestamping for all subscriber flows, thereby causing performance bottlenecks, fragmentation, or outages.

As discussed in previous sections, NSH timestamping relies on an underlying time synchronization protocol. Thus, by attacking the time protocol an attack can potentially compromise the integrity of the NSH timestamp. A detailed discussion about the threats against time protocols and how to mitigate them is presented in [RFC7384].
8. Open Items for WG Discussion
   o Specification and operation of TSCP
   o AOB

9. IANA Considerations

TLV Class Allocation

TLV classes are defined in [NSH].

IANA is requested allocate a new TLV class value:

0x10 KPI General Monitoring and timestamping type.

NSH Timestamping TLV Type

IANA is requested to set up a registry of "NSH Timesamping TLV Types". These are 7-bit values. Registry entries are assigned by using the "IETF Review" policy defined in [RFC5226].

IANA is requested to allocate two new types as follows:

   o Type = 0x00 Reserved.
   o Type = 0x01 Timestamp.

10. Acknowledgments

    The authors would like to thank Ron Parker of Affirmed Networks and Seungik Lee of ETRI for their reviews of this draft.

    This document was prepared using 2-Word-v2.0.template.dot.

11. References

11.1. Normative References


11.2. Informative References


Authors' Addresses

Rory Browne
Intel
Dromore House
Shannon
Co.Clare
Ireland

Email: rory.browne@intel.com