J. Kumar
INTERNET-DRAFT J. Kumar
Intended Status: Informational S. Anubolu
Z. He
R. Manur
Broadcom Limited
D. Cai
H. OU
AliBaba Inc.
Y. Li
S. Suwei
Huawei
Expires: September 6, 2018 March 5, 2018

Inband Flow Analyzer
draft-kumar-ifa-00

Abstract

Inband Flow Analyzer (IFA) records flow specific information from smart NIC and/or switches across the network. This document discusses the method to collect the data on a per hop basis across the network and perform localized analytics operations on it. This document also describes transport agnostic header definition for tunneled and non tunneled flows.

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/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at

http://www.ietf.org/shadow.html

Table of Contents

1. Introduction 3
1.1 Terminology 3
2. Scope, Applicability, and Motivation 3
3. IFA Operations 4
3.1 IFA Zones 5
3.2 IFA Function Nodes 5
3.2.1 Initiator Function Node 5
3.2.2. Transit Function Node 6
3.2.3. Terminating Function Node 6
3.3 IFA Header 6
3.4 IFA Metadata 9
3.5 IFA Analytics 12
3.6 IFA Ordered Set 12
3.7 IFA False Positives 12
3.7.1 Prevention Model - Filters at the edge of IFA Zone . . . 13
3.7.2 Detection and Drop Model - No configuration 13
4. Interoperability Considerations 13
5. Security Considerations 14
6. References 14
6.1 Normative References 14
6.2 Informative References 14
Authors' Addresses 14

1. Introduction

This document describes an Inband Flow Analyzer (IFA) mechanism to mark a packet to enable the collection of analyzed meta data with the analyzing flow. IFA defines an IFA header to mark the flow and mandate the collection of analyzed meta on per marked packet per hop basis across the network. This ability to mark the packet using IFA OAM header can be leveraged to create synthetic flows meant for network data collection. This document describes mechanism to emulate the live traffic and/or create synthetic flows. IFA avoids defining protocol header specific modifications for collecting and analyzing flows. IFA puts minimal requirements on the switching silicon. This document also describes IFA zones, IFA reports and IFA meta data.

1.1 Terminology

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 [RFC2119].

E2E: Edge to Edge

IFA: Inband Flow Analyzer

Geneve: Generic Network Virtualization Encapsulation [I-D.ietf-nvo3-geneve]

IOAM: In-situ Operations, Administration, and Maintenance

MTU: Maximum Transmit Unit

2. Scope, Applicability, and Motivation

This document describes the IFA deployment, type of traffic supported, header definitions, analytics and data path functions.

Scope: IFA deployment involves defining a IFA zone, understanding the requirements in terms of traffic overhead and points of data collection. Given that IFA provides ability to perform local analytics on the collected data, this document describes the scope of analytics function as well. Scope of IFA is from the smart NIC and/or ToR to any/all node in the network and can terminate in network and/or the smart NIC.

Applicability: IFA analyzes traffic and is encapsulation agnostic. Simple TCP and UDP flows as well as tunneled flows can be monitored. IFA can be enabled on smart NIC or can be just enabled on the network nodes. Enabling IFA on smart NIC provides better scalability and

visibility in the traffic. IFA best performs when there is a hardware assist for deriving the flow data in the data path. This document describes data path functions for IFA.

Motivation: Main motivation for IFA is to collect analyzed metadata on a per packet per flow basis for a given application. Since changing the application L4 header is not permissible, IFA attempts to create a sampled stream of application traffic and use it to collect the metadata along the application path. This sampled stream is later discarded. Provision is made to support inband insertion of metadata with flexibility to do payload or tail stamping. This draft attempts to define a marking of sampled or native packets using the IFA header so as to collect meta data in hardware. This draft also provides ability to handle false positives for the application traffic to be analyzed.

3. IFA Operations

IFA performs flow analysis and possible actions of the flow data inband. Once a flow is enabled for analysis, node with the role of "Initiator" makes a copy of the flow and tags them for analysis and data collection. Copying of the flow is done by sampling or cloning the flow. These new packets are representative packets of the original flow and poses exact same characterization as the original flow. This means that representative packets also called as IFA flow traverse the same path in the network and same queues in the networking element. Figure 1 show the IFA based Telemetry Framework. The terminating node is responsible to terminate the IFA flow by summarizing the metadata of the entire path and send it to Collector.

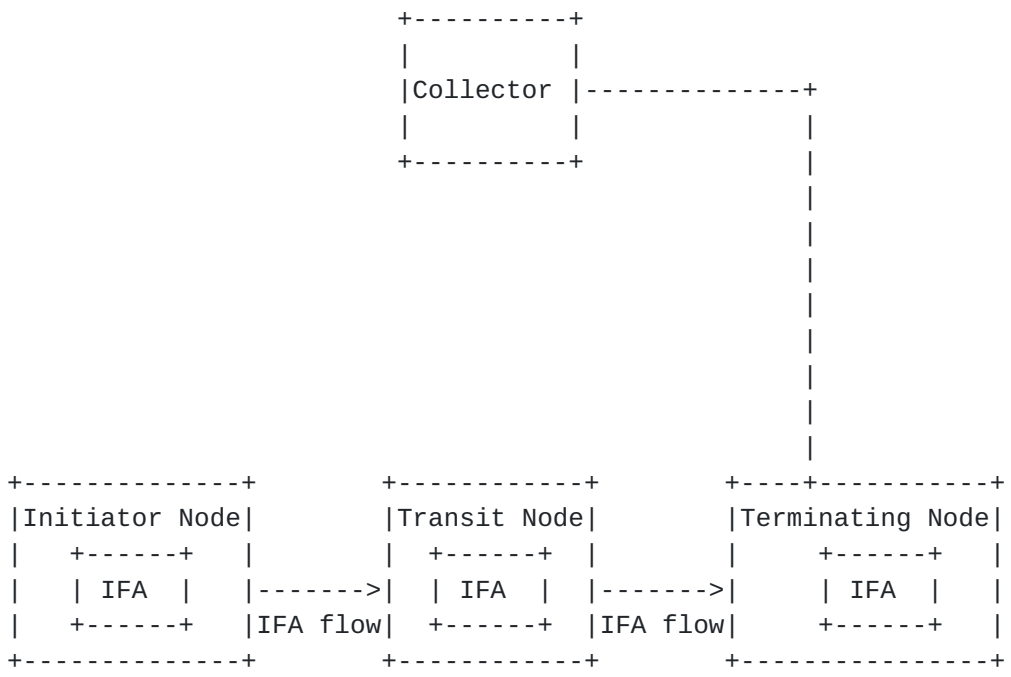

Figure 1 IFA Zone Framework

3.1 IFA Zones

IFA zone is the domain of interest where IFA monitoring is enabled. IFA zone MUST have designated IFA function nodes. IFA zone can also be controlled by setting appropriate TTL value in L3 header. Initiator and Terminating function nodes are always at the edge of the IFA zone. Internal nodes in the IFA zone are always Transit function nodes.

3.2 IFA Function Nodes

There are three IFA functional nodes.

3.2.1 Initiator Function Node

Smart NIC or ToR or any other node can perform the function of initiator. Better design is keep this role closest to the application if possible else flow visibility will be coarse. IFA initiator node performs following functions,

- Samples the flow traffic of interest based on a configuration.
- Converts the traffic into IFA flow by adding IFA header.
- Updates the packet with initiator node metadata.
- Re-inject the IFA flow in the network.
- May mandate a specific template id metadata by all networking

elements
- May mandate tail stamping of metadata by all networking elements

3.2.2. Transit Function Node

This node is responsible for inserting transit node metadata in the IFA packet.

3.2.3. Terminating Function Node

This node is responsible for following
- Insert terminating node metadata in the IFA packet
- Perform local analytics function on one or more segment of metadata for e.g. threshold breach for resident time, congestion notifications and so on.
- Terminate the IFA flow by summarizing the metadata of the entire path and send it to collector
- Drop the IFA flow

3.3 IFA Header

IFA header is a variation of https://tools.ietf.org/html/draft-lapukhov-dataplane-probe-01 header.

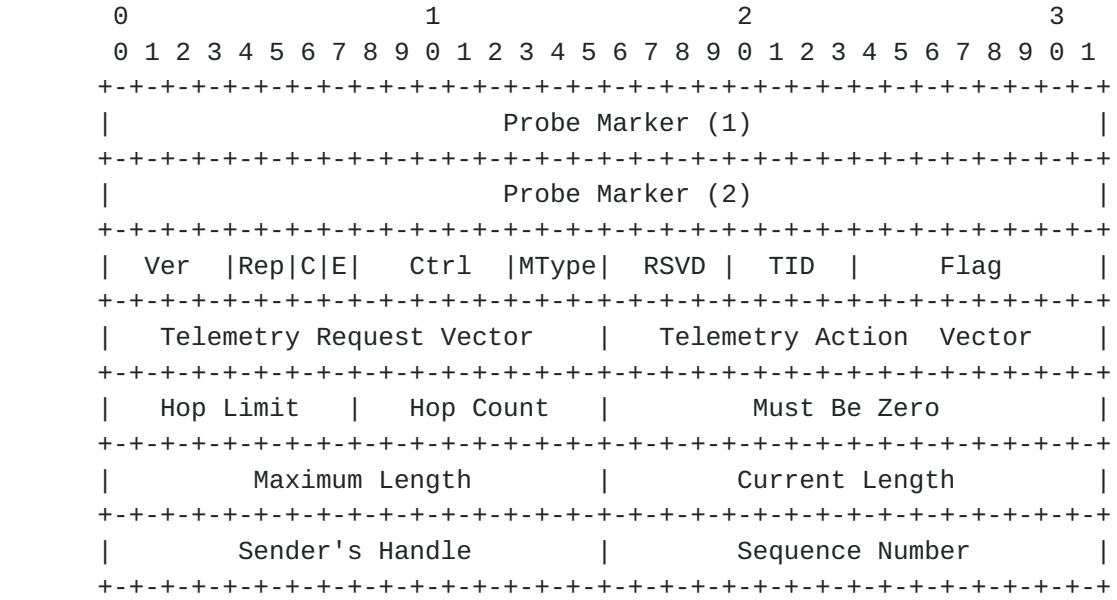

Figure 2 IFA Header Format

(1) The "Probe Marker" fields are arbitrary 32-bit values generally used by the network elements to identify the packet as a probe packet. These fields should be interpreted as unsigned integer values, stored in network byte order. For example, a network element

may be configured to recognize a UDP packet destined to port 31337 and having 0xDEAD 0xBEEF as the values in "Probe Marker" field as an active probe, and treat it respectively. These fields are initialized to a configured value.

(2) "Ver" is version and currently set to 1.

(3) "Rep" is replication requested. These bits are used to indicate replication of packet. This field is used to explore all the valid forwarding paths and is set to 00.

0: No replication requested.
1: Port level replication requested. This is application to LAG interfaces.
2: Next hop level replication requested. This is application to L3 ECMP paths.
3: Port and Next hop level replication requested. This is application to L3 ECMP over LAG interfaces

(4) "C" is copy requested. This bit is set for all the replicated packets to distinguish from the original packets. This bit is set to 0.

(5) "E" is max hop count exceeded. This bit is set when device can not add metadata because it has exceeded the hop count limit.

(6) "Ctrl" are the bits for local optimizations. For eg these bits can contain the instruction count in the "Telemetry Request Vector" field.

(7) The "MType" is message type and field value could be either "1" - "Probe" or "2" - "Probe Reply". This field is set to 1.

(8) "RSVD" are the reserved bits and must be initialized to 0.

(9) "TID" is the mandated template id and must be honored by all the networking elements in the path

(10) The "Flags" field is 8 bits, and defines the following flags:

1) Bit 0: "Overflow" (O-bit). This bit is set by the network element if the number of records on the packet is at the maximum limit as specified by the packet: i.e. the packet is already "full" of telemetry information.

2) Bit 1: "Inband" (I-bit). This bit if set indicates IFA is inband with terminating device disposing the IFA header, metadata stack and forwarding the packet. This bit is set by the

initiator node.

3) Bit 2: "TailStamp" (T-bit). This bit if set mandates all the network elements in the path to add the metadata at the tail of the packet. This bit is set by the initiator node.

4) Bit 4: "Template ID" (TID-bit). This bit if set mandates all the network elements in the path to insert specified template id of the metadata. This bit is set by the initiator node.

(11) "Telemetry Request Vector" is a 16-bit long field that requests well-known inband telemetry information from the network elements on the path. A bit set in this vector translates to a request of a particular type of information. The following types/bits are currently defined, starting with the least significant bit first. Telemetry request vector is always in context of a given template id. For eg template id 1 will have telemetry request vector as follows.

1) Bit 0: Device identifier.
2) Bit 1: Ingress port ID + egress port ID.
3) Bit 2: Hop latency.
4) Bit 3: Queue ID + Queue occupancy.
5) Bit 4: Ingress timestamp.
6) Bit 5: Egress timestamp.
7) Bit 6: Queue ID + Queue congestion status.
8) Bit 7: Egress port tx utilization

(12) "Telemetry Action Vector" is a 16-bit long field that requests inband telemetry metadata to be inserted based on the action indicated from the network elements on the path. A bit set in this vector translates to an action rule of a particular type of information. When the network node is able to perform some on-premises intelligence in deciding whether to insert metadata based on the criteria indicated by some vector bit, this vector can be set. The following types/bits are currently defined, starting with the least significant bit first:

1) Bit 0: Insert(1)/Ignore(0).
2) Bit 1: Queue depth exceed watermark for ECN.
3) Bit 2: Queue depth exceed watermark for PFC.
4) Bit 3: Resident delay breach.

(13) "Hop Limit" is treated as an integer value representing the number of network elements. See the Section 4 on the intended use of the field.

(14) The "Hop Count" field specifies the current number of hops of capable network elements the packet has transit through. It begins

with zero and must be incremented by one for every network element that adds a telemetry record. Combined with a push mechanism, this simplifies the work for the subsequent network element and the packet receiver. The subsequent network element just needs to parse the template and then insert new record(s) immediately after the template.

(15) The "Max Length" field specifies the maximum length of the telemetry payload in bytes. Given that the sender knows the minimum path MTU, the sender can set the maximum of payload bytes allowed before exceeding the MTU. Thus, a simple comparison between "Current Length" and "Max Length" allows to decide whether or not data could be added. Value of "0" means ignore.

(16) The "Current Length" field specifies the current length of data stored in the probe. This field is incremented by each network element by the number of bytes it has added with the telemetry data frame.

(17) The "Sender's Handle" field is set by the sender to allow the receiver to identify a particular originator of probe packets. Along with "Sequence Number" it allows for tracking of packet order and loss within the network.

3.4 IFA Metadata

This is the information inserted by each hop after the IFA header. IFA metadata can be inserted at following offsets

- Payload Stamping: After the layer 4 header
- Tail Stamping: After the end of packet

This document does not talk about merits or demerits of either approaches.

Each hop MUST provide "device id" and "TID" (template ID) to be able to identify the source and template of metadata it is inserting. A node may have multiple sensors corresponding to different sets of telemetry data collection. The contents and format of such set of telemetry data is defined in a template that is identified by template ID (TID). Each hop may support different "TID" and may insert metadata as per its own "TID". Collector must have a list of all supported "TID" in a network path to be able to decode the metadata. Templates must be published with its assigned TID. TID enables networking element to support diverse set of metadata and helps collector to decode the data appropriately.

Following is done at each hop before inserting IFA metadata.

(1) Overflow bit in Flags - Check if bit is set. If yes then do not insert the metadata and forward the packet.

(2) Hop Count - Increment by 1.

(3) Hop Limit - Decrement by 1. Check if the value has reached 0. If yes and it is a terminating node then perform the terminating node functions else just drop it.

(4) Current Length - Increase the current length by the size (in Bytes) of metadata added by network element. If "Max Length" field is non zero, perform the size exceeded check. If size is exceeded then set the "Overflow Bit" in "Flags field.

Data field of IFA metadata is shown below:

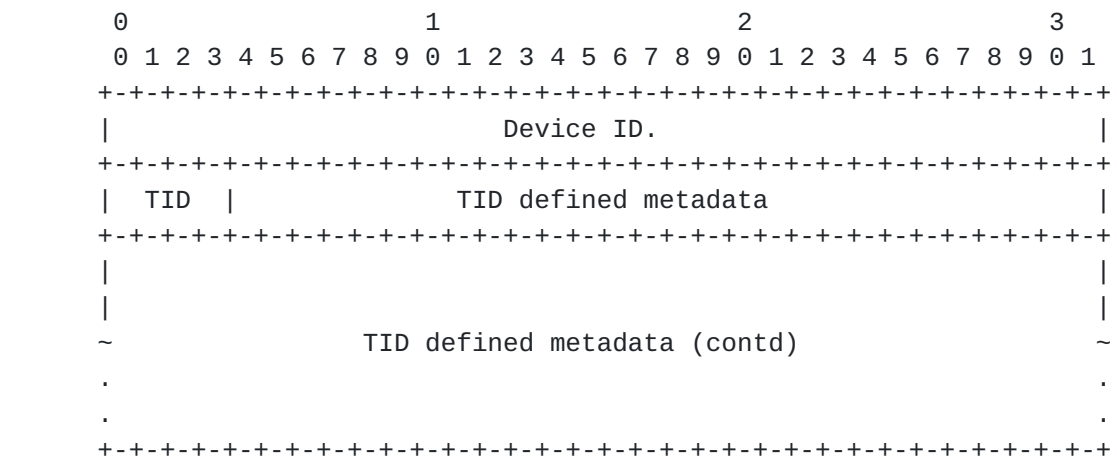

Figure 3 IFA Metadata Format

(1) Device ID: 32-bit unique device identifier. Note that Device ID is at fixed location at offset 0.

(2) TID: 4-bit template ID. Defines the following flexible format of metadata header.

(3) TID defined metadata: Variable length field. This data field is defined by the template identified by the TID. Some of the TID defined metadata is defined as follows.

When TID is 1, IFA metadata format is specified below.

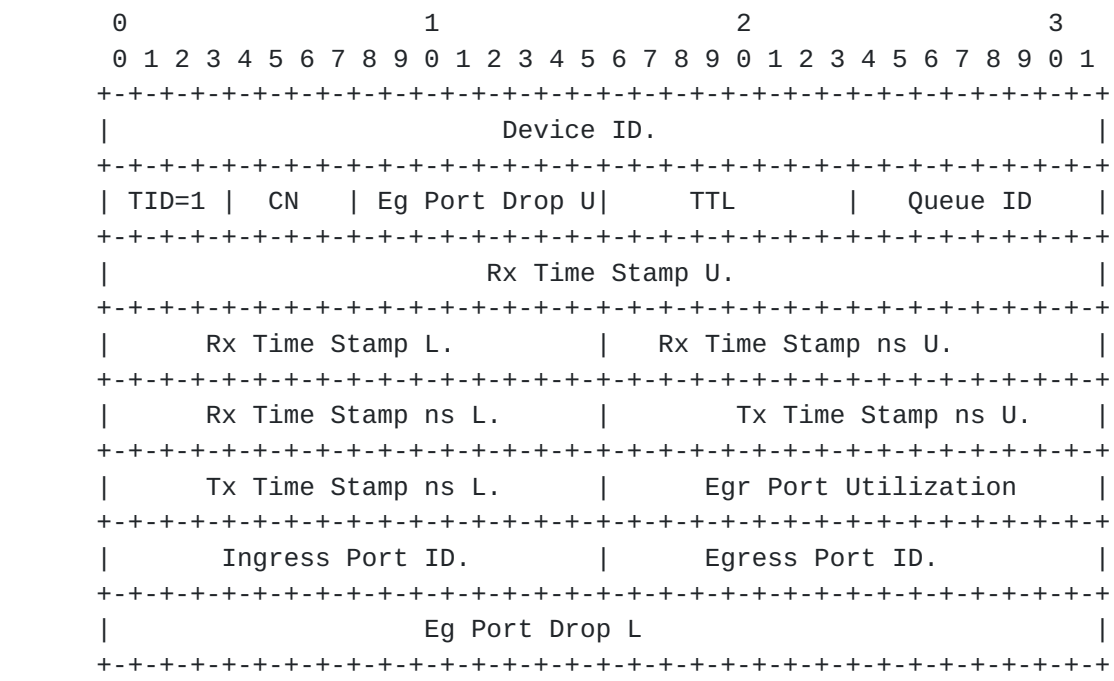

Figure 4 IFA Metadata Format when TID = 1

(1) Device ID: 32-bit unique device identifier. Note that Device ID is at fixed location at offset 0.

(2) TID: 1

(3) CN: 4b field indicating the state of congestion.

(4) Eg Port Drop U: Upper 8b of 40b egress drop count.

(5) TTL: 8b field initialized by Initiator function node.

(6) Queue ID: 16b egress queue ID.

(7) Rx Time Stamp U: Upper 32b of 48b seconds of Rx Time stamp.

(8) Rx Time Stamp L: Lower 16b of 48b seconds of Rx Time stamp.

(9) Rx Time Stamp ns U: Upper 16b of 32b nano sec of Rx Time stamp.

(10) Rx Time Stamp ns L: Lower 16b of 32b nano sec of Rx Time stamp.

(11) Tx Time Stamp ns U: Upper 16b of 32b nano sec of Tx Time stamp.

(12) Tx Time Stamp ns L: Lower 16b of 32b nano sec of Tx Time stamp.

(13) Egr Port Utilization: 16b egress port utilization (in %)

(14) Ingress Port ID: 12b ingress port id.

(15) Egress Port ID: 12b egress port id.

(16) Eg Port Drop L: Lower 32b of 40b egress drop count.

3.5 IFA Analytics

Once network path data is collected for a flow, IFA provides ability to act on the data. There are two kind of actions considered in this proposal.

(1) Action Bit MAP in IFA Header - This is encoded in the IFA header. Node in the path will use the action bitmap to insert or not insert the metadata based on threshold breach. Not insert operation is setting the field value to -1.

(2) Terminating Node Actions - Terminating node may decide to perform threshold or other actions on the set of metadata in the packet. This information is not encoded in the IFA header

3.6 IFA Ordered Set

TTL field in the IFA metadata is used to create an ordered set for the cases where network node is not capable of inserting the IFA metadata or inserts at the a different offset for e.g. as a Tail Stamp metadata.

Copying of TTL from outer IP header will be skipped for the IFA non compatible nodes. This will create a hole in TTL values in the set of IFA metadata in a packet. These holes are identified and can be used to either create null metadata for the receiver. If there is Tail Stamp metadata present then these holes are filled with the Tail Stamp metadata. This mechanism is implemented by terminating function to create a IFA metadata ordered set for the receiver.

3.7 IFA False Positives

One of the challenge of using probe signature in IFA header is a false positive. This draft proposes following actions to avoid any false positives.

False positive happens when payload of the packet matches the IFA

probe markers. This will trigger insertion of metadata and IFA header updates at each hop thereby corrupting the packet. If this is a packet belonging to real traffic then this corrupted packet will get forwarded to the application. If this is a sampled IFA packet it will result in drop of real traffic.

To avoid this condition, following two deployment models are considered.

3.7.1 Prevention Model - Filters at the edge of IFA Zone

This model requires installing global filters on all ports on the edge nodes of an IFA zone. Note that edge nodes are the initiator or terminator function nodes. This model requires careful configuration.

1) Initiator node MUST install a match rule attached to the port/flows being monitored for probe marker.

2) Initiator node MUST transition the detected packet to IFA packet by inserting IFA header with "O" and "I" Flag bits set. This will result in no metadata insertion.

3) Terminating node MUST detect the "I" flag in the IFA header. If set, it MUST dispose the IFA header and forward the packet per forwarding rules.

3.7.2 Detection and Drop Model - No configuration

This model doesn't require any configuration and relies on the fact that the any false positives will be dropped by the terminator node.

Drop and mirror functionality can be used to report these dropped packets.

Initiator node can install global rules for detection and reporting.

4. Interoperability Considerations

Some encapsulations use protocol specific identifications, e.g., a VXLAN-GPE Next Protocol value ([I-D.brockners-inband-oam-transport]) or a Geneve Option Class value ([I-D.draft-brockners-nvo3-ioam-geneve-00]) to indicate the presence of metadata. Similar approach can be used for IFA flow identification.

If the hardware supports IFA flow creation directly to live traffic and non-sampling based metadata collection from the terminating node

has no performance concern, IFA header and metadata can be inserted to live data packet without sampling, and the initiating and terminating nodes should work consistently and coordinately in inserting and stripping the metadata. The intermediate nodes are no change.

5. Security Considerations

TBD

6. References

6.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

6.2 Informative References

[I-D.brockners-inband-oam-transport] Brockners, F., Bhandari, S., Govindan, V., Pignataro, C.,Gredler, H., Leddy, J., Youell, S., Mizrahi, T., Mozes, D., Lapukhov, P., and R. Chang, "Encapsulations for In-situ OAM Data", draft-brockners-inband-oam-transport-05 (work in progress), July 2017.

[I-D.draft-brockners-nvo3-ioam-geneve-00] Brockners, F., Bhandari, S., Govindan, V., Pignataro, C.,Gredler, H., Leddy, J., Youell, S., Mizrahi, T., Mozes, D., Lapukhov, P., and R. Chang, "Geneve encapsulation for In-situ OAM Data", draft-brockners-nvo3-ioam-geneve-00 (work in progress), October 2017.

[INT Specification] https://p4.org/assets/INT-current-spec.pdf

Authors' Addresses

Jai Kumar
Broadcom Limited
Email: jai.kumar@broadcom.com

Surendra Anubolu
Broadcom Limited

Email: surendra.anubolu@broadcom.com

Zongying He
Broadcom Limited
Email: zongying.he@broadcom.com

Rajeev Manur
Broadcom Limited
Email: Rajeev.manur@broadcom.com

Dezhong Cai
AliBaba Inc.
Email: d.cai@alibaba-inc.com

Heidi OU
AliBaba Inc.
Email: heidi.ou@alibaba-inc.com

Yizhou Li
Huawei Technologies
EMail: liyizhou@huawei.com

Sun Suwei
Huawei Technologies
EMail: sunsuwei@huawei.com