Moonshot: Implementing the Next Generation of Network Telemetry Technologies

Andy Gospodarek
Broadcom Corporation

Abstract

Current network monitoring and telemetry applications require host-based collectors across all nodes in a network (both on servers/hypervisors and traditional switches and routers). These can be effective solutions, but just as datacenter deployment patterns have evolved new technology to track traffic as it moves through the network had emerged. Newer specifications like Inband Network Telemetry (INT) [2] and Inband Flow Analyzer (IFA) [1] propose standards to add metadata to packets or clone and add metadata as they flow through a network to allow collectors/agents to gather data at the network edges. Hardware that supports INT/IFA can add metadata automatically with application/flowlevel/virtual-port granularity which allows more detailed network monitoring and assurance to customers that service levels for applications are being met.

1 Introduction to Network Telemetry

INT and IFA are all designed create a generic method of reporting and collecting network state information on individual flows as the packets traverse a network. This allows for collection of data from individual hosts or applications as frames that are part of those flows are marked with telemetry headers as they entry a telemetry domain. Network devices can interpret telemetry header fields as *telemetry instructions* and a capable device will update packet headers and header-fields *In-Situ* – as a frame traverses the network. Marking frames as they travel through a network allows detailed reporting of the exact data-plane used by packets on the network as well as enables real-time feedback loops and event detection. This information can also be sent to an external collector for post-processing if desired.

1.1 Network Telemetry Components

Despite using slightly different nomenclature, the fundamental components of the two main telemetry technolgies covered in the paper are similar.

1.1.1 Source or Initiator Node

This is a trusted entity that creates the initial telemetry header and places it into packets that are transmitted.

1.1.2 Transit Hop or Transit Node

Any network element that adds telemetry metadata to a packet that that contains supported telemetry instructions.

1.1.3 Sink or Terminating Node

This is a trusted entity that removes telemetry headers from frames to make the existence of the the headers transparent to applications. This trusted entity will use the headers and other local configuration to determine if information needs to be sent to a collector.

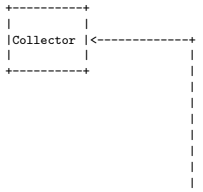
1.1.4 Collector

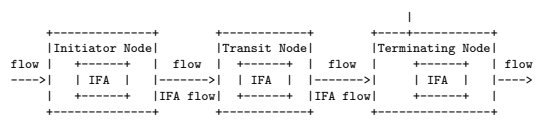
An application that will receive telemetry data collected by a Sink or Terminating node.

1.1.5 Typical Packet Flow

Below is a diagram of a typical packet path of a network flow and IFA flow through the components described in the previous section. In this case the IFA flow is a sample of the flow, so two frames travel between the Initiator Node and the Terminating Node.

This was adapted from the latest IFA specification and the time fo this writing:





In the case where an Initiator and Terminating nodes are switches or other forwarding elements on a network, flows could originate from an external device and exit to another device. Initiator and Terminating Nodes could also be servers with supported hardware and software stacks. In that case flows may originate or teminate on the IFA/INT node rather than originating/terminating from an external device.

This paper does not intend to cover the full scope of each telemetry technology and feature; anyone who would like to learn more should consider reading the latest INT and IFA specifications. It will cover some basic frame formats of each proposal in order to provide context for an implementation discussion.

2 Inband Network Telemetry (INT)

Inband Network Telemetry is a framework suggested by those interested in using P4 to create a programmable pipeline for networking forwarding elements. INT has multiple methods for collecting information about the network: frames are updated as they traverse the network or special *probe packets* are used to collect information about the network. In addition to defining the frame format and fields, the latest INT specification also conveniently provides a P4 program specification for INT Transmit.

2.0.1 INT Frame Format

The current INT specification describes the following formats as being able to support additional encapsulation headers to support INT:

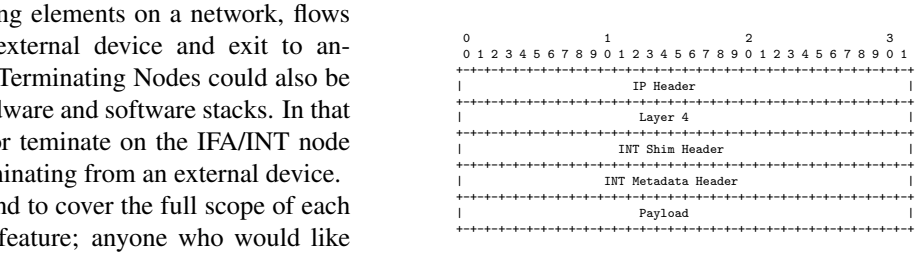
- INT over VXLAN (as VXLAN payload, per GPE extension)
- INT over Geneve (as Geneve option)
- INT over GRE (as a shim between GRE header and encapsulated payload)
- INT over NSH (as NSH payload)

Additionally the INT specification also describes how DSCP bits or *probe markers* can be placed in the payload of packet (after the Layer4 header) to support these packet formats.

- INT over TCP (as payload)
- INT over UDP (as payload)

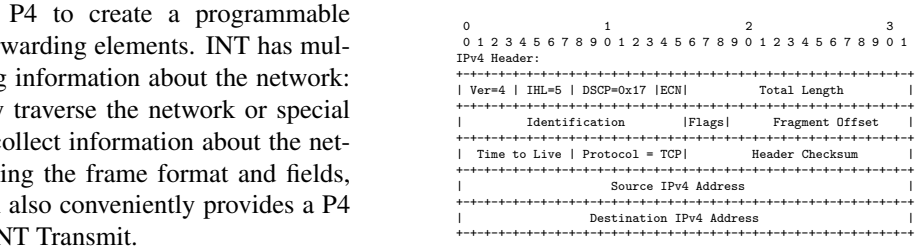
Though many datacenter networks use encapsulated traffic (VXLAN, Geneve, or GRE), the fact that INT does not have native support unencapsulated traffic (standard IPv4/IPv6 and TCP/UDP) could be an issue for some deployments.

The specified frame format for an INT IPv4/TCP frame would be as follows:



Remember that the INT Headers and Payload together are viewed as the full payload to any non-INT-aware device, so anytime INT headers are added to a packet any fields that account for the size of the packet or payload will need to be adjusted.

If the decision to use a reserved DSCP mark (0x17 in this case) to indicate a packet contained INT headers would cause the IPv4 header to look like this:



The INT specification also outlines suggestions for how to deal with frames as they grow beyond the MTU, how to deal with false detection of *probe markers* contained in payload data of non-INT frames, as well as other deployment scenarios.

3 Inband Flow Analyzer (IFA)

The initial IFA specification was drafted later than other initial telemetry technologies and while similar, it aims to address some of the shortcomings of INT and IOAM. One of the main differences is the ability to send telemetry metadata via a cloned frame rather than via the original datagram.

Allowing cloned frames provides benefits over In-Situ modification of frames. One benefit of cloning is administrators do not need to be concerned about frames growing beyond the MTU since there is also support to allow truncation of frames that are beyond the size of the MTU. The IFA specification also indicates that adding metadata to live traffic is a requirement but this cloning feature is a nice addition to avoid disruption of PMTU discovery.

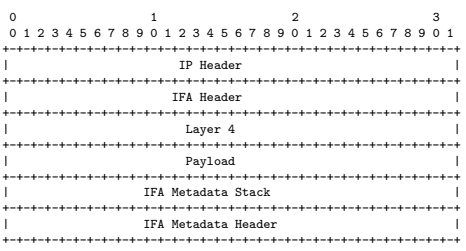
The proposed frame/header format was modified significantly from the INT specification. The goal was to make

the frame format more acceptable to devices that were not IFA-aware.

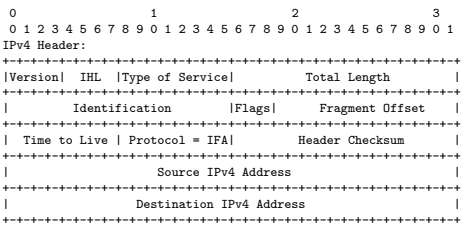
3.0.1 IFA Frame Format

The IFA spec outlines a significantly different scheme for the location to telemetry metadata. From the start IFA aims to interoperate with unencapsulated IPv4 and IPv6 traffic. This is accomplished by using the IPv4 *Protocol* and IPv6 *Next Header* fields to specify that this frame is an IFA frame. (There is no current reservation for IFA protocol, so testing currently uses one of the experimental protocol numbers.)

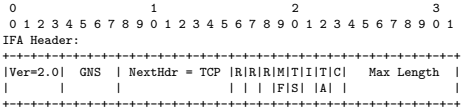
The specified frame format for an IFA IPv4/TCP frame would be as follows:



A closer look at the IPv4 header demonstrates that Protocol=IFA would be used to signal that this frame is an IFA frame:



Additionally the IFA Header provides a Next Header field that would indicate that TCP is the next protocol:



4 Hardware Requirements

Hardware requirements/implementations for handling IFA/INT today and the challenges facing those wanting to implement them in hardware and software.

5 Possible Configuration Methods

Proposals for configuring INT/IFA in both software dataplane and hardware dataplane environments on supported hardware.

5.1 Host-based configuration

5.2 Network-based configuration

6 Risks and Rewards

Assessing and minimizing risk associated with the maintenance of deploying an early draft of a standard.

7 Acknowledgments

Special thanks to all those invoved in IFA and INT specs as well as those at Broadcom who did not particiapte in the writing of the standard, but did provide code, feedback, etc during this process.

References

[1] J. KUMAR, E. A. Inband Flow Analyzer. *https://tools.ietf.org/pdf/draft-kumar-ippm-ifa-01.pdf*.

[2] KIN CHANGHOON, E. A. In-band Network Telemetry (INT). *https://p4.org/assets/INT-current-spec.pdf*.

Notes