Moonshot: Implementing the Next Generation of Network Telemetry Technologies

Andy Gospodarek
Broadcom Corporation
gospo@broadcom.com
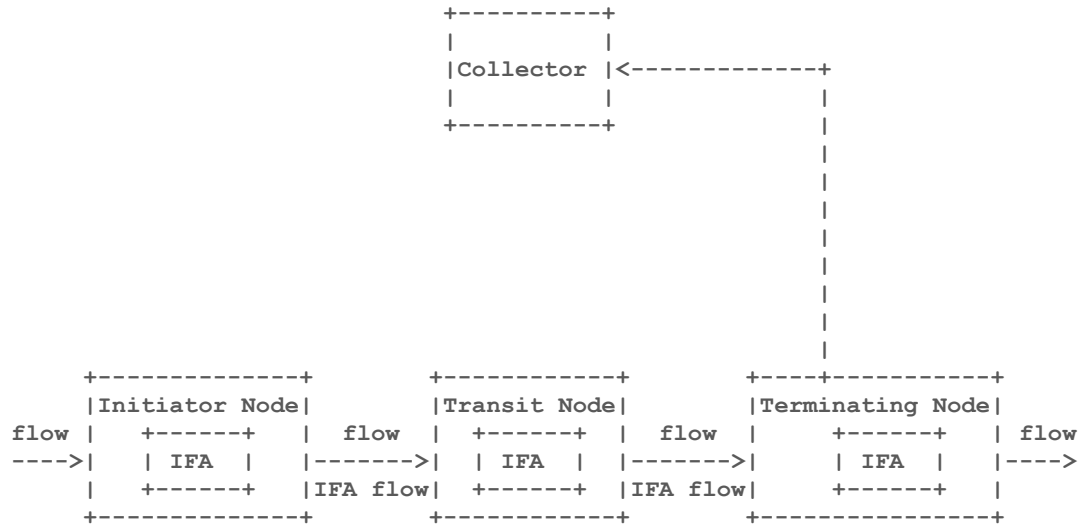
Introduction to Network Telemetry

Introduction to Network Telemetry

- Generic network-wide method for reporting and collecting network state information on packets as they traverse the network
- Packets contain *telemetry instructions* and each supported device responds to those instructions by adding metadata to frames
- *Telemetry instructions* can be added *In-Situ* to existing frames or to sampled frames that are transmitted simultaneously with existing traffic
- Standards like IOAM, INT, and IFA all describe ways to implement Network Telemetry

Network Telemetry Components

```
                        +----------+
                        |          |
                        |Collector |<------------+
                        |          |             |
                        +----------+             |
                                                 |
                                                 |
                                                 |
                                                 |
                                                 |
                                                 |
                                                 |
                                                 |
     +--------------+       +-----------+       +----+-----------+
     |Initiator Node|       |Transit Node|      |Terminating Node|
flow |    +------+  | flow  |   +------+ |  flow |    +------+    | flow
---->|    | IFA  |  |------->|   | IFA  | |------>|    | IFA  |    |---->
     |    +------+  |IFA flow|   +------+ |IFA flow|   +------+    |
     +--------------+       +-----------+       +----------------+
```

Source or Initiator Node

```
     +--------------+
     |Initiator Node|
flow |    +------+   |  flow
---->|   | IFA  |   |------->
     |    +------+   |IFA flow
     +--------------+
```

Node that adds telemetry instructions to frames -- in-band or out-of-band
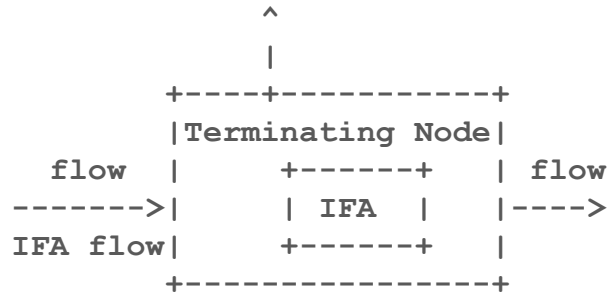
Transit Hop or Transit Node

```
        +------------+
        |Transit Node|
  flow  |  +------+  |   flow
------->|  | IFA  |  |------->
IFA flow|  +------+  |IFA flow
        +------------+
```

Any node that adds telemetry metadata to frame that contains telemetry instructions
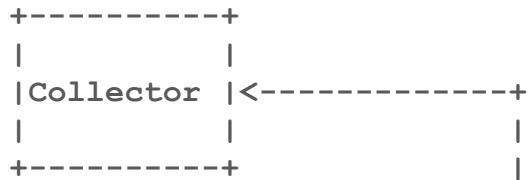
Sink or Terminating Node

```
               ^
               |
      +----+-----------+
      |Terminating Node|
 flow |      +------+   | flow
------->|     | IFA  |   |---->
IFA flow|     +------+   |
      +----------------+
```

Node that removes telemetry headers from frames and sends them to a collector

Collector

```
+----------+
|          |
|Collector |<------------+
|          |             |
+----------+             |
                         |
```

Any application that will receive and process telemetry data collected by Sink or Terminating Node
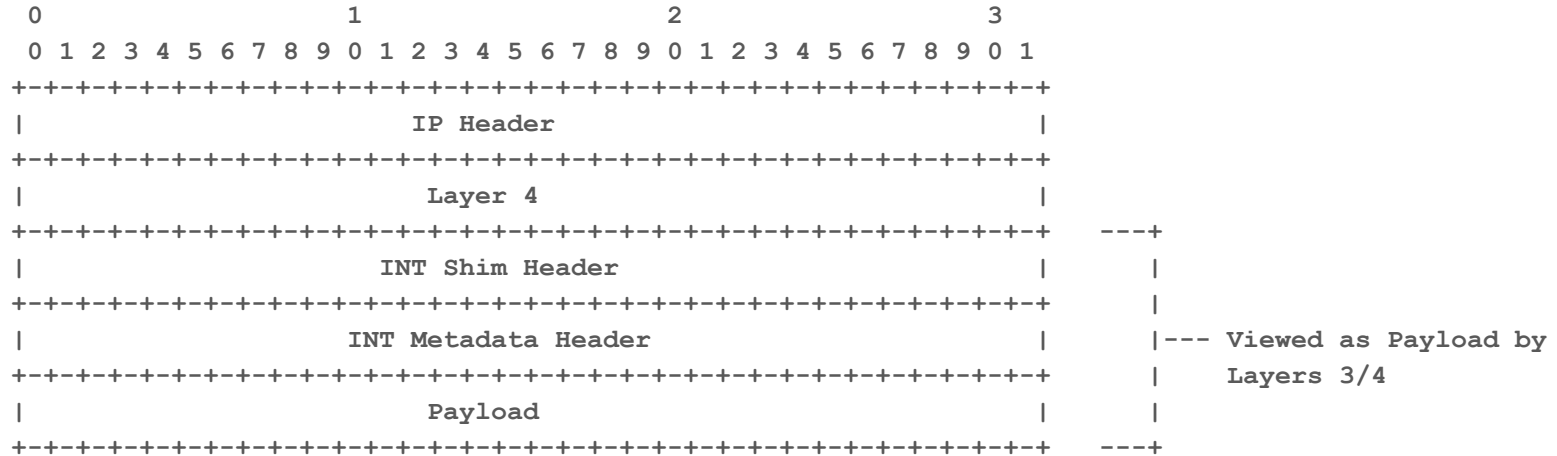
Inband Network Telemetry (INT)

Inband Network Telemetry (INT)

- Framework suggested by P4.org
- True In-band monitoring as frames are modified as they traverse the network
- Compatible with encapsulation formats that have extension/option capabilities
    - 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)
- Support for IP/IPv6 TCP/UDP traffic not as easy since there is not room for extensions
    - INT header and metadata transparently added to payload

INT IPv4/TCP Frame 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          IP Header                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Layer 4                             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  ---+
|                       INT Shim Header                        |      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+      |
|                     INT Metadata Header                      |      |--- Viewed as Payload by
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+      |     Layers 3/4
|                          Payload                            |      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+  ---+
```
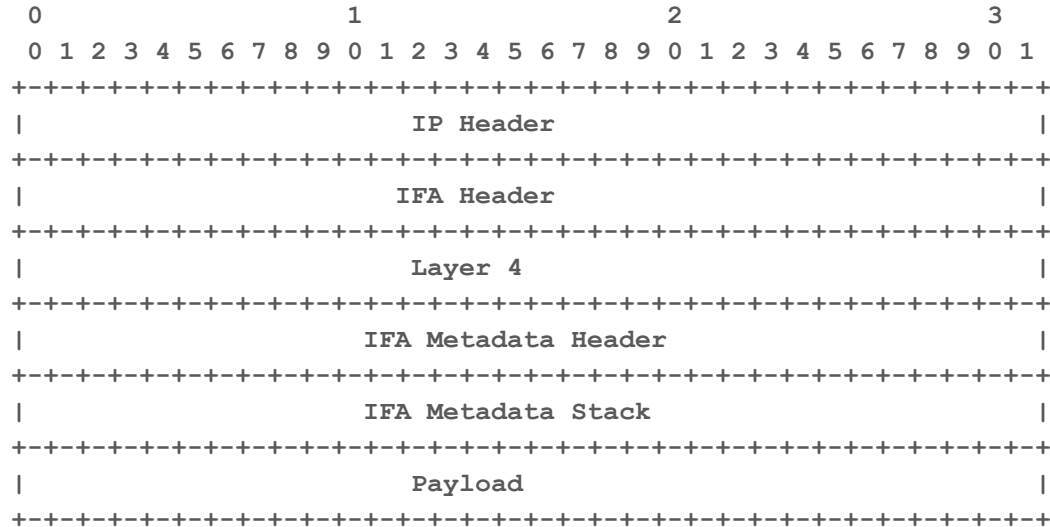
Inband Flow Analyzer (IFA)

- Specification released later to address perceived shortcomings of IOAM and INT
- Adds ability to sent telemetry instructions in sampled/copied frames rather than original frames
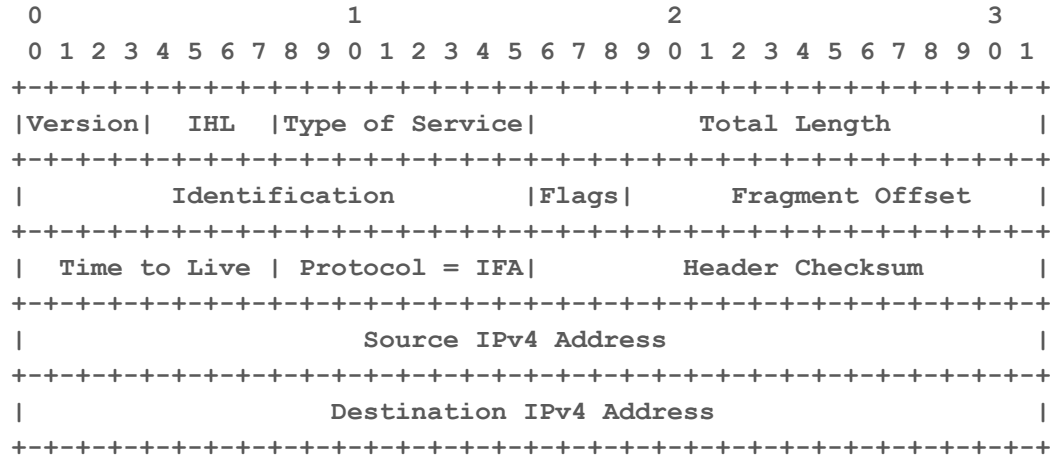- Frame format aims to be compatible with more hardware

IFA IPv4/TCP Frame 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          IP Header                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          IFA Header                           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Layer 4                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       IFA Metadata Header                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                       IFA Metadata Stack                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                          Payload                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

IFA IPv4 Header

```
 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
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version|  IHL  |Type of Service|          Total Length         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Identification        |Flags|      Fragment Offset    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time to Live | Protocol = IFA|         Header Checksum        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                        Source IPv4 Address                    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Destination IPv4 Address                 |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Telemetry Software Architecture

Source or Initiator Node

Desired Functionality	Kernel Implementation
Packet Sampling	TC or netfilter
Packet Mirroring (IFA)	TC
Packet Redirect	TC or XDP/eBPF
Encapsulate and Transmit	new lwtunnel type or XDP/eBPF

Transit Hop or Transit Node

Desired Functionality	Kernel Implementation
Match on INT/IFA Frame	TC
Update Metadata and Transmit	new lwtunnel type or XDP/eBPF

Sink or Terminating Node

Desired Functionality	Kernel Implementation
Match on INT/IFA Frame	TC or netfilter
Collect metadata from frame	new lwtunnel type or XDP/eBPF
Send frame to collector	new lwtunnel type or XDP/eBPF
Transmit original frame (INT)	new lwtunnel type or XDP/eBPF

Thinking Beyond the Linux Kernel Datapath

Hardware Support

- Switch Hardware
  - INT is by P4
  - Switch vendors and switch ASIC vendors are active in INT and IFA standards writing
- NIC Hardware
  - NIC vendors are active in INT and IFA standards writing
  - NIC+FPGA solutions for IOAM exist -- ASICs will add support soon

Network-based configuration

- Kernel datapath examples all presume configuration by netlink
- As hardware support for IFA is added, expect configuration over network
  - Common in switches
  - Becoming more common in servers used for baremetal deployments

Userspace Implementations

- Unlikely to be a DPDK application just for INT/IFA
- Libraries for DPDK-based applications could be created to add INT/IFA functionality to existing dataplane applications

Development and Deployment Risks

User Risk

- Protocols and packet types can change
- Risk is minimized when single organization controls entire infrastructure

Community Risk

- Early acceptance to Linux kernel could be problematic if software not kept up to date across infrastructure
- Risk is minimized when single organization controls entire infrastructure

Děkuji!