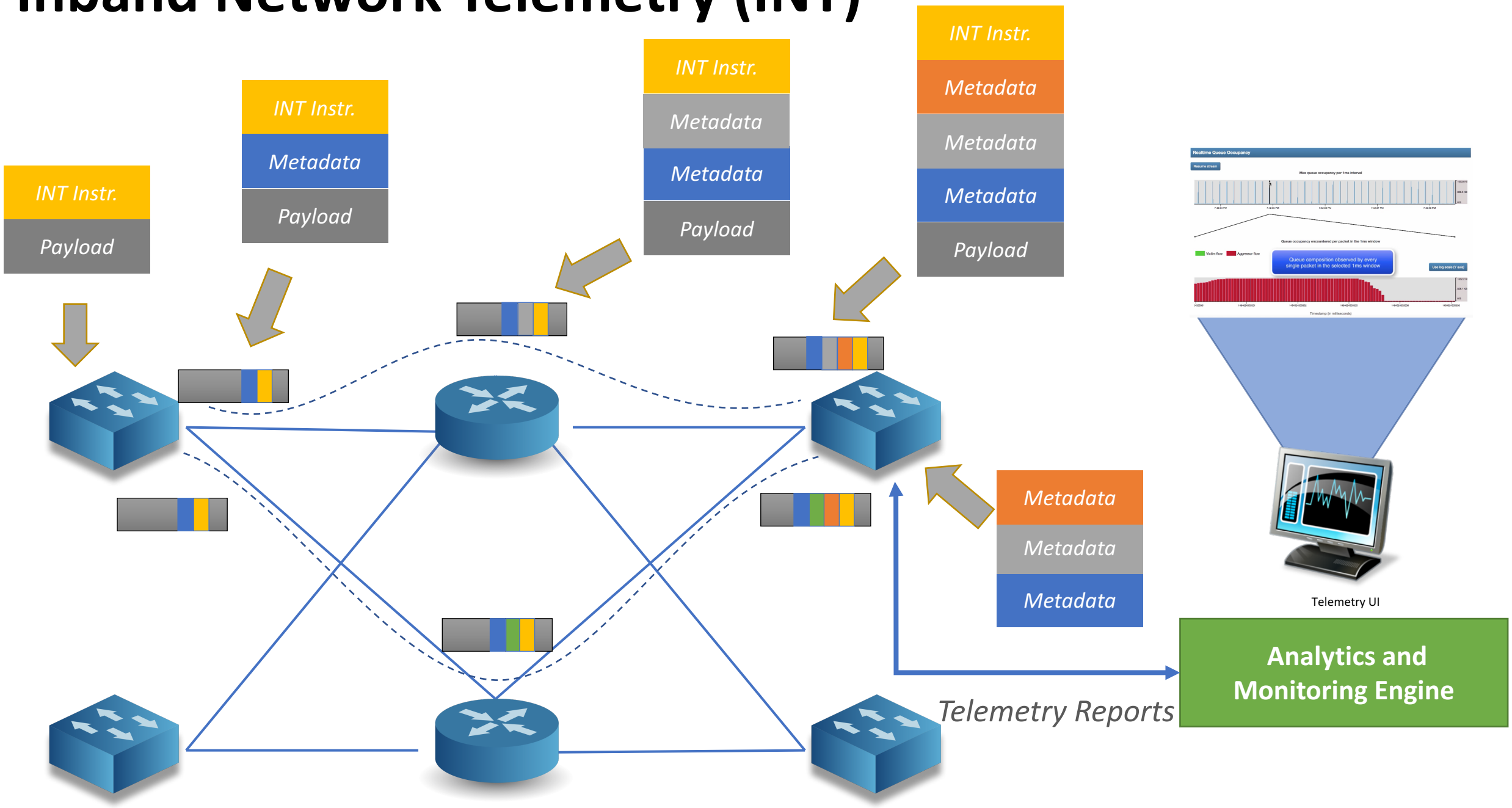
Proposed Telemetry Report Format

December 4, 2017

Mickey Spiegel, Jeongkeun Lee: *Barefoot Networks*
Gordon Brebner: *Xilinx*
Mukesh Hira: *VMware*

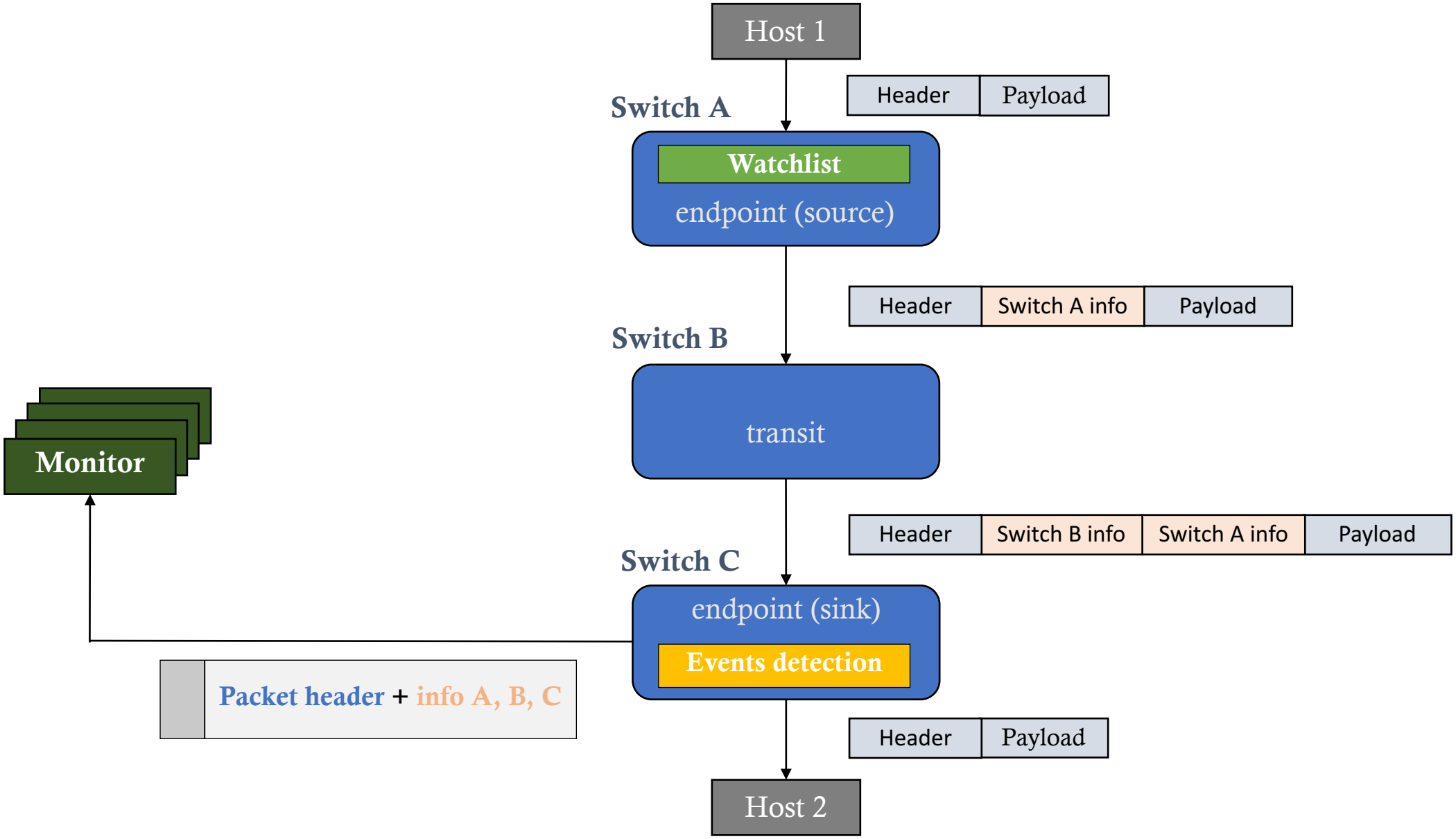
Inband Network Telemetry (INT)



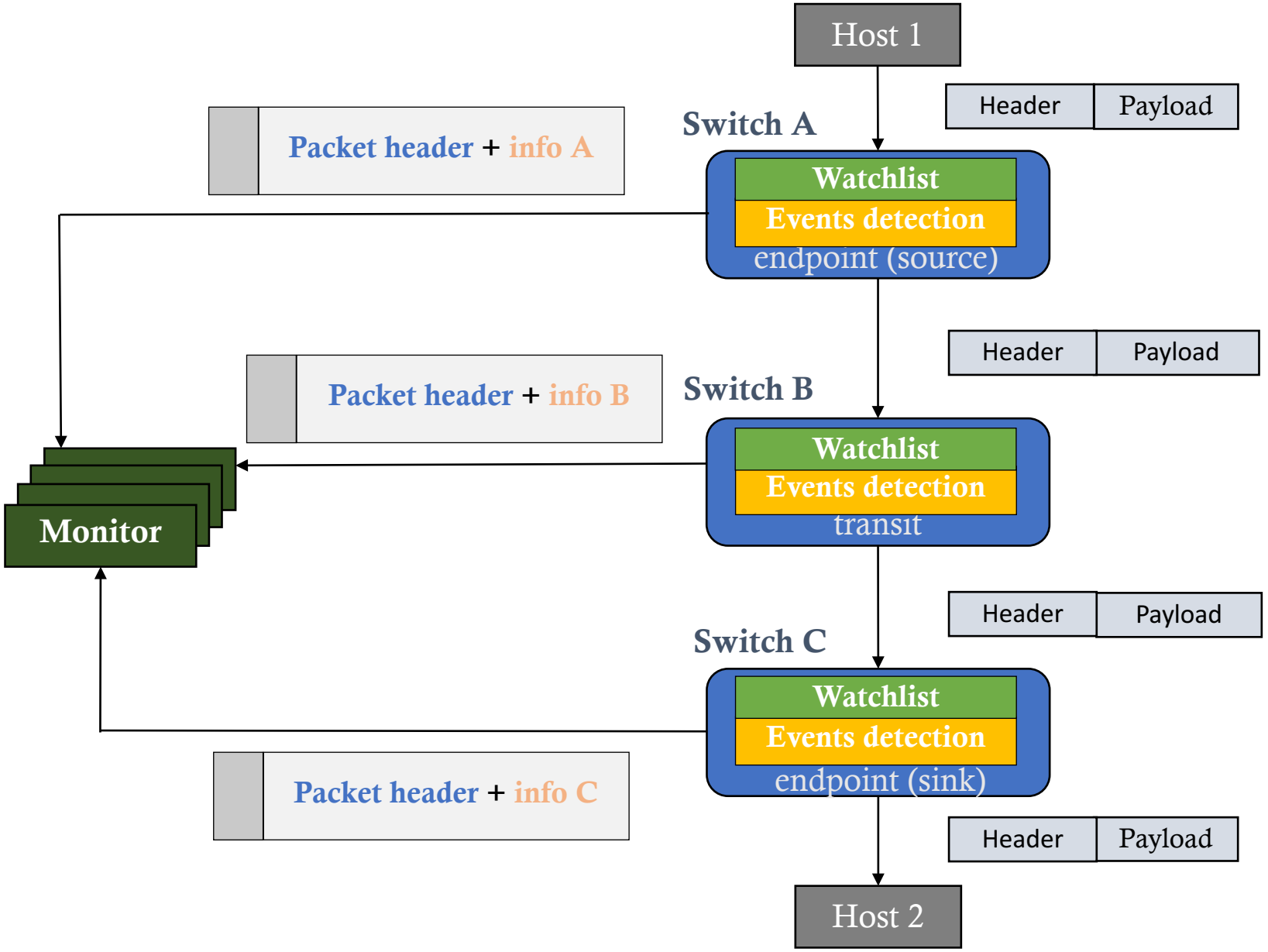
Scope

- Define packet formats for telemetry reports from data plane network devices (e.g. switches) to a distributed telemetry monitoring system
- Support two modes:
  - In-band (In-situ) Telemetry mode
    - Telemetry metadata is embedded in between the original headers of data packets as they traverse the network
    - Telemetry reports are generated by network devices at the edges of the network
    - Use any of the telemetry data plane specifications such as [INT] or [iOAM]
  - Postcard mode
    - Each network device generates its own telemetry reports
    - The distributed telemetry monitoring system will receive reports from different network devices, each describing the telemetry metadata (such as switch IDs, port IDs, latency) for one hop
    - No change to data packets traversing the network

In-band (In-situ) Telemetry Mode



Postcard Mode



Out of Scope

- Configuration of network devices so that they can determine when to generate telemetry reports, and what information to include in those reports, such as [SAI_DTel]
- Events that trigger generation of telemetry reports
- Selection of particular destinations within distributed telemetry monitoring systems, to which telemetry reports will be sent
- Export format for flow statistics or summarized flow records such as [IPFIX]

Key Concepts

- Telemetry Report
  - A message that a network device sends to the monitoring system, carrying
    - A snapshot of the original data packet (mostly the inner + outer headers) that triggered the report, which can be used for flow identification
    - Telemetry metadata collected from the reporting network device
    - Possibly telemetry metadata from upstream network devices (e.g. INT or IOAM)
- Telemetry Report Associations
  - *Tracked Flows*: Monitor data packets matching *flow watchlist*
  - *Dropped Packets*: Monitor dropped packets matching *drop watchlist*
  - *Congested Queues*: Packets entering a specific queue during a period of queue congestion
  - Different telemetry modes may be used for different associations
    - For example, INT for tracked flows, Postcard for dropped packets and congested queues

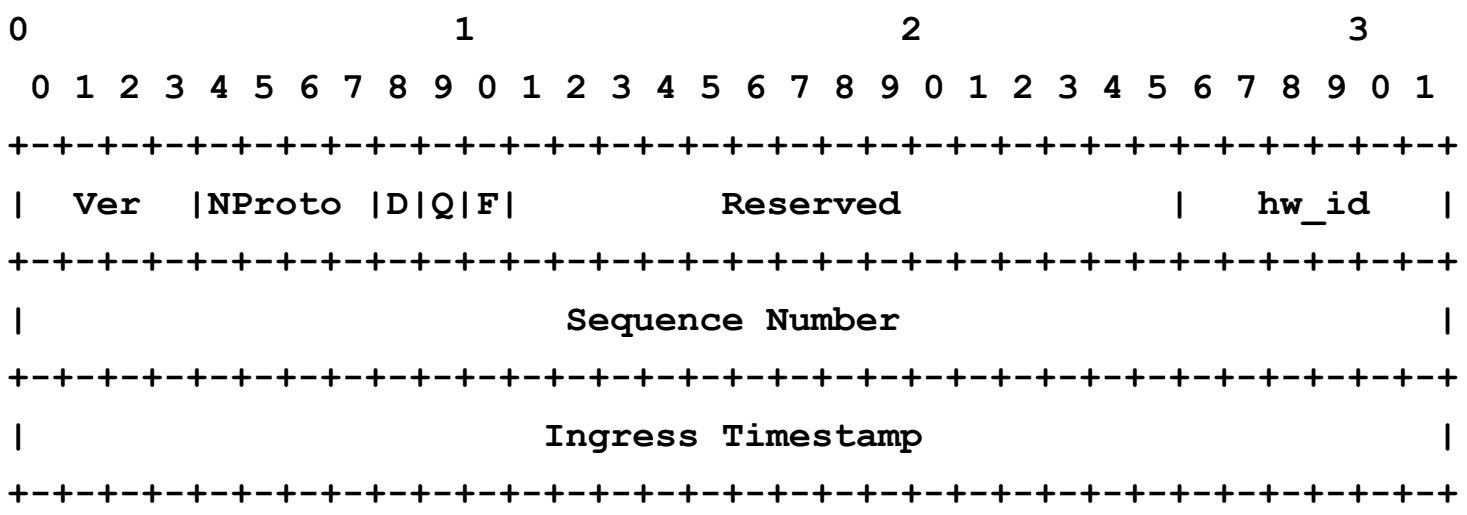
Key Concepts (2)

- Event Detection
  - Not every inspected packet needs to trigger a telemetry report
  - Network devices may apply filters to determine when significant events occur
    - For example, whenever a packet matching a tracked application flow is received or transmitted on a different path than previous packets, or
    - If a significant change in latency is experienced at one particular hop
  - This is left open for implementations to differentiate themselves
    - Beyond the scope of this specification
- Correlation of telemetry reports
  - Telemetry reports for a specific application flow may be received from multiple network devices
    - When using postcard mode, each hop will generate a separate telemetry report
    - When using in-band (in-situ) telemetry mode, in case of path change or in case of dropped packets
  - The distributed telemetry monitoring system may want to correlate these telemetry reports
    - Based on the original packet header fields included in each telemetry report
    - Telemetry reports include one association bit for each telemetry report category, e.g. to apply certain types of telemetry report correlation only when the corresponding bits are set
  - The mechanisms for correlation are left to each implementation
    - Beyond the scope of this specification

Outer Encapsulations

- UDP-based encapsulation
  - Various outer encapsulations may be used to transport the UDP telemetry report
  - Typically ethernet, followed by IPv4 or IPv6, followed by UDP
- IPv4 header fields
  - Source IP address identifies the network device that generates the telemetry report
  - Destination IP address identifies a location in the distributed telemetry monitoring system that will receive the telemetry report
  - In case of IPv4, as is the case for any other IP packet, either the Don't Fragment (DF) bit must be set, or the IPv4 ID field must be set so that the value does not repeat within the maximum datagram lifetime for a given source address/destination address/protocol tuple
- UDP header fields
  - Source Port may optionally be used to carry flow entropy, for example based on a hash of the inner 5-tuple
    - Otherwise, it should be set to 0
  - Destination Port is user configurable
    - Expectation is that the same Destination Port value will be used for all telemetry reports in a particular deployment

Telemetry Report Fixed Header (12 octets)



NProto: Next Protocol

- 0 Ethernet
- 1 Telemetry Drop header, followed by Ethernet
- 2 Telemetry Switch Local header, followed by Ethernet

D: Dropped - At least one packet matching a drop watchlist was dropped

Q: Congested Queue Association - Indicates the presence of congestion on a monitored queue

F: Tracked Flow Association - Matched a flow watchlist somewhere (INT or iOAM) or locally (postcard)

hw_id: Identifies the hardware subsystem that generated this report, e.g. a specific linecard in a chassis

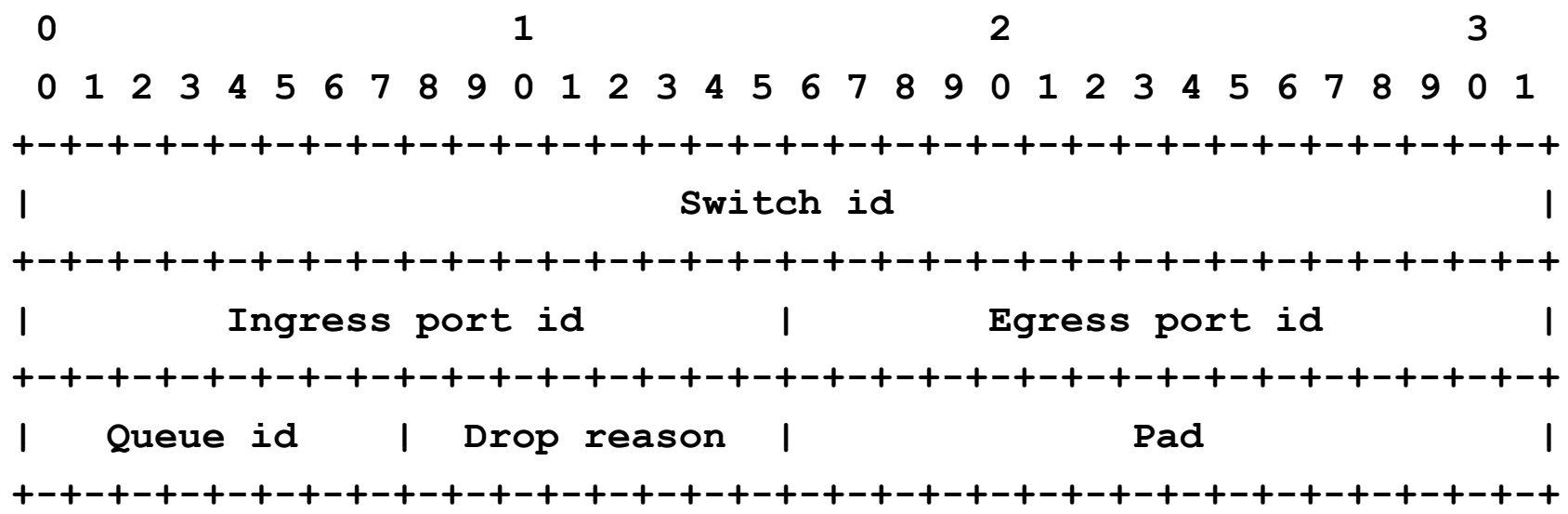
Sequence Number: From a specific hw_id to a particular destination, used to detect loss of telemetry reports

Ingress Timestamp: In nanoseconds

Next Protocol

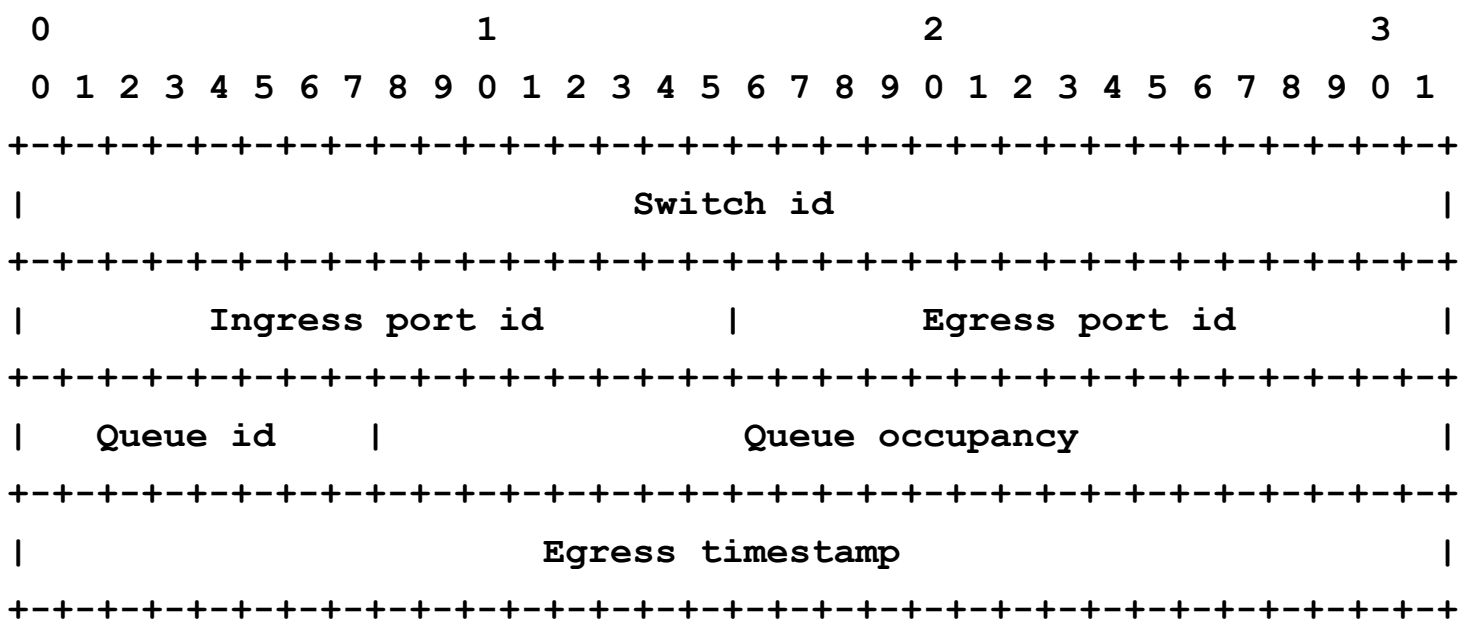
- Indicates the header format that immediately comes next: Drop, Switch Local, ethernet
- All you need for parsing the report packet, assuming that you parse header by header
- Ethernet includes many current and future cases:
  - INT over L4 (inner ethernet, inner IP, inner UDP/TCP, INT, payload)
  - INT over VXLAN (ethernet, IP, UDP, VXLAN-GPE, INT, inner eth, inner IP, payl)
  - IOAM over VXLAN
  - INT may include hop-by-hop header and/or destination header
  - IOAM may include Pre-allocated Trace Option, Incremental Trace Option, Edge-to-Edge Option, and/or Proof of Transit Option
  - INT/IOAM over Geneve
  - INT/IOAM over NSH

Telemetry Drop Report Header (12 octets)



- Defined in [INT]:
  - Switch id
  - Ingress port id
  - Egress port id
  - Queue id
- Drop reason
  - An enumeration that indicates the reason why a packet was dropped, for example as defined in github.com/p4lang/switch.

Telemetry Switch Local Report Header (16 octets)

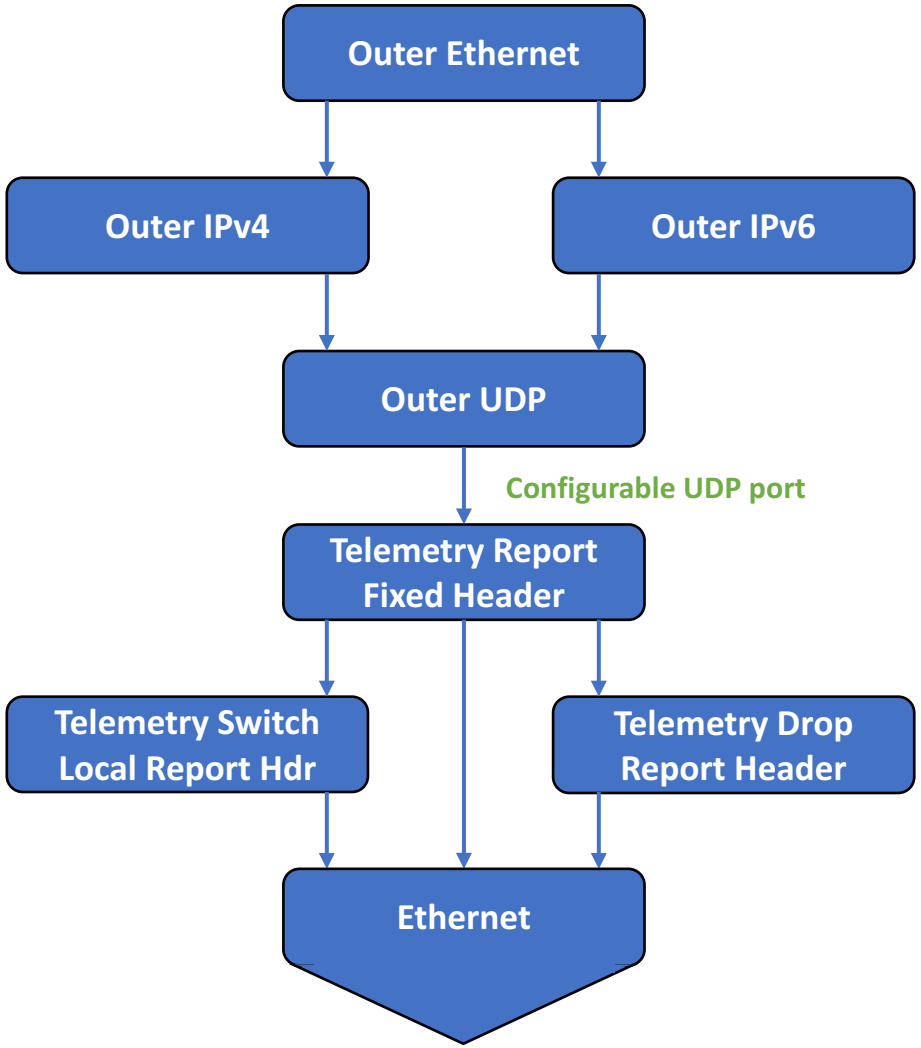


- Defined in [INT]:
  - Switch id
  - Ingress port id
  - Egress port id
  - Queue id
  - Queue occupancy
- Egress timestamp
  - The device local time when the packet was sent out the ***egress*** physical or logical port, in nanoseconds

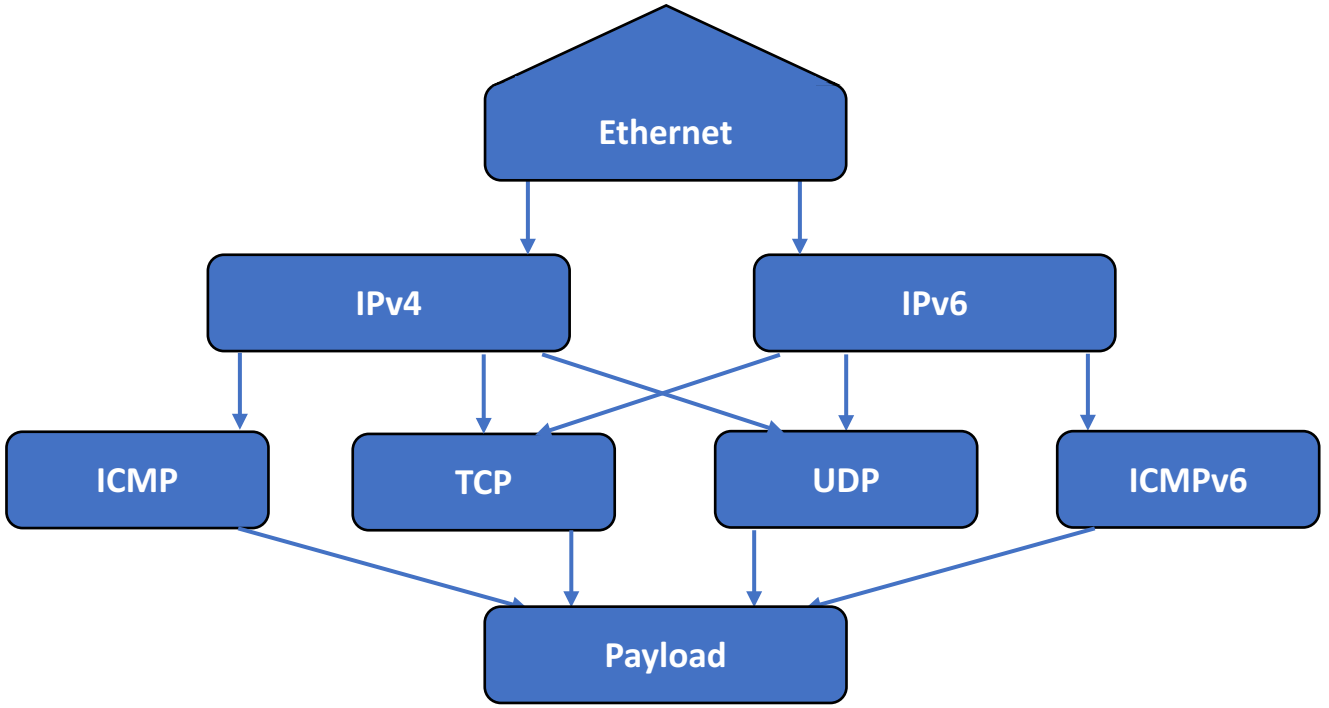
Embedded Telemetry Metadata

- Telemetry metadata may be embedded within the payload after the Telemetry Report headers
  - Typically for metadata from hops prior to the network device generating the report
  - Format is not specified as part of telemetry report format
    - Typically encoded using a defined data plane format such as [INT] or [iOAM]
- Tracked Flow Association bit provides a hint
  - If set to 0 => no embedded telemetry metadata
  - If set to 1 => may or may not be any embedded telemetry metadata
- A network device generating a telemetry report may include its local telemetry metadata in any of:
  - Embedded telemetry metadata
  - Telemetry Switch Local Report header or Telemetry Drop Report header
    - In the same telemetry report as embedded telemetry metadata from previous hops
    - In a separate telemetry report from embedded telemetry metadata from previous hops

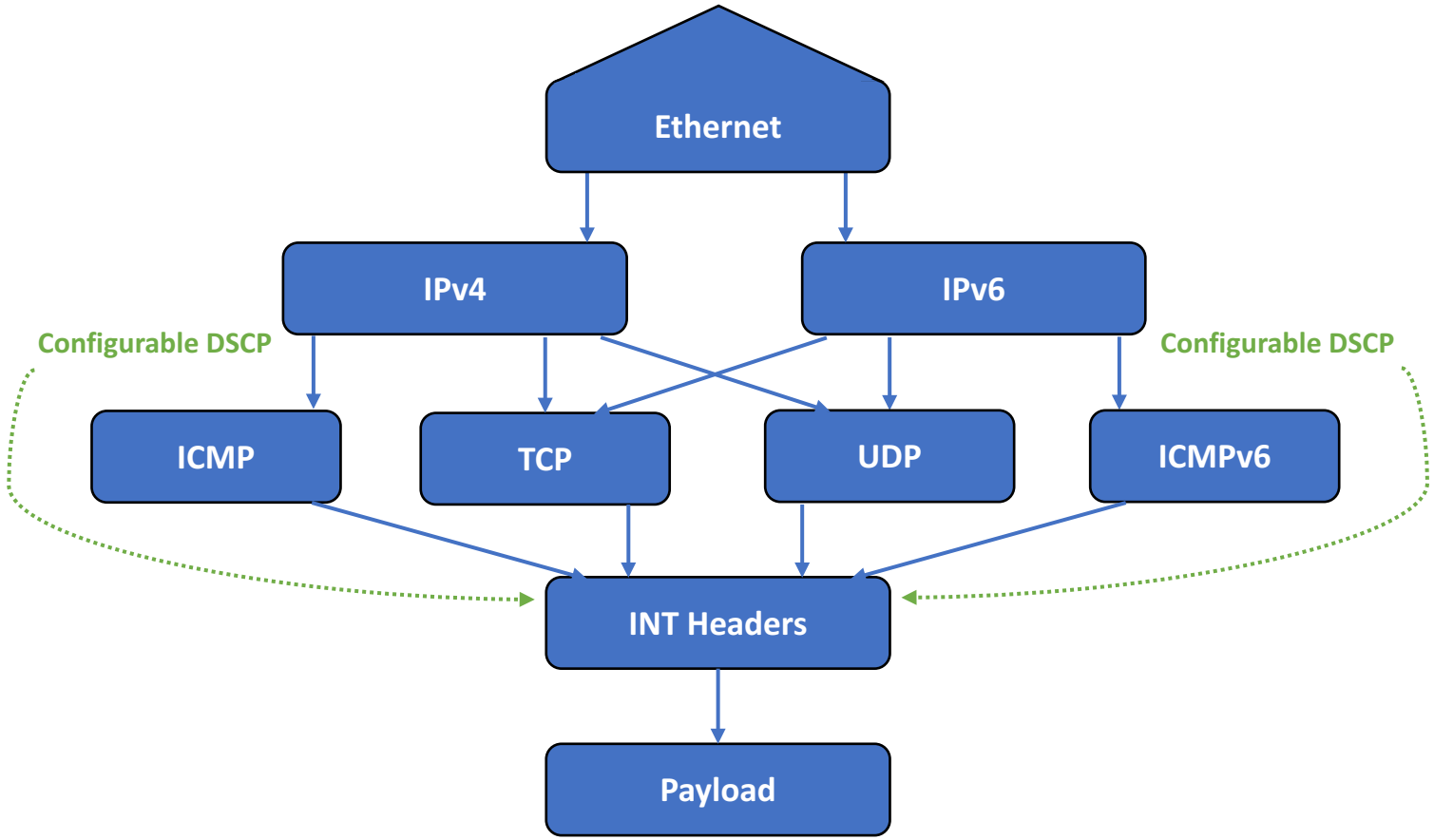
Simple Ethernet/IP Transport of Telemetry Report



Remaining Packet Format: Flat Packet

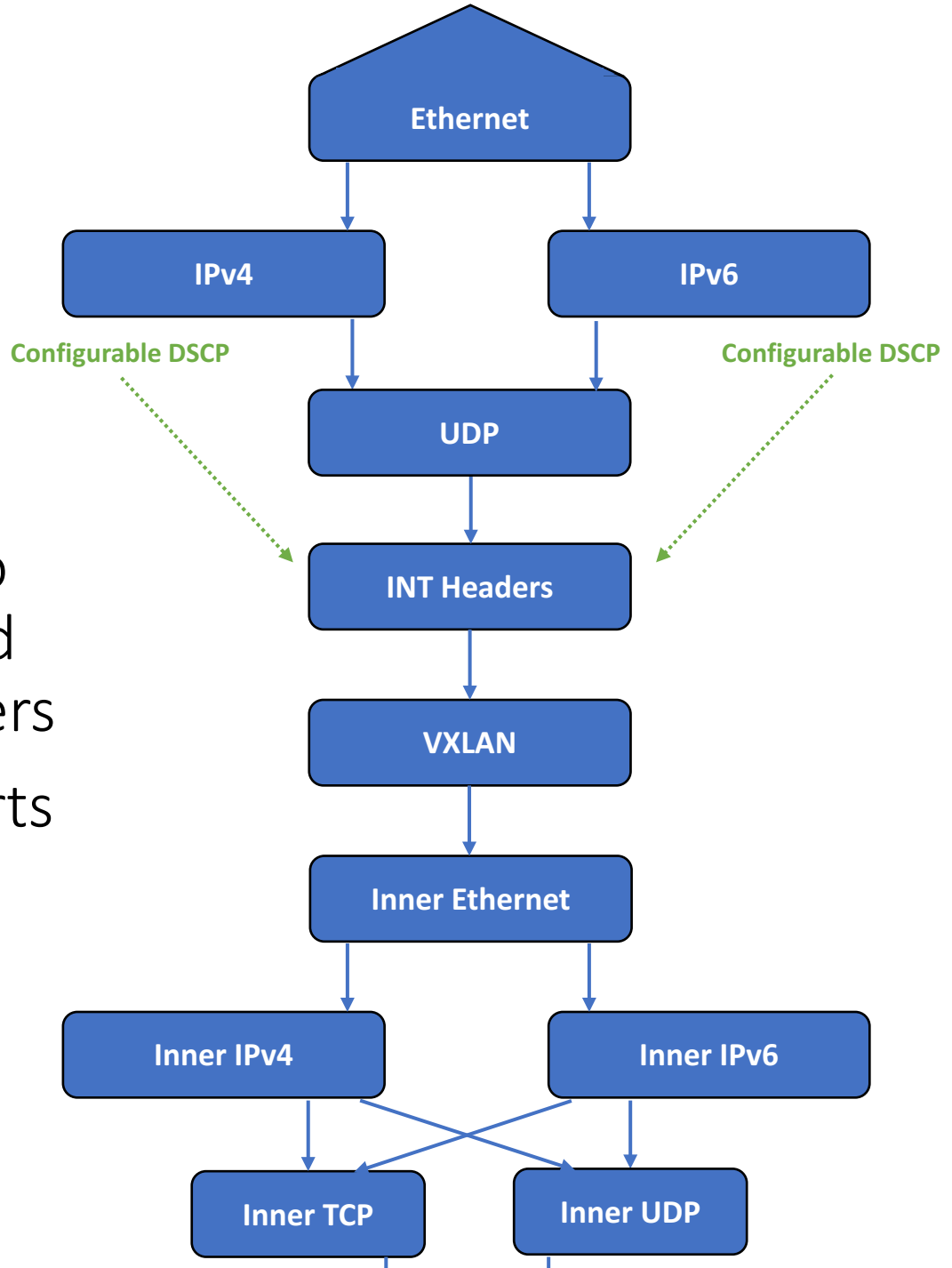


Remaining Packet Format: INT over TCP/UDP/ICMP



Remaining Packet Format: VXLAN with INT over TCP/UDP/ICMP

- Monitor may desire to categorize flows based on inner packet headers
- Drop and queue reports from INT transit switches may include embedded INT metadata



Remaining Packet Format: VXLAN with IOAM Trace

- Monitor may desire to categorize flows based on inner packet headers
- Drop and queue reports from IOAM transit switches may include embedded IOAM metadata