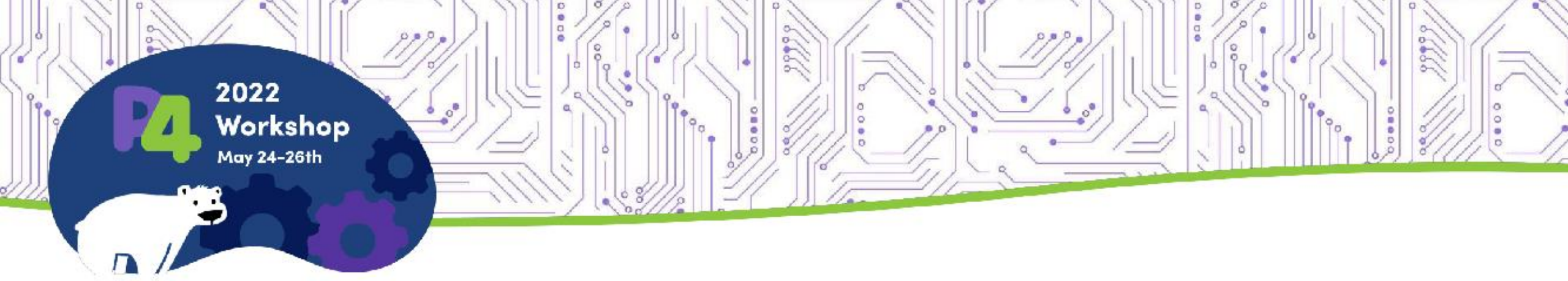


Inband Network Telemetry (INT): History, Impact and Future Direction

Jeongkeun “JK” Lee, Sr. Principal Engineer, Intel
Mukesh Hira, Principal Engineer, VMware

Former co-chairs of P4 Applications Working Group

Agenda

- Introduction and history
- Key Principles
- INT Modes
- Industry impact
- Direction: telemetry for control
- Summary

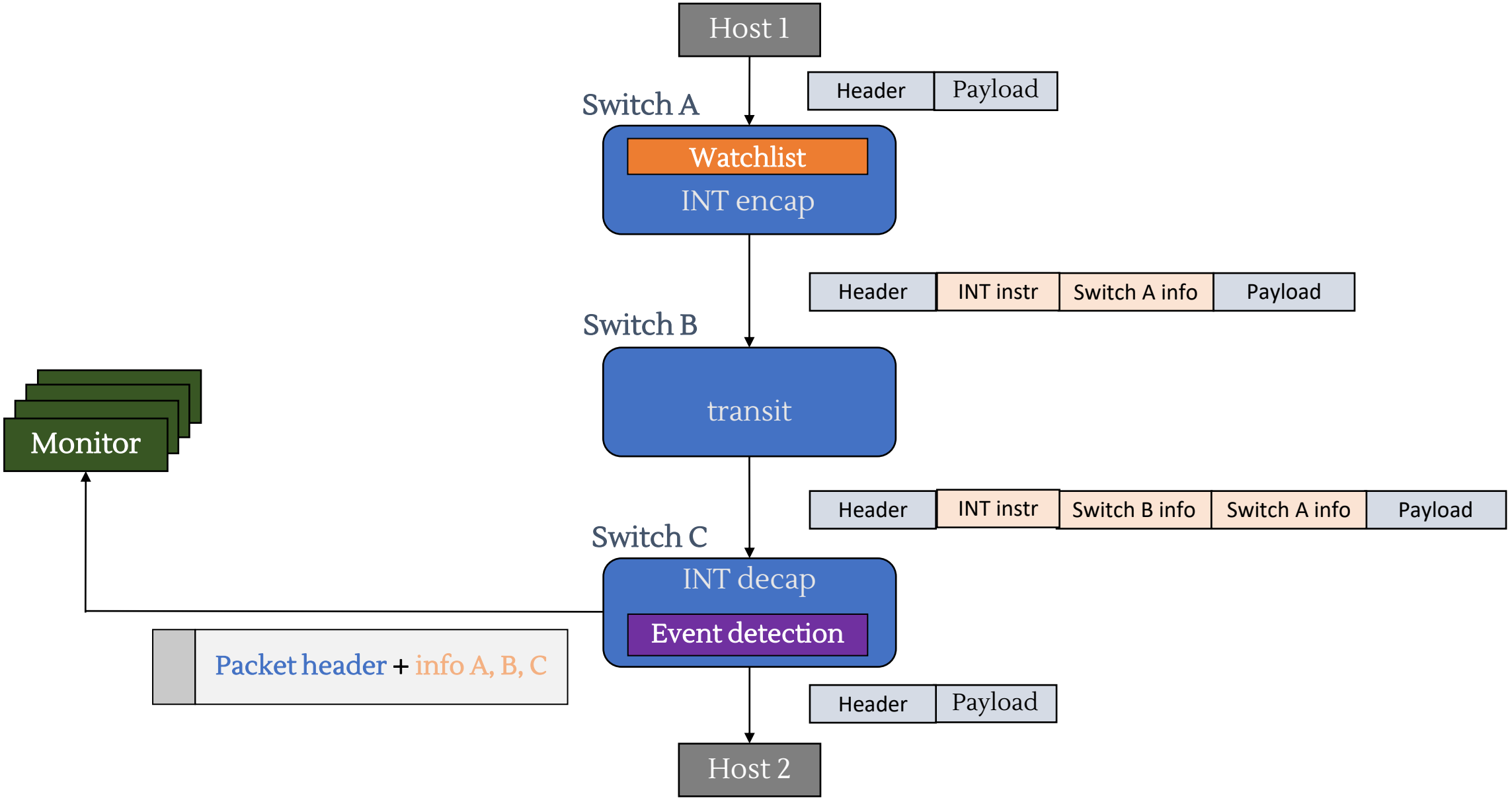
Introduction to INT

- Lack of fine-grained visibility into Networks has been a significant pain point for Network Operators

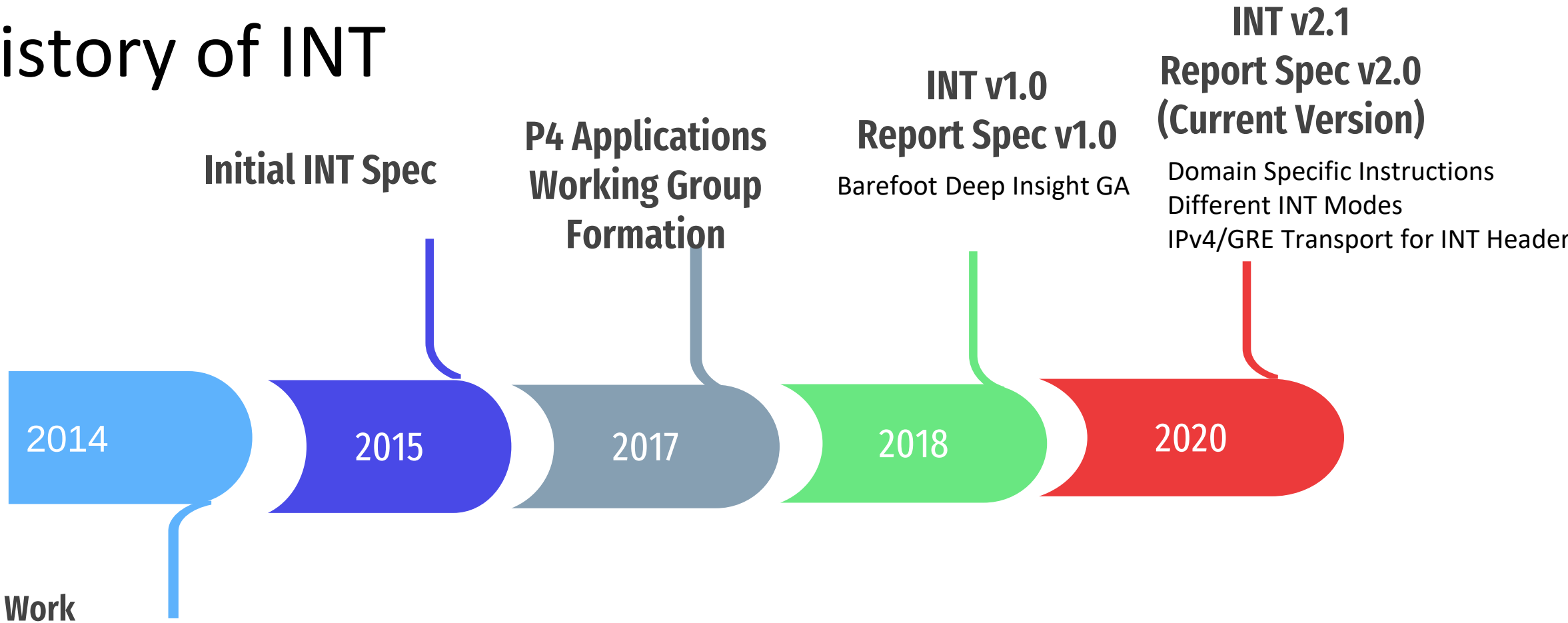
- Troubleshooting and monitoring Networks has been challenging
  - Root-causing transient problems is especially challenging
  - Troubleshooting tools have remained primitive for years: Ping, Traceroute, sFlow, Traffic Mirroring, Packet Capture

- INT changes the game
  - Fine-grained visibility on a per-packet basis
    - What path did a packet traverse?
    - What state did it experience at each hop
    - What other flows contributed to the state?

INT: Initial Concept



History of INT



- Millions of Little Minions: Using packets for low latency network programming and visibility, Sigcomm 2014.
- In-band Network Telemetry via Programmable Data Planes, Sigcomm 2015 Demo

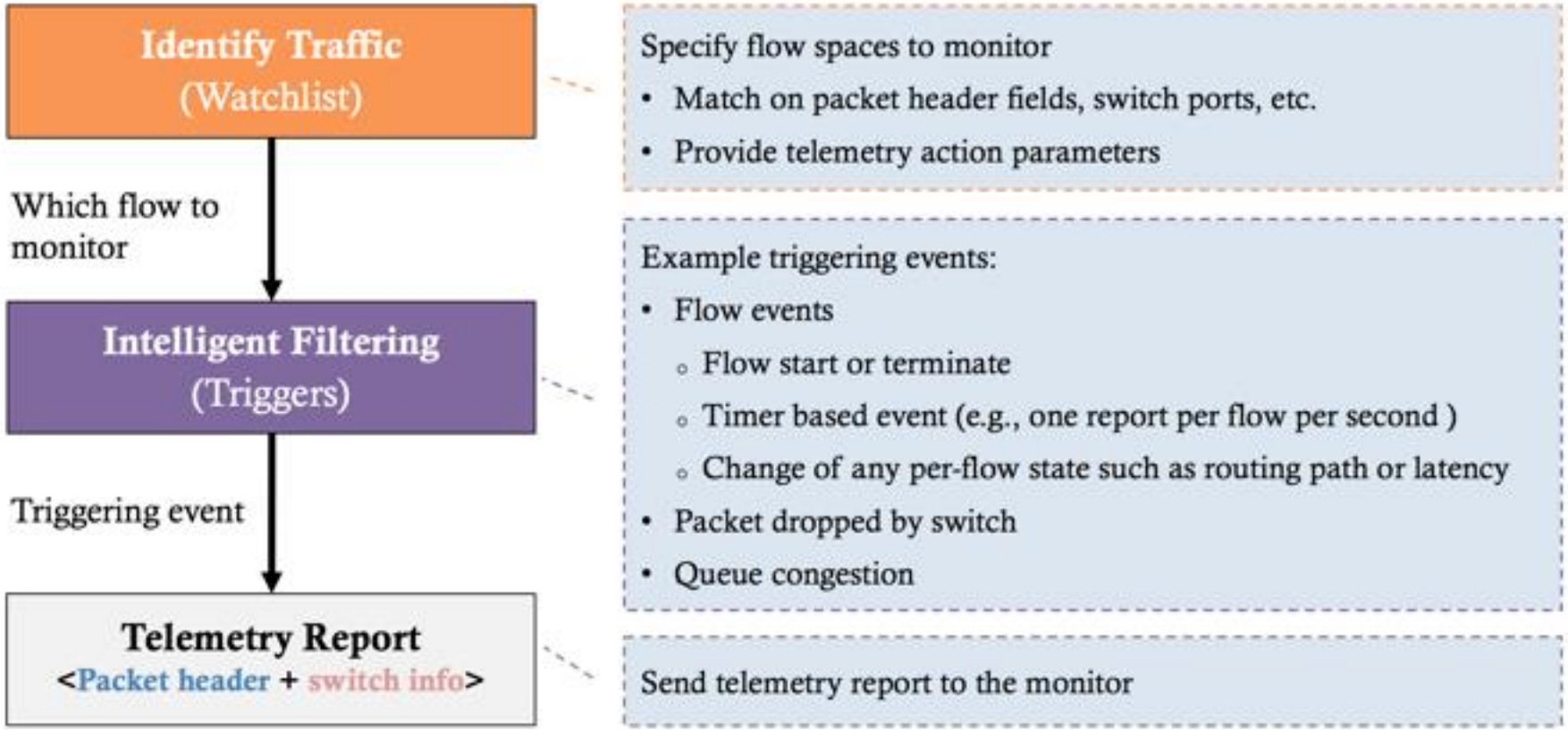
INT Development Principles

- Leverage ***Programmable*** P4-based Data Plane
- Release Features and Capabilities at Software Development Velocity
- Applications / End Points completely agnostic of INT underneath
- Flexibility is of utmost importance for rapid innovation and minimum barriers to adoption
  - Flexibility in INT Header Location: INT over TCP/UDP, VXLAN/Geneve, IPv4 GRE
  - Flexibility in Modes of Operation
  - Flexibility in Instruction Definition: Domain Specific Instructions
  - Flexible Report Format Definition
  - Flexible Metadata Semantics: YANG model based reporting of metadata to INT monitor

INT Impact

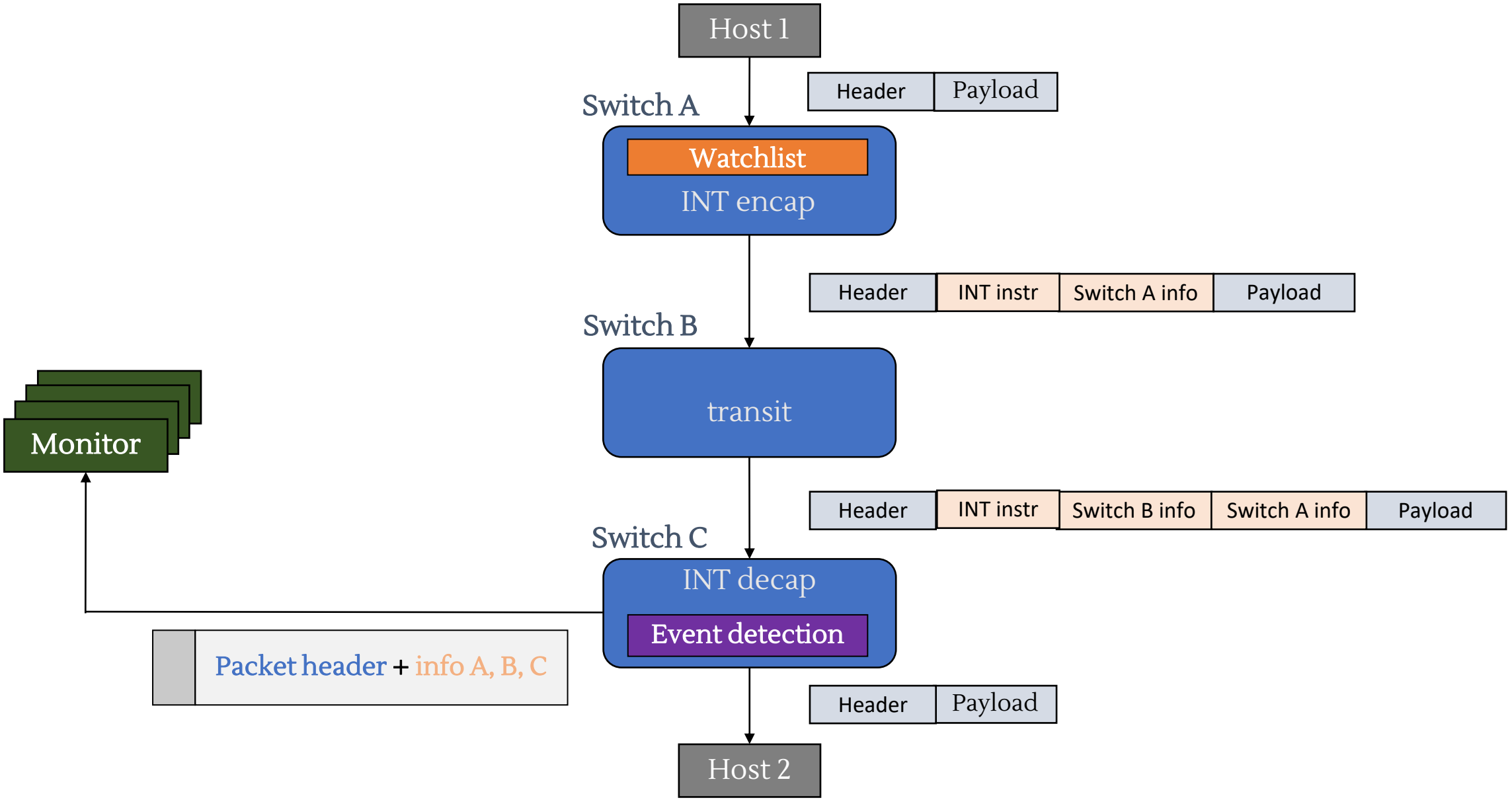
- Real-Time Fine-grained visibility is a game-changer and opens up a spectrum of use cases
  - Simplified Troubleshooting
  - Intelligent Path Selection: Choosing the best among Equal-Cost paths for optimal performance at flow / flowlet granularity
  - Congestion Control
  - Network Management and Capacity Planning → reduce OPEX

INT system aspects: logical functions and placement

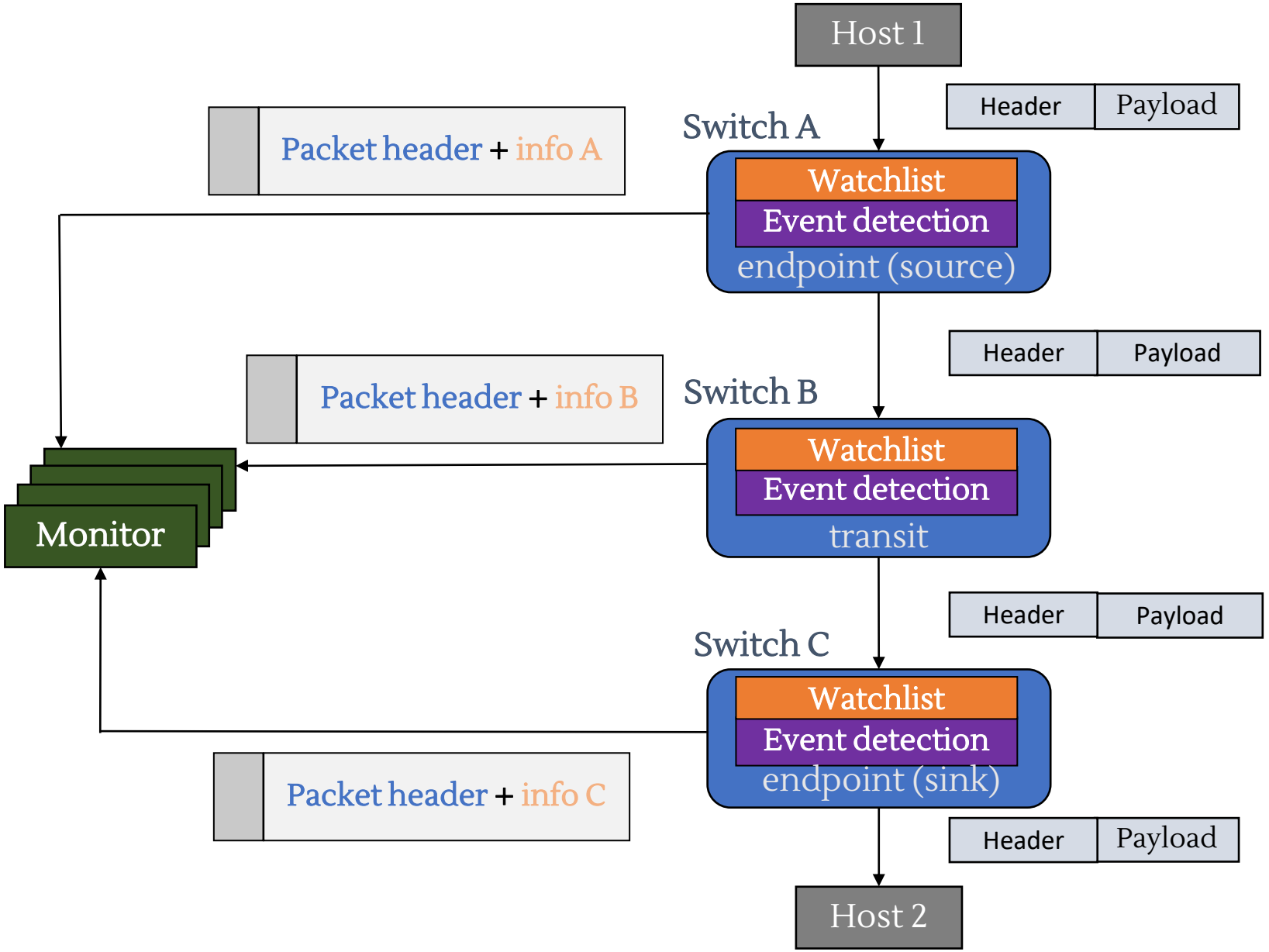


Three logical functions (above) can be placed at different network locations (**different INT modes in next slides**) to best meet the requirement and deployment constraints, such as live data packet modification.

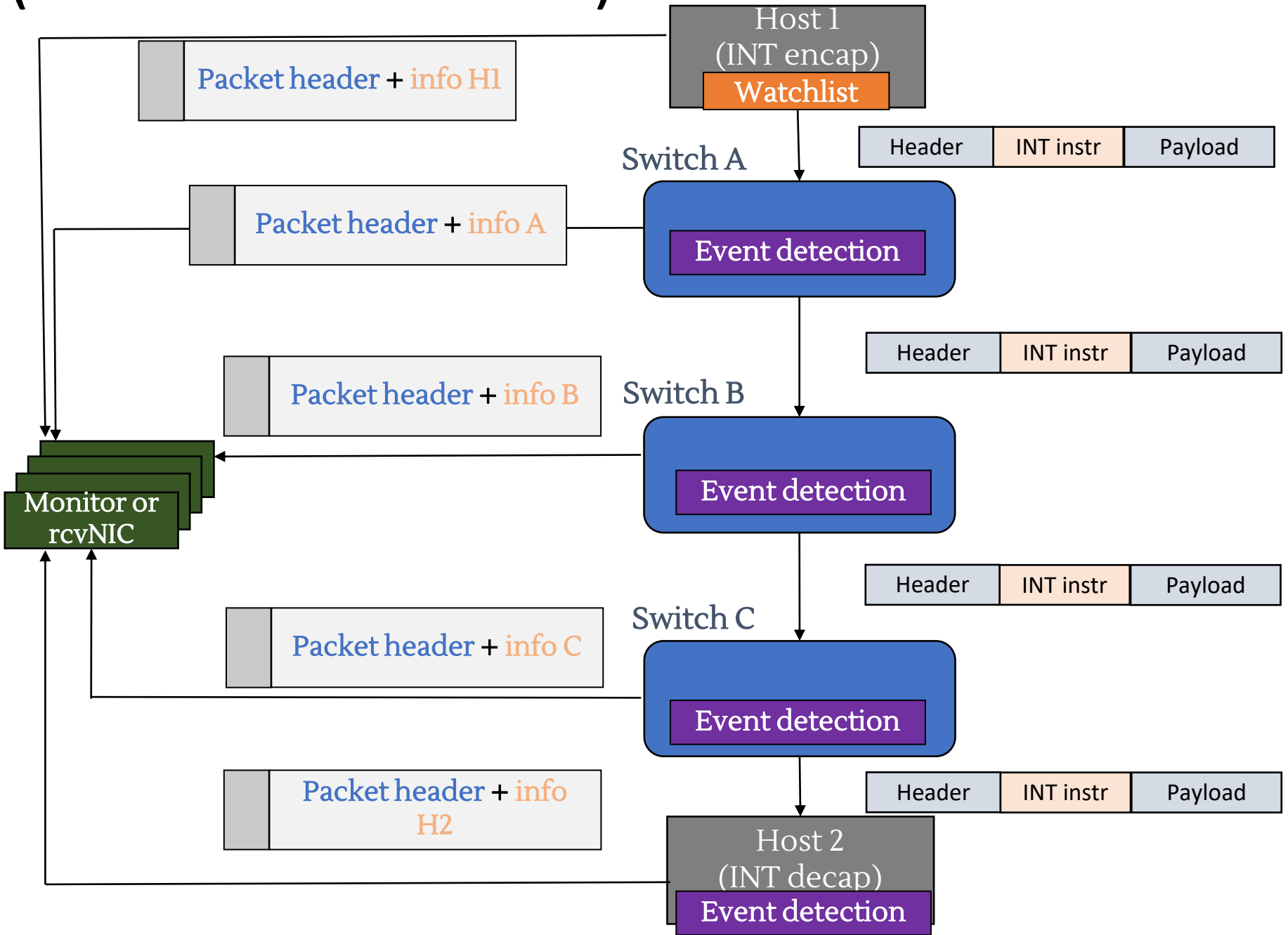
INT-MD (eMbed Data)



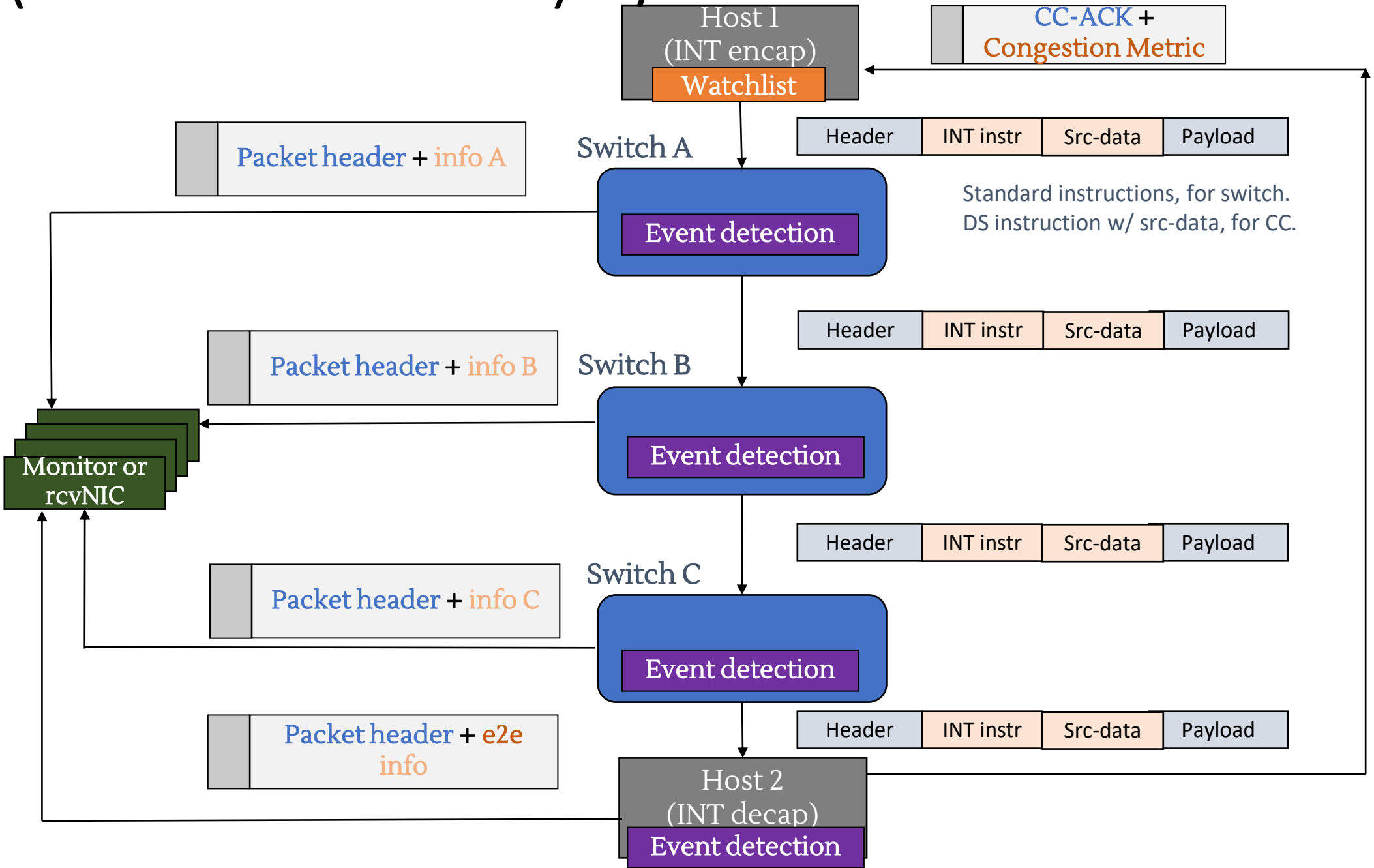
INT-XD (eXport Data) … aka Postcards



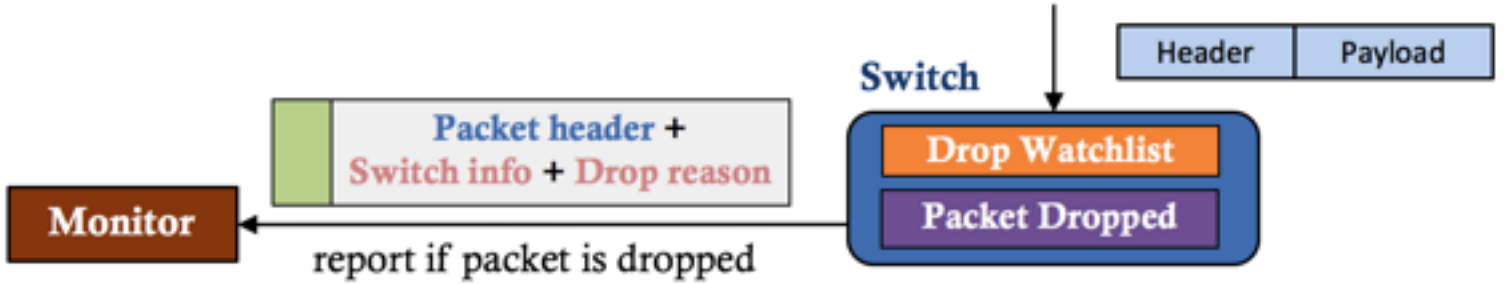
INT-MX (eMbed instruction)

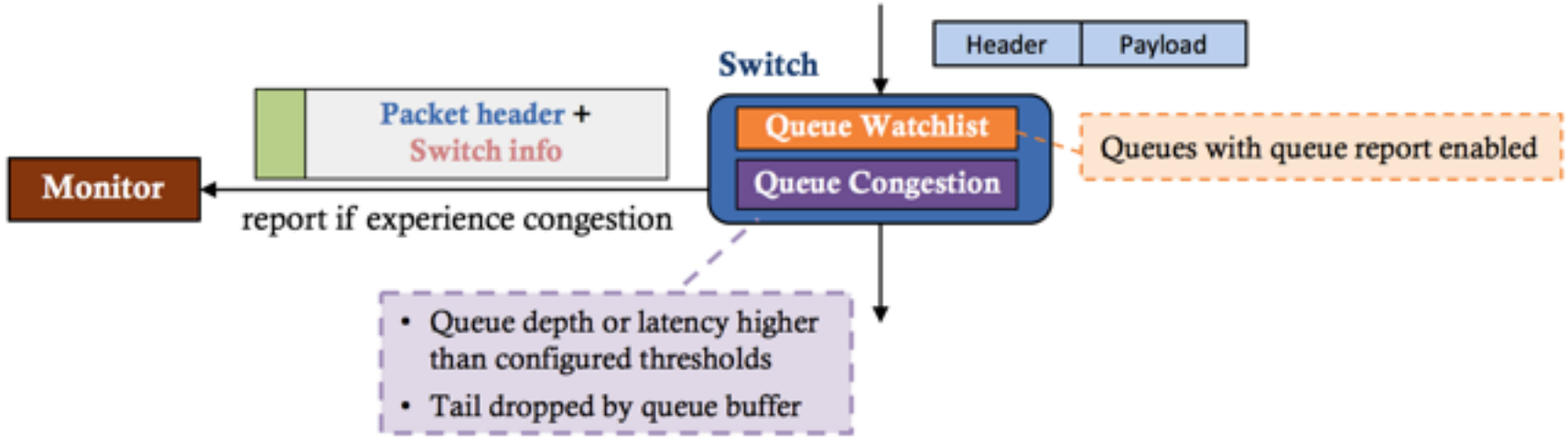


INT-MX (eMbed instruction) w/ src-data

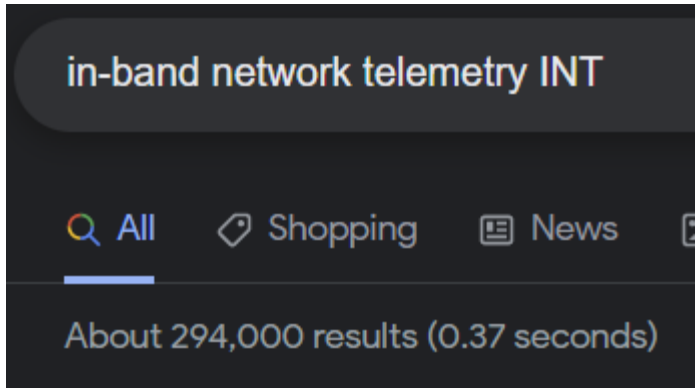


Drop report & Queue report
(local events, supported by Telemetry Report spec)



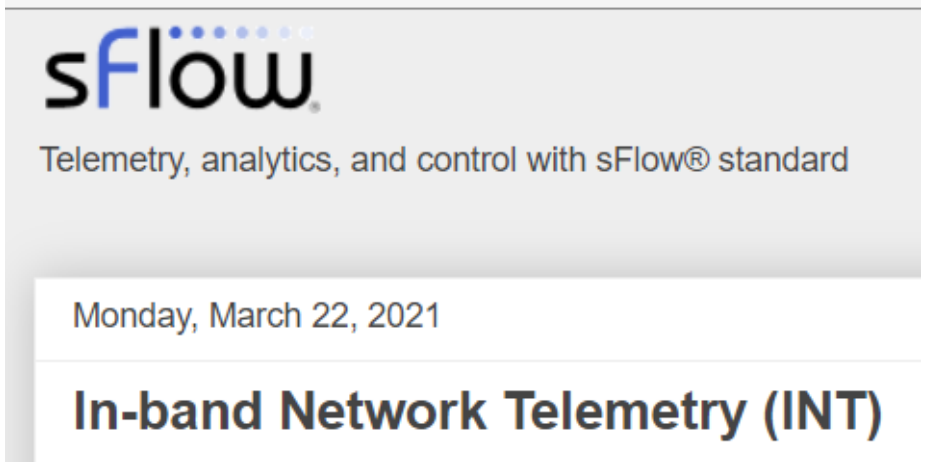


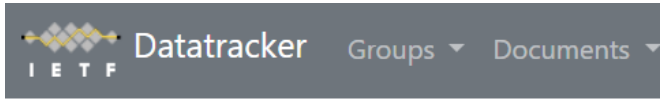
INT adoptions



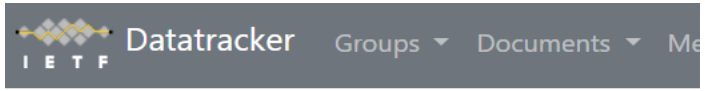
- HW/SW implementations in
  - Cisco
  - Xilinx
  - Arista
  - Intel
  - ONF
  - …

- Variants in standard forums





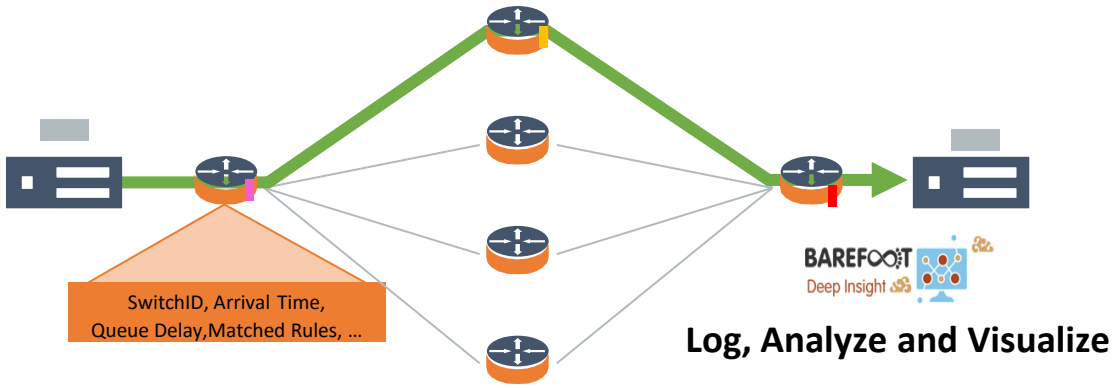
In-situ OAM (ioam)



Inband Flow Analyzer

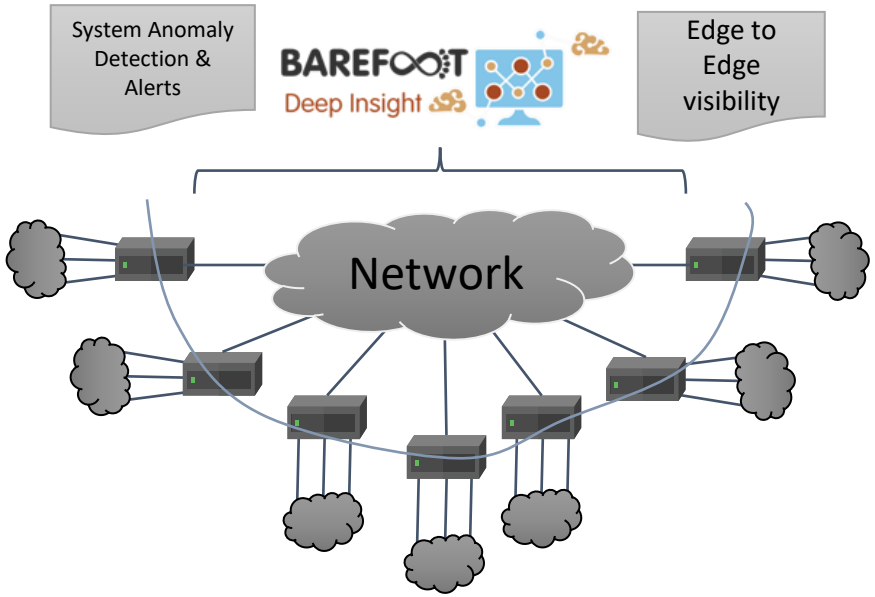
INT deployment options

Switch INT



- Telemetry about network fabric

Host INT



- Telemetry from network edges

https://github.com/intel/host-int

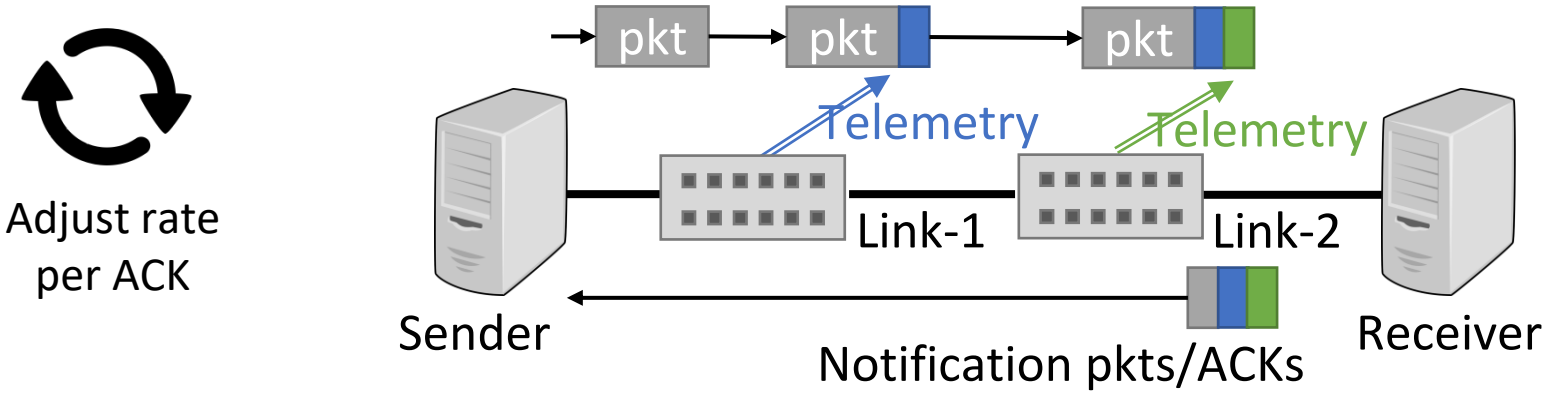
Direction: Telemetry for Control @scale

For control
- SDN control
- Transport, multi-pathing, congestion control

@Scale
- In terms of #flows and #reports
- Direct integration/correlation with application

HPCC: High Precision Congestion Control (Sigcomm'19)

- Use in-band telemetry as precise feedback and control



- HPCC++ @IETF

https://datatracker.ietf.org/doc/draft-miao-iccrg-hpccplus/

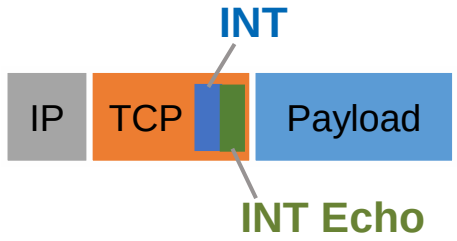
Authors:

R. Miao — *Alibaba Group*
H. Liu — *Alibaba Group*
R. Pan — *Intel Corporation*
J. Lee — *Intel Corporation*
C. Kim — *Intel Corporation*
B. Gafni — *Mellanox Technologies, Inc.*
Y. Shpigelman — *Mellanox Technologies, Inc.*
J. Tantsura — *Microsoft Corporation*

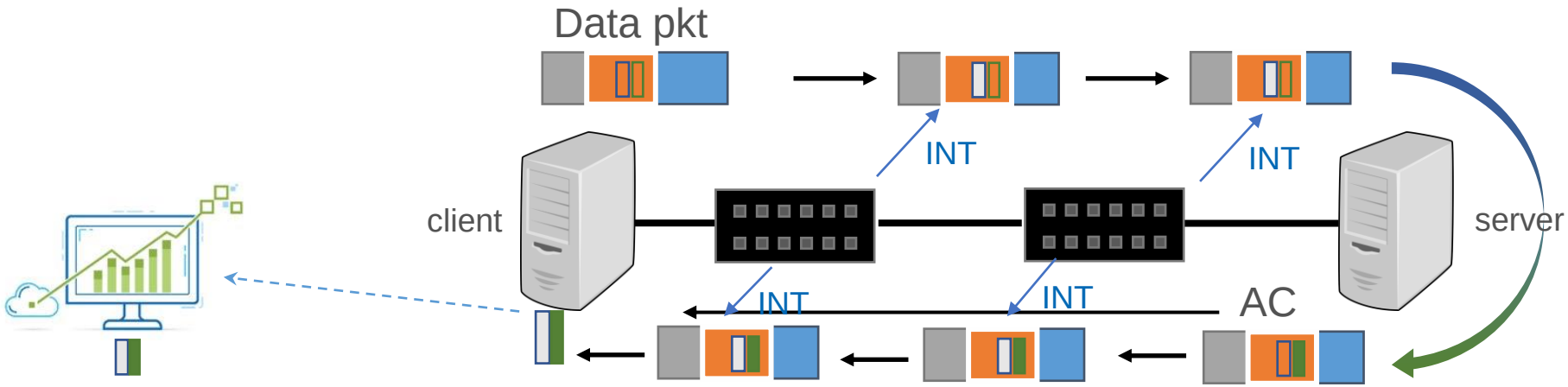
New types of INT for control & transport

- Problem: large stack overhead in header and RX Transport processing
- In-network INT stack reduction
  - min/max/sum ‘aggregate’ metrics
  - Constant size, fitting in INT-MX src-data or existing L2/L3/L4 header
- Back-To-Sender (BTS) telemetry
  - INT-XD (postcard) to sender, no change on data pkts
  - Fastest possible signaling of emergent congestion/failure events
  - Progress from previous P4 workshop talks
    - 2020: Advanced Congestion & Flow Control with Programmable Switches
    - 2021: Realizing One Big Switch Performance Abstraction using P4
    - Key is signaling from switch ingress (pre-queueing)

TCP-INT (in-network reduction)

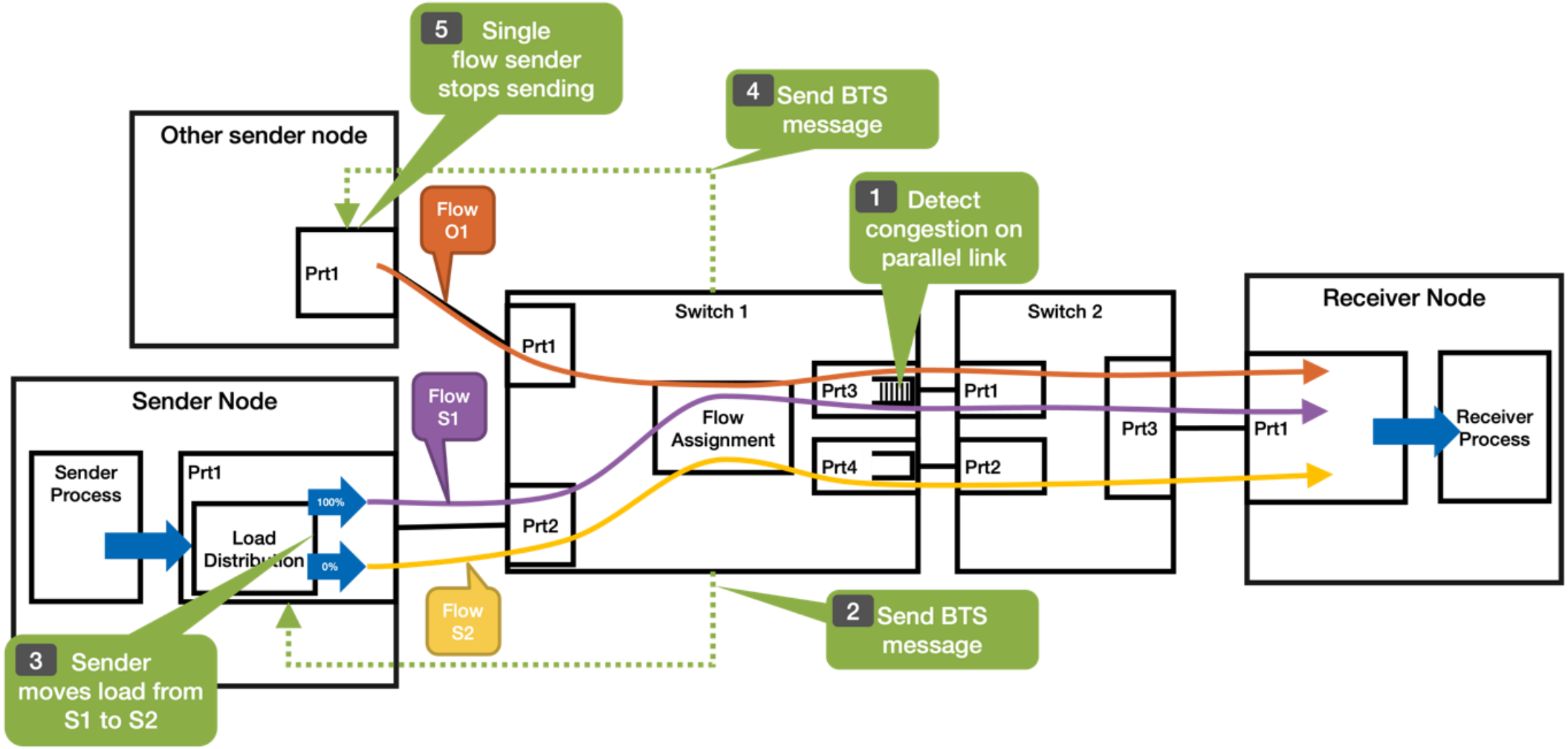


- INT stack reduced and embedded in TCP Options header
  - Works with NIC HW offloading of TCP checksum and segmentation
- Fabric telemetry (max qdepth, min BW) with TCP states (cWND, sRTT)



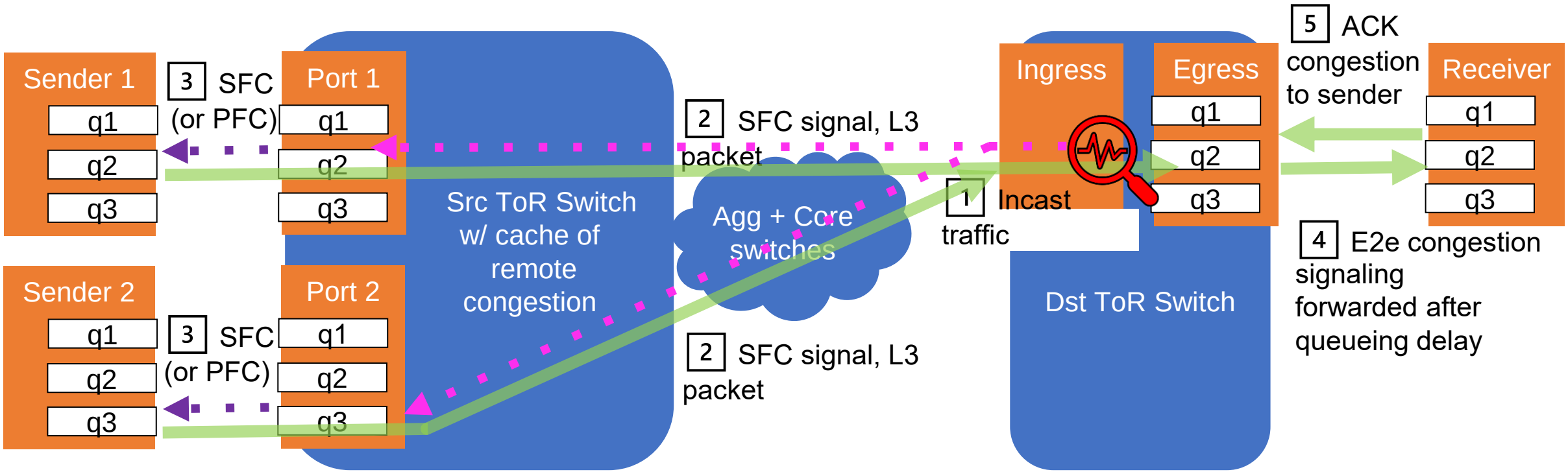
Demo @
P4 WS'22

Back-To-Sender (BTS) telemetry



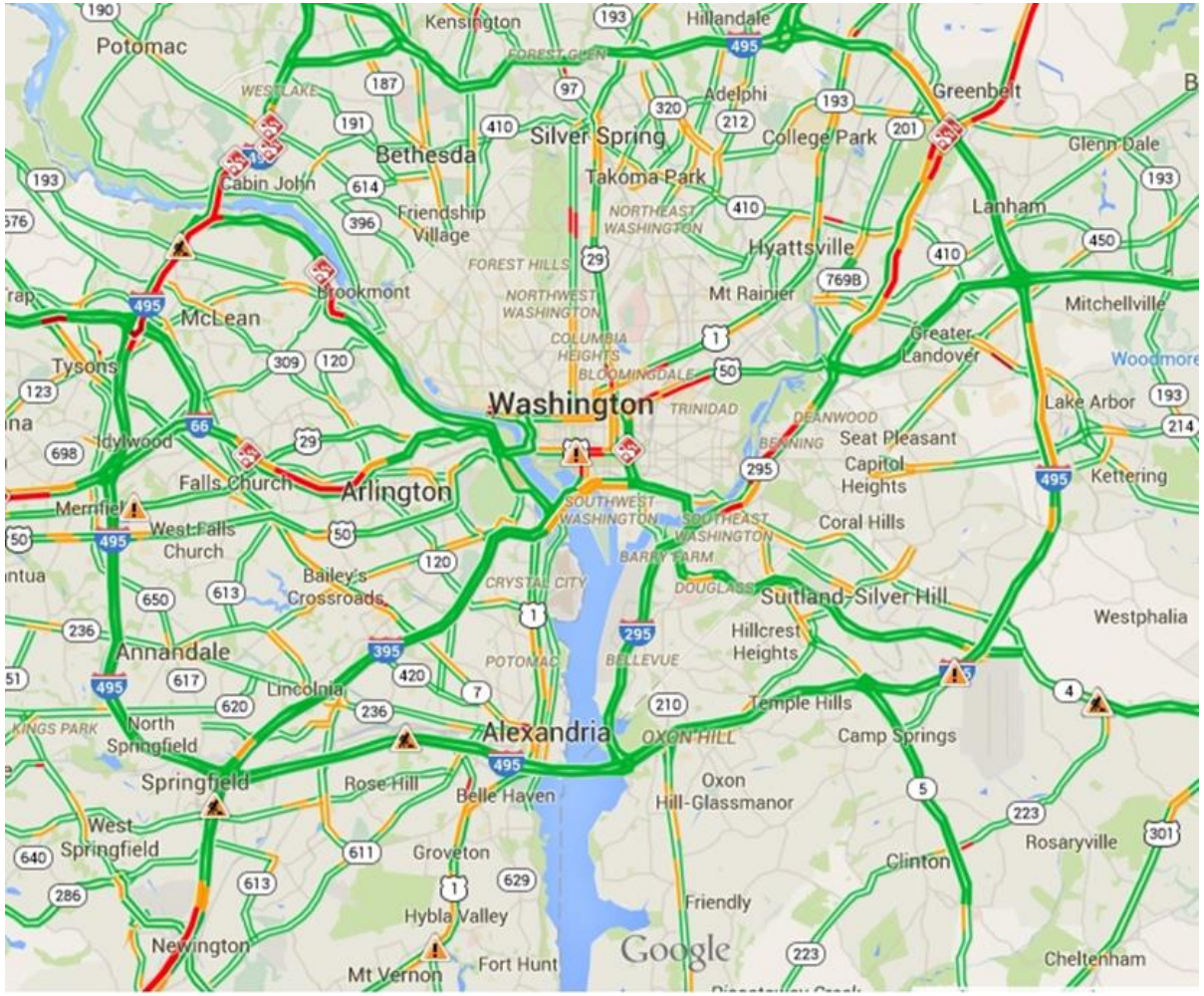
Use case of BTS: Source Flow Control @ IEEE

- Signaling of expected qdelay + flow control reaction within congestion-free sub-RTT
- Standardization process in IEEE 802.1
  - Motion for PAR/CSD approved: link to the material.



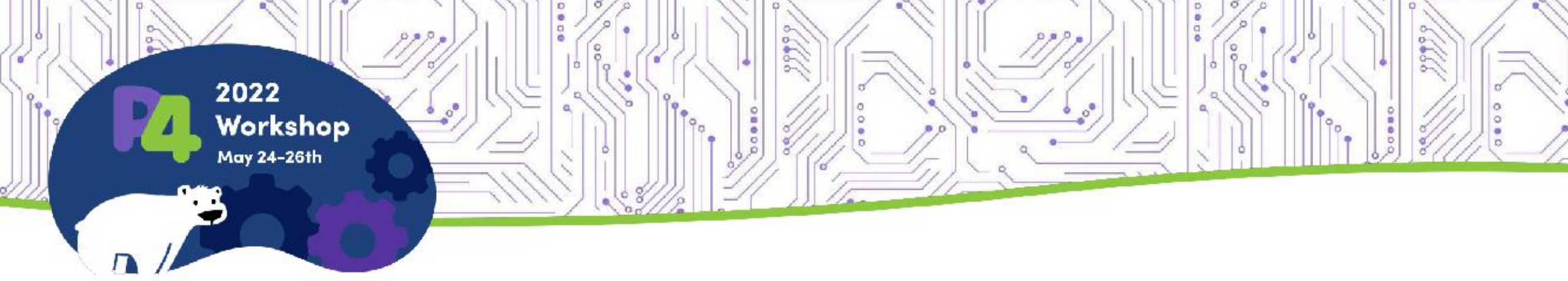
Summary

- INT: representative P4 app
  - Became industry term
  - Basic network building block
  - Many research work
- Towards the “Maps” service
  - Fast and precise telemetry
  - Push the network utilization and app latency to the theoretical optimal



“If you can’t measure it, you can’t improve it”

-- Peter Drucker



Thank You

This deck includes contributions from Changhoon Kim, Roberto Mari, Mickey Spiegel, and Jeremias Blendin.

Greatly appreciate P4 Apps WG members who directly, indirectly
contributed to the specs of INT and Telemetry Report Format.