tnc21 Making Waves
21-25 June 2021

GÉANT

In-band Network Telemetry at AmLight

Lessons Learned after two years playing with INT

Jeronimo Bezerra & Arturo Quintana

TNC 2021

June 23rd, 2021

Outline

- What is In-band Network Telemetry?
- How are we using INT?
- Use Case
- Moving to production

In-band Network Telemetry (INT)

- INT is a P4 application that records network telemetry information in the packet while the packet traverses a path between two points in the network
- As telemetry is exported directly from the Data Plane, Control Plane is not affected:
  - Translating: *you can track/monitor/evaluate EVERY single packet at line rate and real time.*
- Examples of telemetry information added:
  - Timestamp, ingress port, egress port, queue buffer utilization, sequence #, and many others

INT: How does it work?

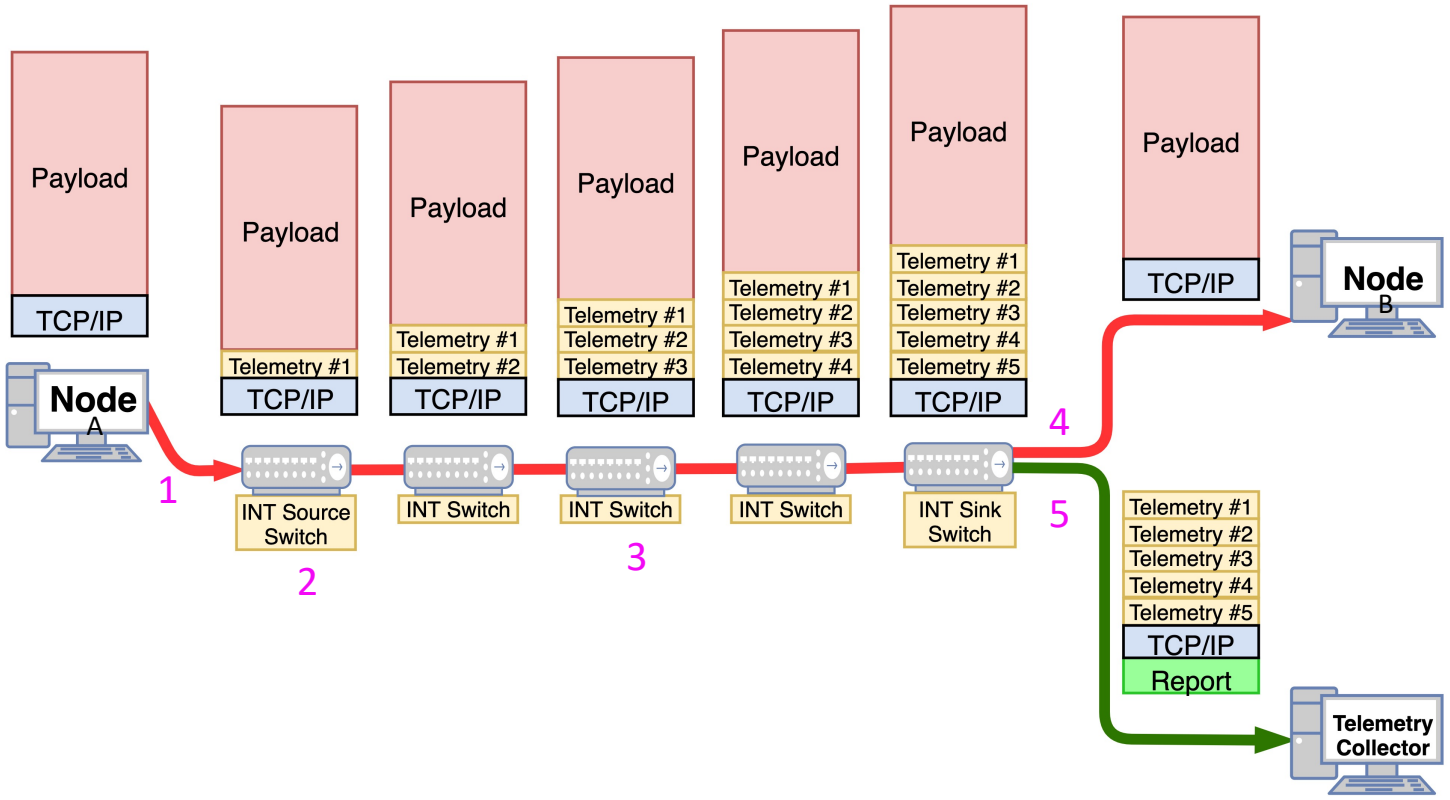
1 – User sends a TCP or UDP packet unaware of INT

2 – First switch (INT Source Switch) pushes an INT header + metadata

3 – Every INT switch pushes its metadata. Non-INT switches just ignore INT content

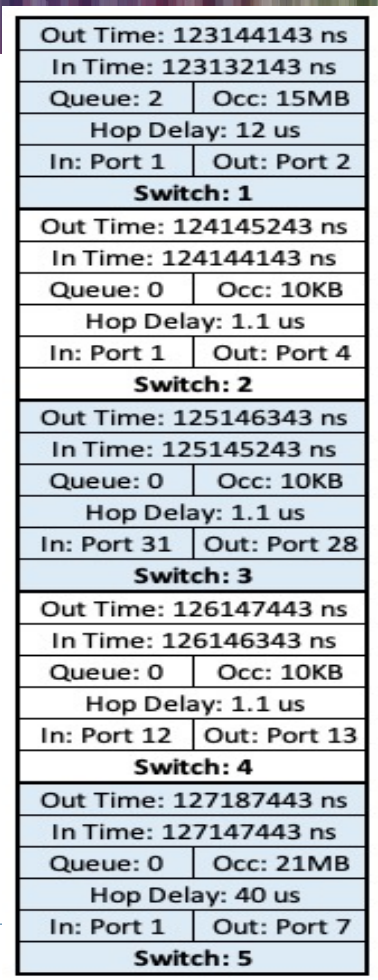
4 – Last switch (INT Sink Switch) extracts the telemetry and forwards original packet to destination

5 – Last switch (INT Sink Switch) forwards the 1:1 telemetry report to the Telemetry Collector



What INT metadata is being used and how? [1]

- The AmLight INT switches leverage the Tofino chip to collect:
  - Per switch:
    - Switch ID
    - Ingress port
    - Egress port
    - Ingress timestamp
    - Egress timestamp
    - Egress queue ID
    - Egress queue occupancy
  - Per telemetry report:
    - Report timestamp
    - Report sequence number
    - Original TCP/IP headers



What INT metadata is being used and how? [2]

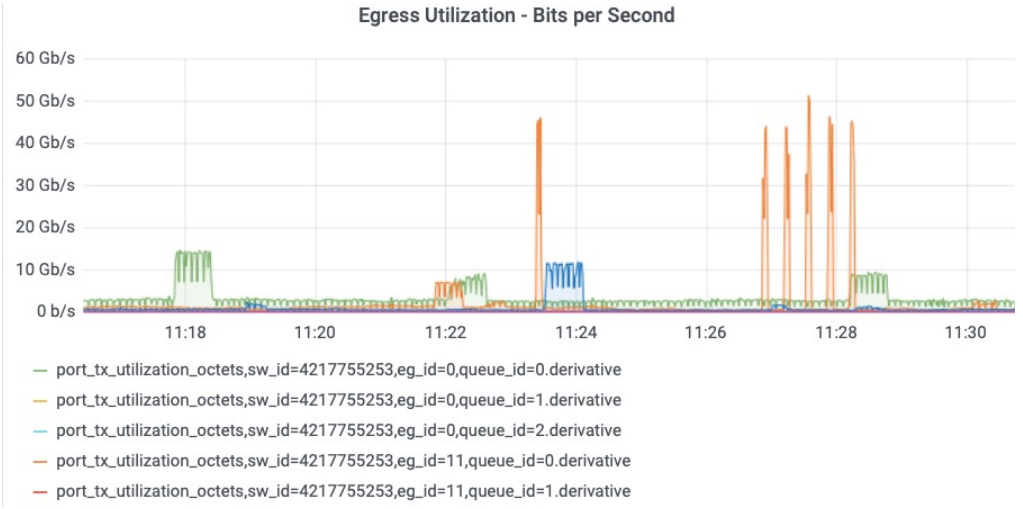
- **Proof of Transit (PoF)** or path taken (L1 traceroute)
  - Metadata used:
    - List of switches
    - Per switch:
      - Switch ID, Ingress port, Egress port, Egress queue ID

- AmLight is capable of path tracing EVERY packet and recording changes
  - Useful for detecting LAG or ECMP hash errors/mismatches
  - Useful for detecting unstable links

- Path taken even indicates *egress queue ID*:
  - Useful for evaluating QoS policies

What INT metadata is being used and how? [3]

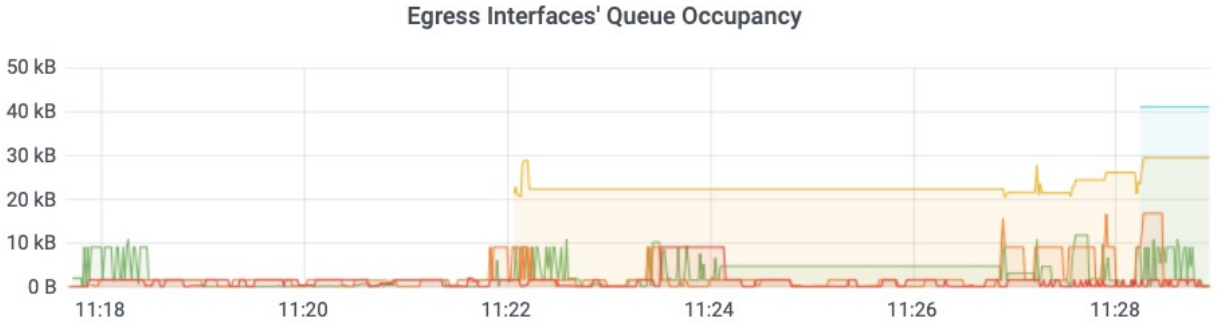
- Instantaneous Ingress and Egress Interface utilization
  - Metadata used:
    - List of switches
    - Per switch:
      - Switch ID, Ingress port, Egress port
    - From the user TCP/IP header:
      - IP length
- Telemetry collector monitors and reports egress interface utilization every 100ms*
  - Useful for detecting microbursts
  - 100ms can be tuned down if needed
  - Bandwidth monitored per interface & queue



What INT metadata is being used and how? [4]

- Instantaneous Egress Interface Queue utilization (or buffer)
  - Metadata needed:
    - List of switches and their metadata
    - Per switch:
      - Switch ID, Egress port, Egress Queue ID, Queue Occupancy

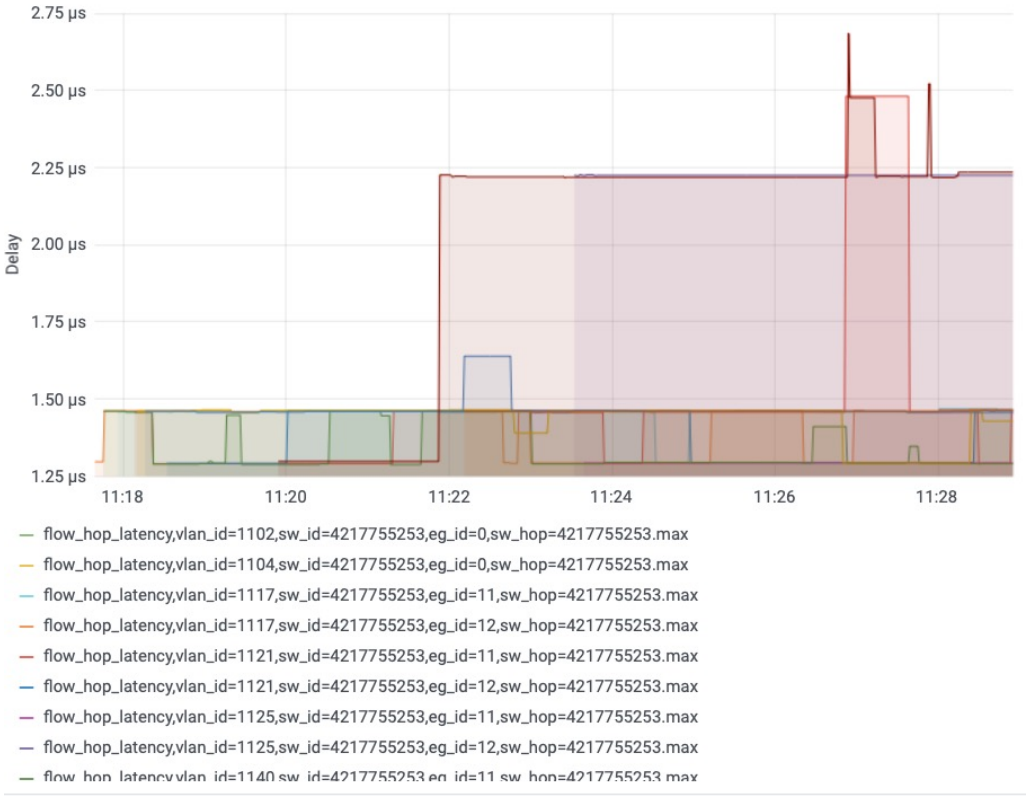
- AmLight monitors every queue of every interface of every switch:
  - Useful for evaluating QoS policies
  - Useful for detecting sources of packet drops



What INT metadata is being used and how? [5]

- Sources of jitter:
  - Metadata needed:
    - List of switches
    - Per switch:
      - Switch ID, ingress timestamp, egress timestamp

- AmLight monitors per-hop per-packet forwarding delay:
  - Useful for evaluating sources of jitter along the path
  - Useful for mitigating QoS policy issues (under provisioned buffers)
  - Useful for mitigating traffic engineering issues (under and over provisioned links)

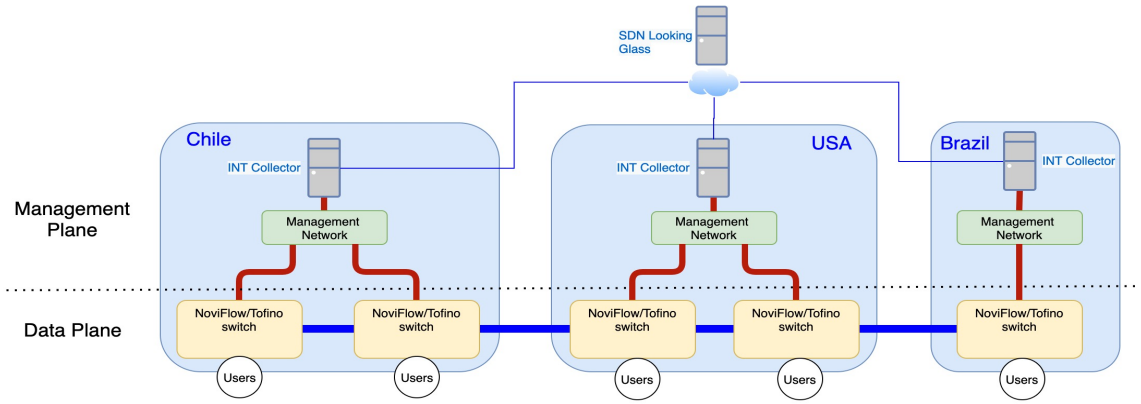


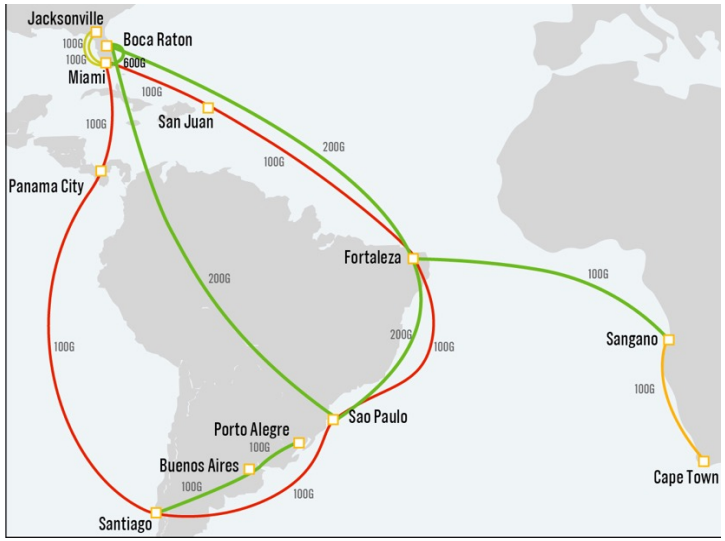
Use Case: Mitigating [malicious] [micro] bursts

- 5 data transfers/bursts of 40-50Gbps for 5 seconds.
- Top: INT metadata exported in real time, per packet
- Bottom: SNMP get running as fast as supported by the switch: 15 seconds.
- *By leveraging legacy technologies, such as SNMP, troubleshooting microbursts – malicious or not – is a complex activity that won't be enough to characterize the microburst and determine its impact.*

Ongoing Deployment at AmLight [1]

- At each AmLight site, NoviFlow/Tofino switches are being deployed
  - These switches are the new AmLight Data Plane
- Each pop has an INT Collector parsing Gbps of telemetry
- InfluxDB & Grafana combo to store and display reports
- Goal is for AmLight to be fully INT-capable by Q2/2022.





Ongoing Deployment at AmLight [2]

- *How many high-priority flows can be handled in real-time by the INT Telemetry Collector?*
  - *Using eBPF/XDP for processing telemetry data*
  - *Currently capable of "processing" 10Mpps**
- *What is the impact caused by INT in a complex network such as AmLight?*
  - *Delay: Pushing INT header takes around 0.00045 ms. No impact in a long-haul network.*
  - *MTU: Each switch adds 24bytes. Tofino chip has MTU of 10K. Legacy devices with shorter MTU in the path have to be considered.*
  - *Colocation: Every AmLight PoP needs colocation for the INT Telemetry Collector (telemetry is processed locally)*
- *How to dynamically enable INT monitoring of specific flows?*
  - *New OpenFlow 1.3 Experimenter Actions created (push_int, add_int_metadata, pop_int, send_report)*
  - *Enables AmLight to be very specific when selecting what to monitor (per-source, per-destination, TCP and UDP, per port, etc.)*

Conclusions

- Monitoring every and any packet is possible with in-band network telemetry!
- INT has increased the network visibility beyond our expectations
- Combining INT and legacy monitoring tools will enable AmLight to track any performance issue and user complain
- Combining INT with learning tools will enable AmLight to create reliable trends and move towards a closed-loop orchestration SDN network.

tnc21 Making Waves
21-25 June 2021

GÉANT

Thank you

Any Questions?

Jeronimo Bezerra jab@amlight.net
Arturo Quintana aquintana@fiu.edu

As part of the GÉANT 2020 Framework Partnership Agreement (FPA), the project receives funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 856726 (GN4-3).