AmLight ExP
Americas Lightpaths Express & Protect

Telemetry and Big Data Workshop
Nov 11th, 2020

[In-band] Network Telemetry at AmLight

Jeronimo Bezerra - IT Associate Director/FIU

Arturo Quintana - Sr. Software Developer/FIU

In-band Network Telemetry (INT) [1]

- INT adds metadata to each packet with information that could be used later for troubleshooting activities.

- As metadata is exported directly from the Data Plane, Control Plane is not affected:

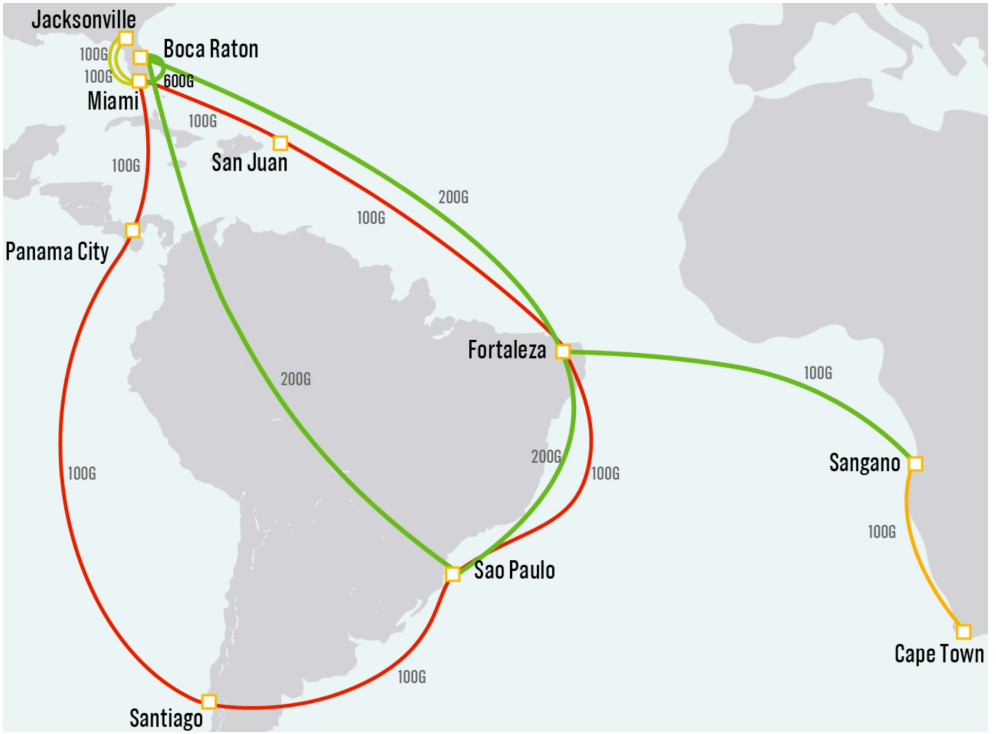
  Translating: *you can track/monitor/evaluate EVERY single packet at line rate.*

- INT metadata being used at AmLight:

  ingress port id, egress port id, ingress timestamp, egress timestamp, queue id, queue occupancy

Introduction to AmLight

- AmLight Express and Protect (AmLight-ExP) (NSF International Research Network Connections (IRNC) Award #1451018) [2]
- 680Gbps of upstream capacity between the U.S. and Latin America and 100Gbps to Africa
- Production SDN Infrastructure since 2014
- NAPs: Florida(2), Brazil(2), Chile, Puerto Rico, Panama, and South Africa
- Main motivation for deploying INT: The Vera Rubin Observatory's SLA

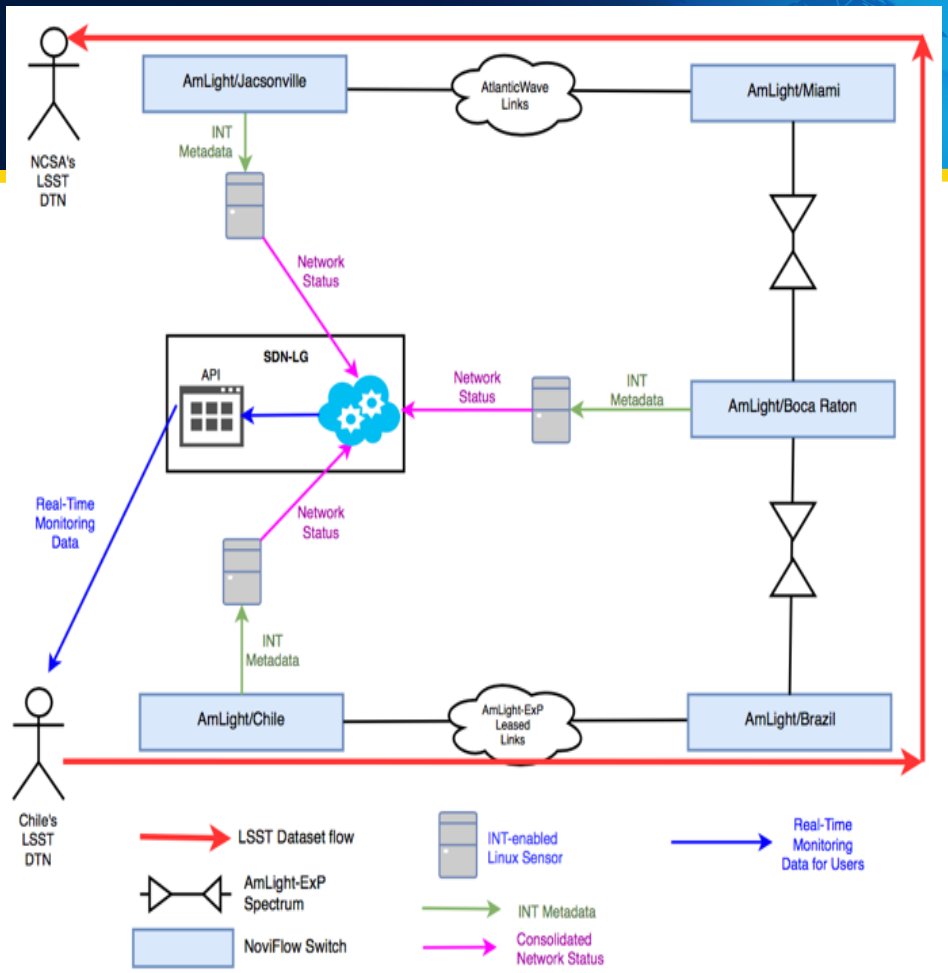


Telemetry at AmLight: Vera Rubin Use Case

- Supporting the Vera Rubin's requirements:
  - Every 27 seconds throughout the night, the telescope at the Base (Chile) site will take a 6.4GB picture of the sky, process it, generate transient alerts (6.3GB) from this picture, and send the 12.7GB data-set to the Archive site in the U.S.
    - From the 27-seconds window, only 5 seconds are available for data transmission [3]
  - Multi traffic types with different priorities (db sync, control, general internet traffic)
  - Full network visibility is required to mitigate issues in real time.

AmLight-INT Project

- NSF Award# OAC-1848746
- Goals:
  - Deploy P4/INT-capable switches
  - Deploy INT Collectors (100G hosts) to collect metadata
  - Develop a new methodology to collect and export INT data in real time to feed SDN controllers and users with monitoring information
  - Create a Network Telemetry Design Pattern to be used by other R&E networks

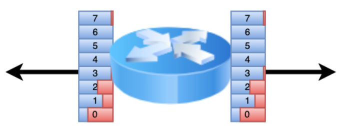


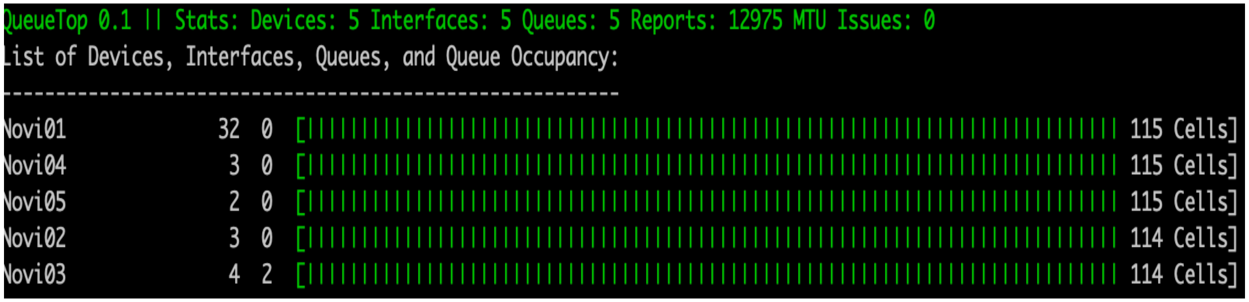
What can we do with INT at AmLight?

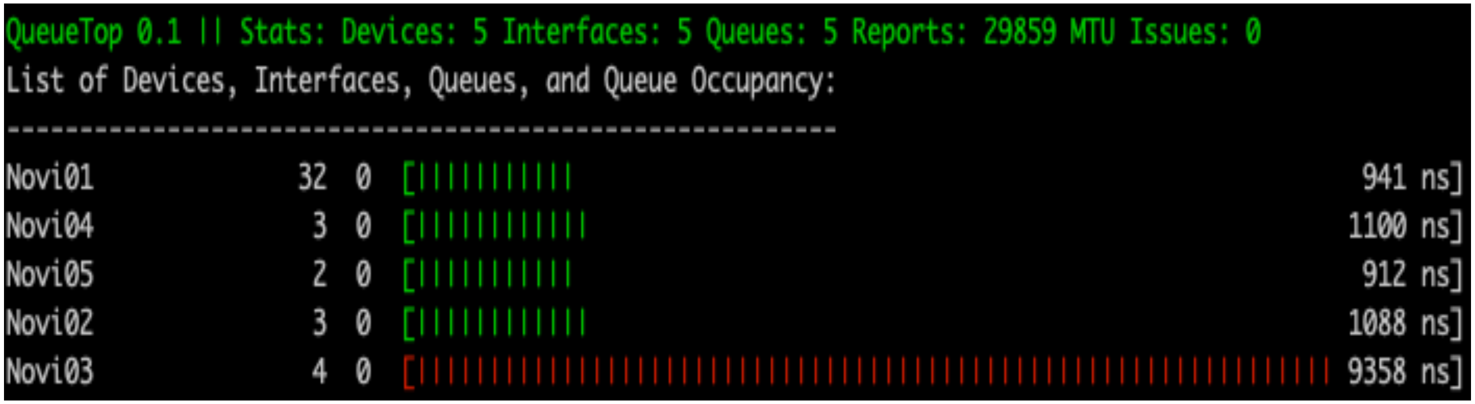
- Selectively enable network telemetry for specific flows:
  - Saves resources
- Track path taken in real-time, including egress queues
  - “Proof-of-transit”: make sure packets are following the forwarding policies
- Track path changes (including LAG port members):
  - Ideal to track link flaps and issues with LAG ports
- Monitor the delay introduced by each network hop in the path
  - Ideal to track per interface and queue high utilization (mitigating micro-bursts)
- Monitor all interfaces’ queue’s being used
  - Ideal to evaluate the QoS policy in place
- Correlate user performance observed with network conditions
  - With user’s TCP/IP headers and INT, we can pinpoint why the performance bottleneck is as is

QueueTop – Queue Occupancy Monitor

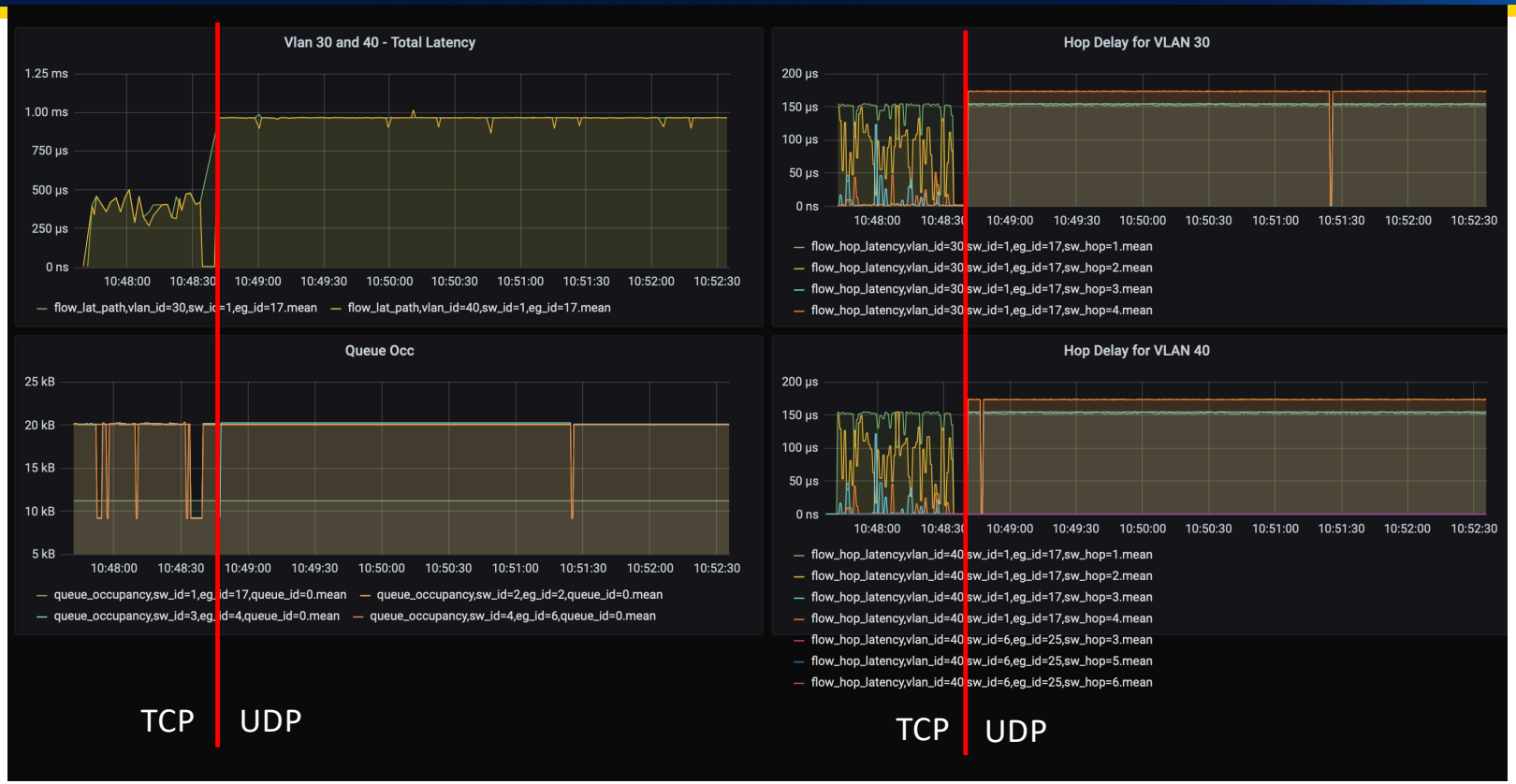
- Monitors all switches' interfaces' queues in real time







QueueTop + InfluxDB + Grafana



Challenge 1: Receiving telemetry reports

- 100Gbps with 9000-Bytes packets ➔ ~1.5M packets per second
- At AmLight 6-10 switches connect Chile to the U.S.
- Telemetry reports have 208-220bytes
- Each user packet creates a telemetry report
- 4.5Gbps of telemetry report for each 100Gbps of traffic
  - Single flow/No hashing
- Solution in place: eBPF/XDP (eXpress Data Path) [4]

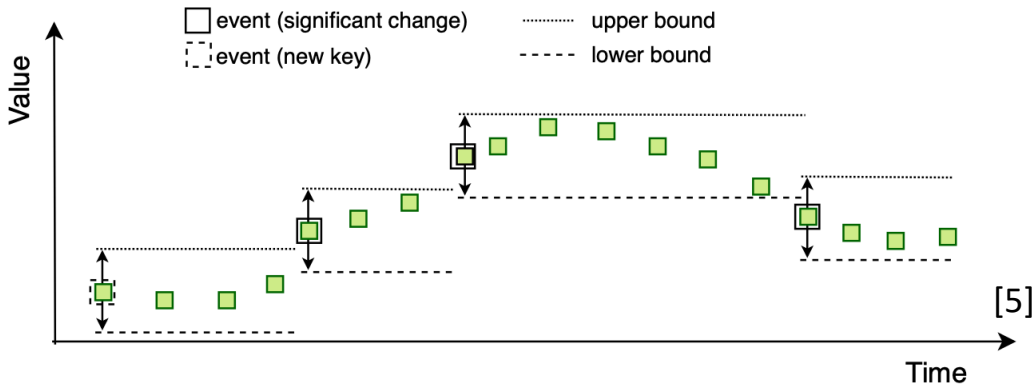
Challenge 2: Storing telemetry reports of interest

- Not feasible to save all telemetry reports (yet)

- Solution: XDP code only “stores” counters that report a change in the traffic behaviour:
  - A queue that increased/decreased more than 20KBytes
  - A flow path that changed
  - A hop delay or total delay that increased above 2 microseconds
  - An egress interface that is using more than 80Gbps

- This data is stored using InfluxDB

- Result:
  - Close to real time processing.
  - Not so granular measurements



[5]

Challenge 3: Storing all telemetry reports for future research (ML/AI)

- Goal: *Store as many telemetry reports as possible for future research to enable ML/AI researchers to have grounding truth for learning algorithms*

- Each Vera Rubin 5-second 13.6GB data transfer will generate ~337MB of telemetry data.
  - 1,334 observations/night: 450GB of telemetry data/night

- Challenges:
  - How to save Gbps of telemetry reports without increasing OPEX (rack space, power consumption, etc.)
  - How/Where/How long to store such data?
  - How to make it available preserving privacy but without compromising research?
  - What data is really necessary from the telemetry report?
  - What has to be combined with reports to give context? Topology?

- Challenge 3 is wide-open. AmLight is a small team and we are looking for collaborations.

Thank you!

- CI Lunch and Learn 2019:
  - https://www.youtube.com/watch?v=RRg9uFz9GkA&feature=youtu.be
- CI Lunch and Learn 2021:
  - Jan 22nd, 2021
  - CI Engineering Lunch & Learn Series
    - https://www.es.net/science-engagement/ci-engineering-lunch-and-learn-series/
  - Demonstration, Interfaces, code
  - Challenges and opportunities for collaborations

References

1. INT Specification: https://p4.org/assets/INT-current-spec.pdf
2. AmLight: http://www.amlight.net
3. Bezerra, J., Arcanjo, V., Ibarra, J., Kantor, J., Lambert, R., Kollross, M., Astudillo, A., Sobhani,S., Jaque, S., Petravick, D., Morgan, H., Lopez, L., "International Networking in support of Extremely Large Astronomical Data-centric Operations". Presented at Astronomical Data Analysis Software and Systems (ADASS XXVII) conference, Santiago, Chile, October 22-26, 2017. http://www.adass.org/
4. NoviFlow : http://www.noviflow.com
5. XDP: https://www.iovisor.org/technology/xdp
6. N. V. Tu, J. Hyun, G. Y. Kim, J. Yoo and J. W. Hong, "INTCollector: A High-performance Collector for In-band Network Telemetry," *2018 14th International Conference on Network and Service Management (CNSM)*, Rome, 2018, pp. 10-18.

Ongoing Deployment

- At each pop, NoviFlow/Tofino switches [4] are being deployed
- Each pop has an INTCollector parsing Gbps of telemetry and uploading Kbps of network state of the SDN Looking Glass
- A new Control Plane was created to replace the legacy environment using OpenFlow 1.3, gRPC, and P4.