AmLight ExP
Americas Lightpaths Express & Protect

TNC / 2019

AmLight-INT: In-band Network Telemetry to support big data applications

Tallinn, Estonia

Jeronimo Bezerra
AmLight Chief Network Engineer

Outline

- Network Monitoring: Current Limitations and Technologies
- Introduction to Network Telemetry
- Current Network Telemetry Efforts
- Telemetry at AmLight: A Use Case

Warming Up!

All networkers in the room have heard at least one of these questions from users:

- *Why am I not getting full bandwidth for my download?*
- *Why is my application achieving such poor network performance?*
- *Where is the packet loss?*
- *May I be placed outside of the firewall?*

Extra challenging if real-time answers are needed!

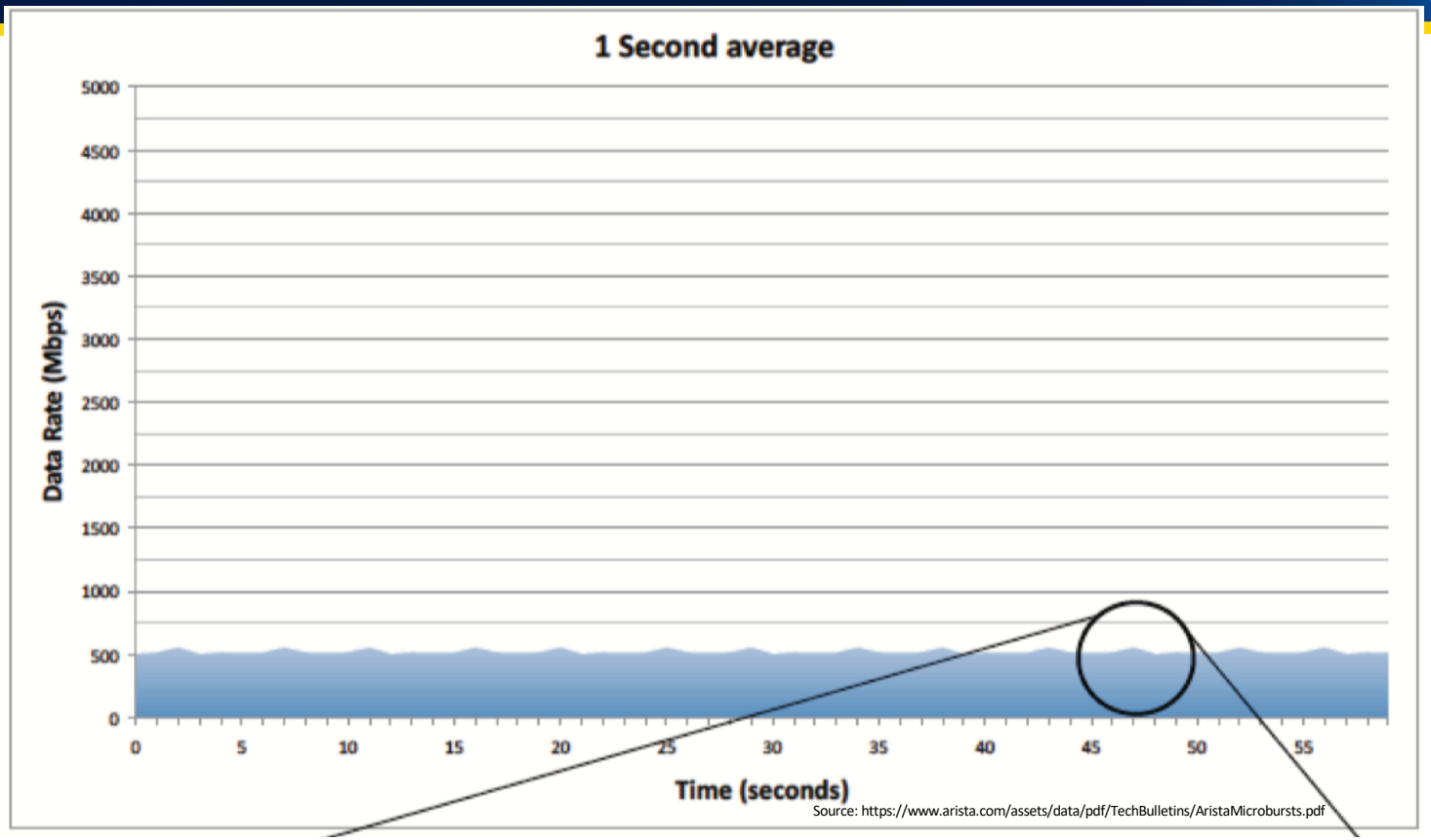
Current Network Monitoring Limitations and Technologies

- Any attempt to track network utilization in *real-time* could become a very complex and expensive task.
  - Polling SNMP or OpenFlow counters is not recommended in a sub-30s interval.
  - Sampling technologies usually export data after a few seconds.
  - Packet sniffing at 100G has a high CAPEX and OPEX.

- Identifying microburst or isolating packet loss in *real time* is not trivial with current technology.

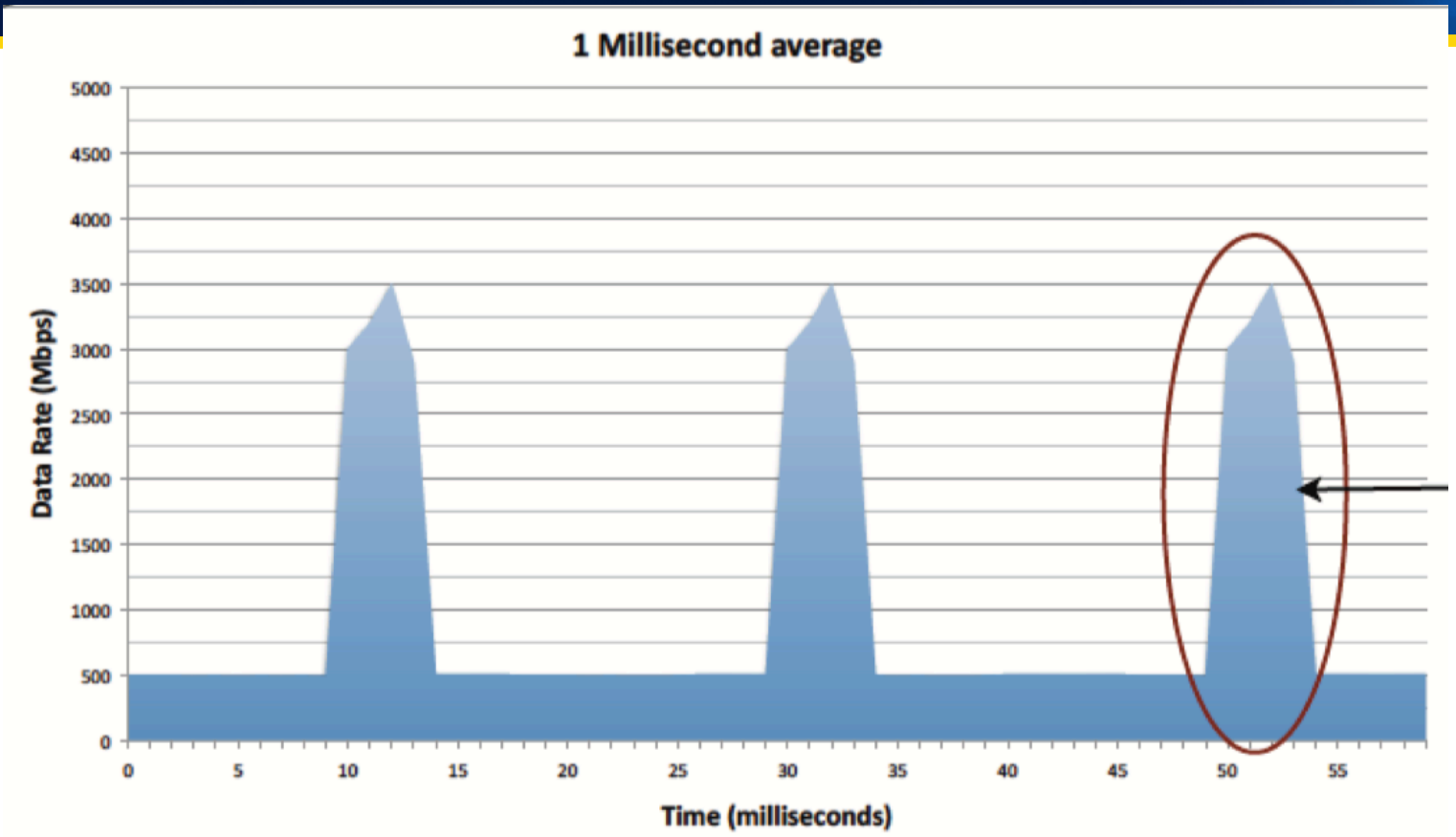
- To add complexity, new *real-time* big-data applications are being created with very strict Service Level Agreements (SLA).

Microbursts



Source: https://www.arista.com/assets/data/pdf/TechBulletins/AristaMicrobursts.pdf

Microbursts [2]



Introduction to *Network Telemetry*

- *Network telemetry is the extension of network reporting to higher granularities and sample rates combined with actionable metrics and alerting [1]*

- Network telemetry technologies define several characteristics [2]:
  - Push and Streaming: Instead of polling data from network devices, the telemetry collector subscribes to the streaming data pushed from data sources in network devices.
  - The data is normalized and encoded efficiently for export.
  - The data is model-based which allows applications to configure and consume data with ease.
  - Network telemetry means to be used in a closed control loop for network automation
  - Also known as *streaming network telemetry* or *streaming telemetry*

- Streaming network telemetry is very useful to detect microburst and queue utilization at a sub-second interval

- With all historic network state, forensic troubleshooting is enabled

[1] https://www.preseem.com/2017/03/network-telemetry/
[2] https://tools.ietf.org/html/draft-ietf-opsawg-ntf-01

New Telemetry Trends @ IETF and ONF

- In 2016, P4.org group create a new P4 application:
  - In-band Network Telemetry (2016)

- IETF Internet Protocol Performance Measurement (ippm) WG:
  - Proof of Transit (2016)
  - Encapsulations for In-situ OAM Data (2017)
  - Data Fields for In-situ OAM (2017)
  - Requirements for In-situ OAM (2018)

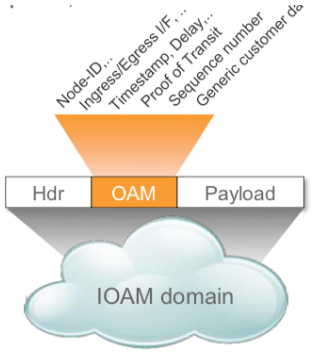
- IOAM, In-situ OAM, In-band OAM, INT, In-band Network Telemetry are used interchangeably in *this* presentation.

In-band Network Telemetry

- INT is an implementation to record operational information in the packet while the packet traverses a path between two points in the network:
  - Complements current out-of-band OAM mechanisms based on ICMP or other types of probe packets.
  - Basically, INT adds metadata to each packet with information that could be used later for troubleshooting activities.

- Example of information added:
  - Timestamp, ingress port, egress port, pipeline used, queue buffer utilization, WiFi link power, CPU utilization, Battery Utilization, Sequence #, and many others

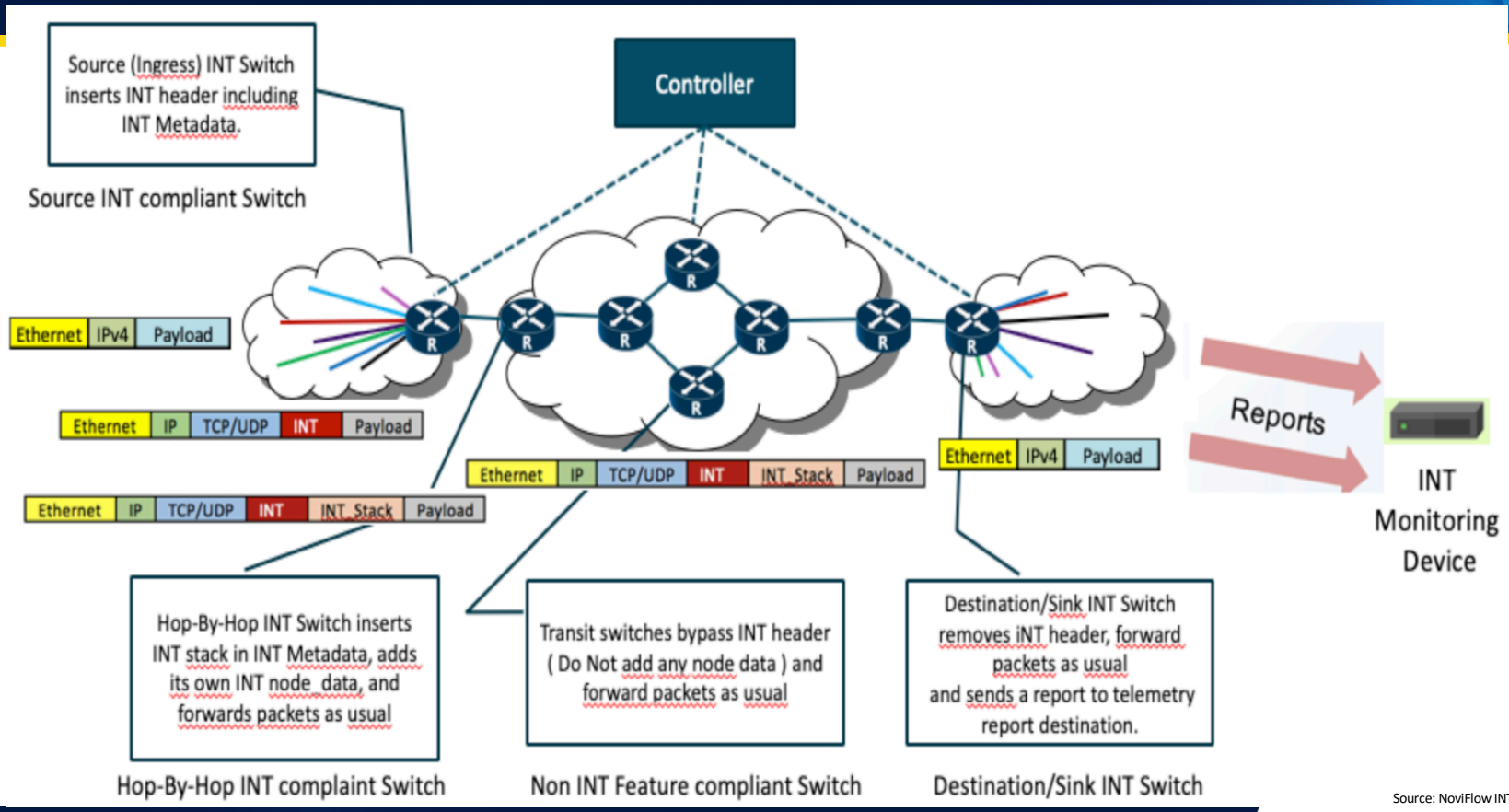
- As metadata is exported directly from the Data Plane, Control Plane is not affected:
  - Translating: *you can track/monitor/evaluate EVERY single packet at line rate.*



Questions addressed by INT

- *How did this packet get here?*
  - The sequence of network devices a packet visited along its path.
    - LAG? No problem. ECMP? No problem. Layer 2 network? No problem!
- *Why is this packet here?*
  - The set of rules a packet matched upon at every switch along the way.
- *How long was this packet delayed?*
  - The time a packet spent buffered in every switch, to the nanosecond, from end-to-end.
- *Why was this packet delayed?*
  - The flows and applications that a packet shared *each queue with*.

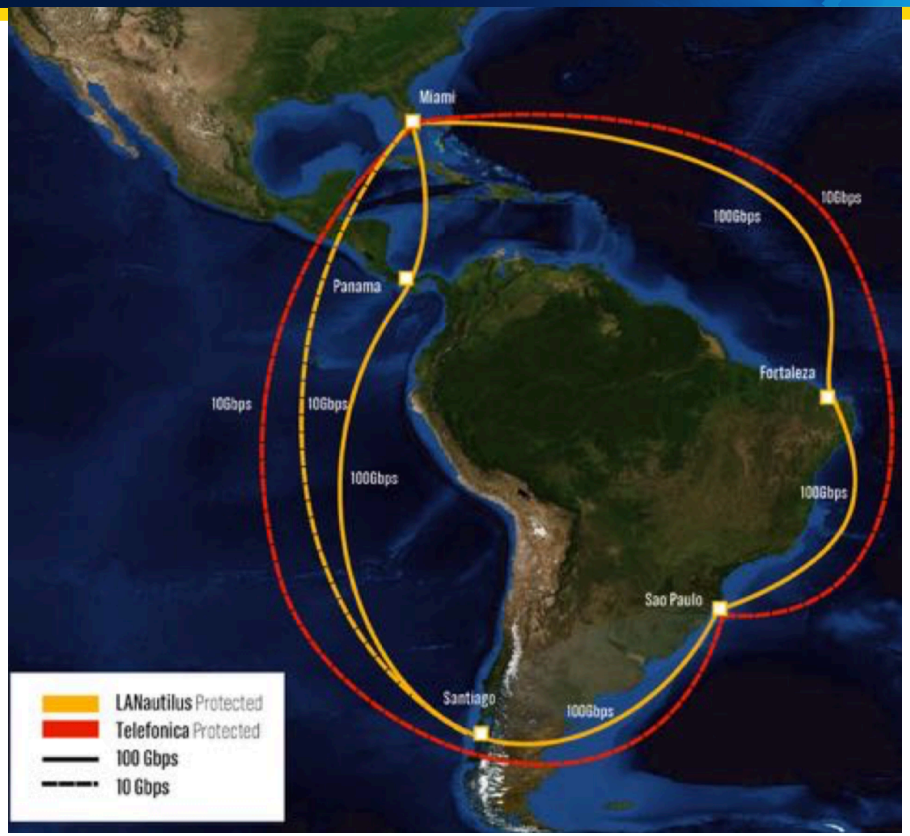
INT: How does it work?



Source: NoviFlow INT SetUp Guide

Introduction to AmLight

- Production SDN Infrastructure since 2014
- NAPs: Miami, Brazil(2), Chile, Puerto Rico, and Panama
- Multiple 10G and 100G links
- Carries Academic and Commercial traffic
- Control Plane: OpenFlow 1.0 and 1.3
- Network Programmability/Slicing
- Inter-domain Provisioning with NSI
- A partnership among FIU, NSF, RNP, ANSP, CLARA, REUNA, and AURA.



Telemetry at AmLight: LSST Use Case

- Support the Large Synoptic Survey Telescope (LSST)'s requirements
  - The LSST will be installed in Chile
  - Every 27 seconds throughout the night, the telescope 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 Illinois/USA
    - From the 27-seconds window, only 5 seconds are available for data transmission
  - Multi traffic types with different priorities (db sync, control, general internet traffic)

Telemetry at AmLight: LSST Use Case

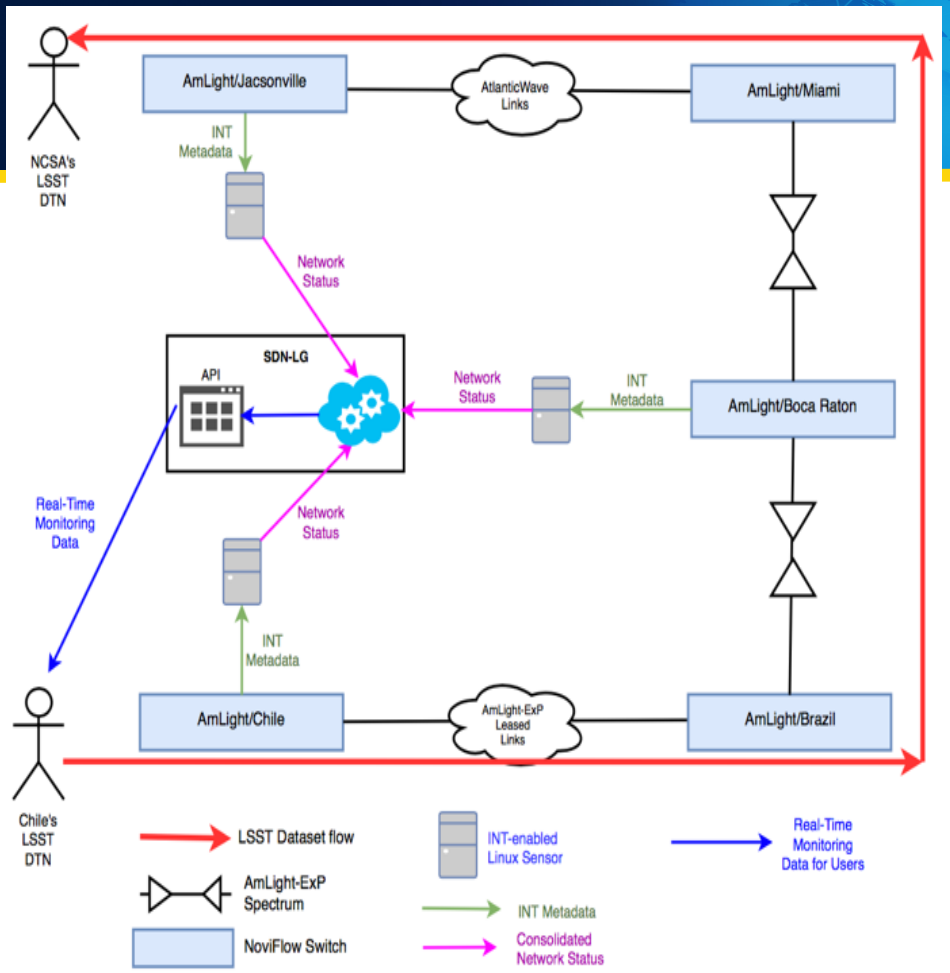
- What if the LSST doesn't manage to send its data in its 5-seconds transfer window?
  - For instance, because of packet loss, lack of capacity, lack of buffers, microburst, DoS attacks?

- If the data transfer window is missed, will AmLight engineering team be able to fix whatever it is happening before the next data transfer window (in less than 22 seconds)?

- How many windows are we going to miss if we have to troubleshoot it manually?
  - ***AmLight-INT Project might be the solution!***

AmLight-INT Project

- NSF IRNC: Backbone: AmLight In-band Network Telemetry (AmLight-INT), Award# OAC-1848746
- AmLight-INT Project Plan:
  - 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



AmLight-INT Project

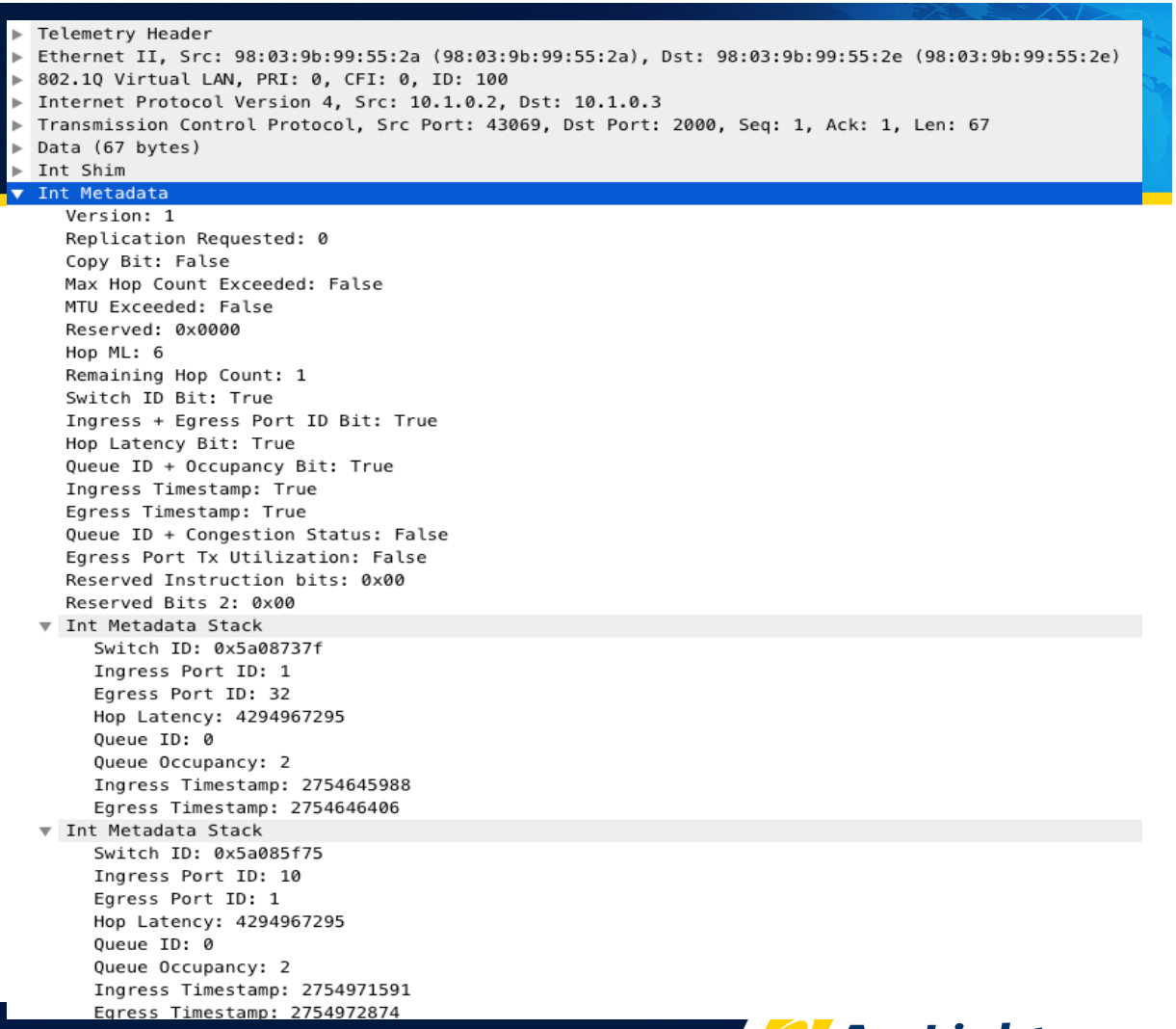
- AmLight-INT is a collaboration between FIU and NoviFlow to expand AmLight SDN network towards an INT-capable domain

- Characteristics of the NoviFlow WB5132 switches @ AmLight:
  - Support OpenFlow 1.3 (also 1.4 and 1.5)
  - Support BFD
  - 32 x 100G (high throughput: 3.2 Tbps)
  - Barefoot/Tofino Chipset:
    - Provides a software-based SDN evolution path to P4-Runtime

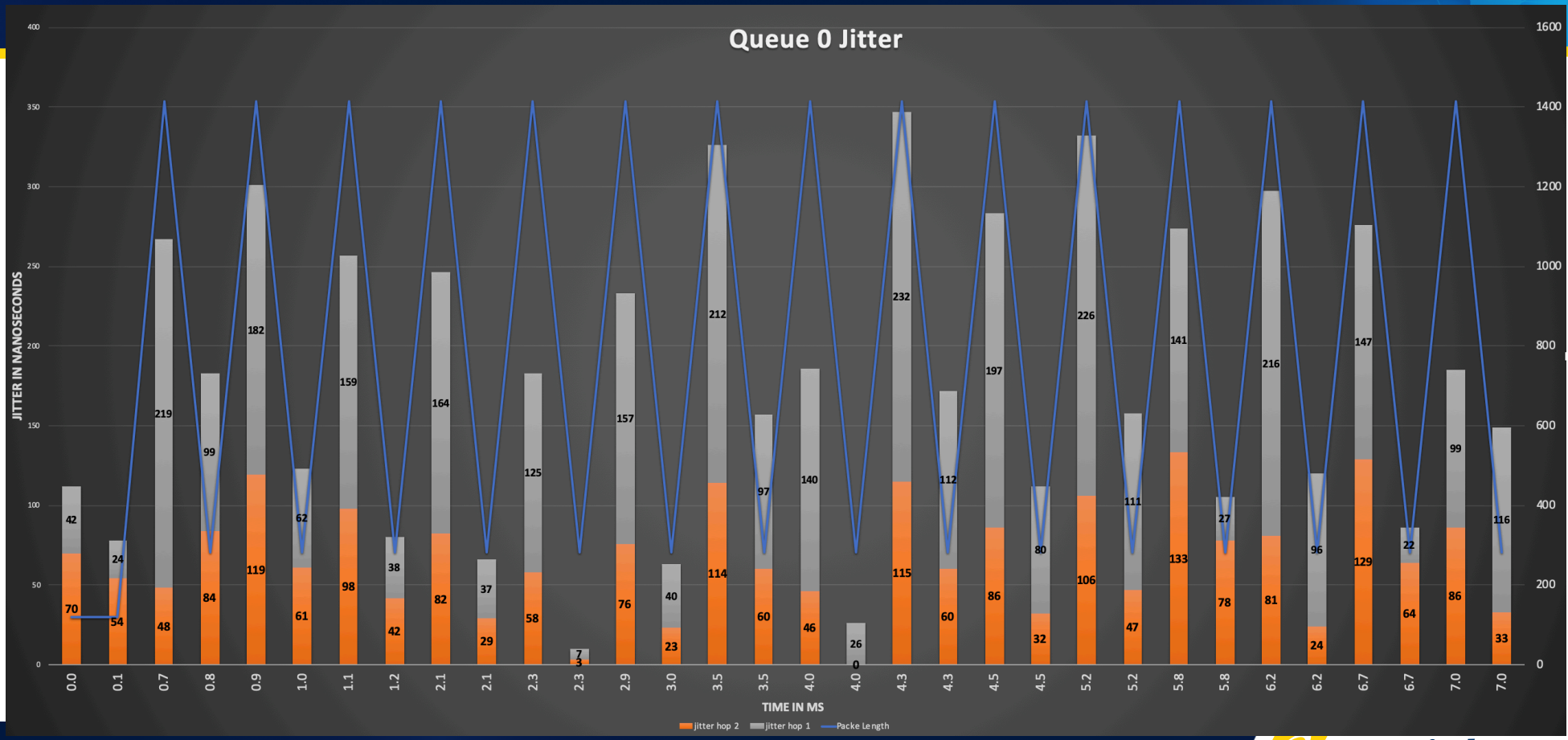
- NoviFlow has already provided two NOS code to enable INT
  - P4/INT specification being followed
    - Nothing is proprietary or strictly created to support the LSST project

First Results

- Wireshark Dissector created by NoviFlow



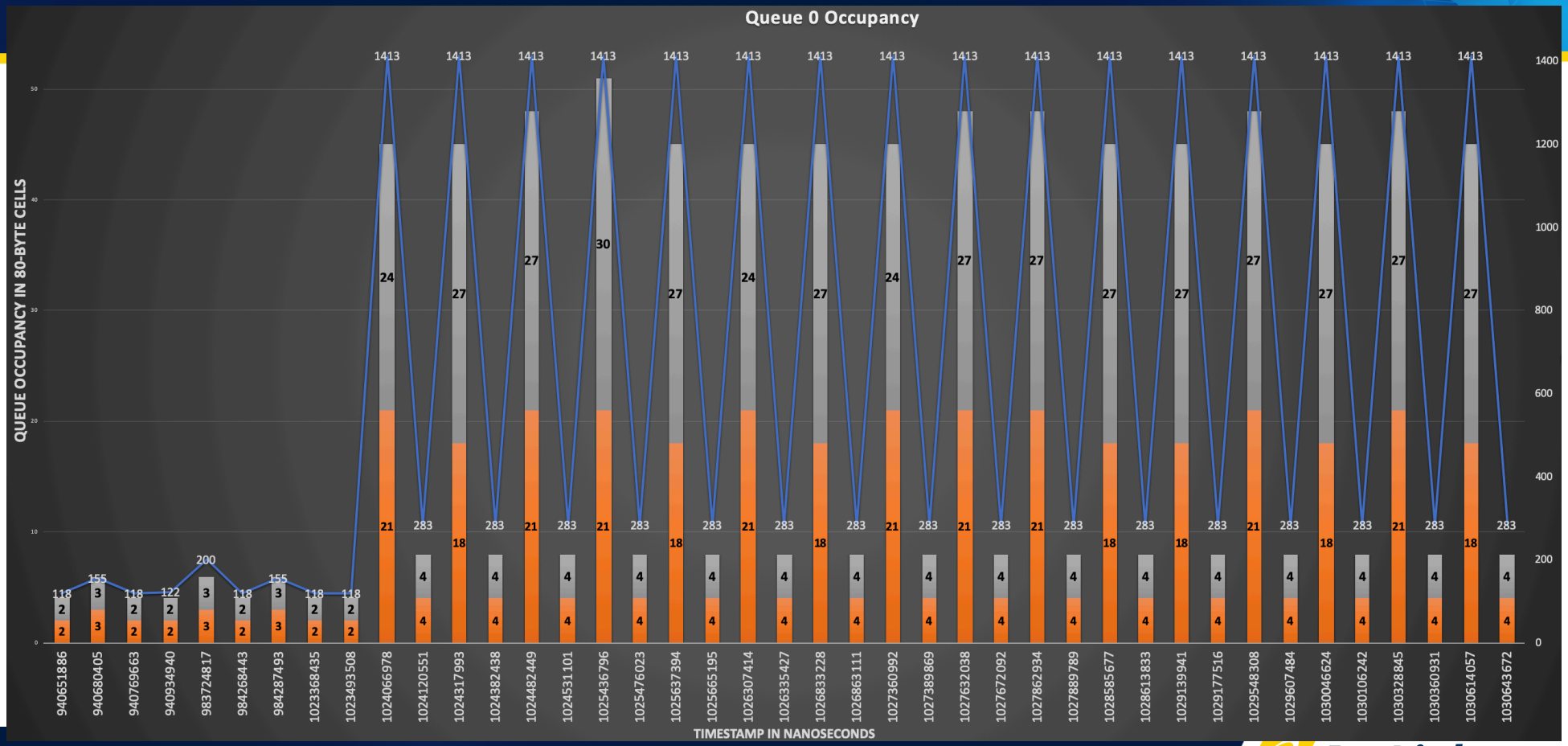
First Results [2] – Queue 0's Jitter



18

AmLight ExP
Americas Lightpaths Express & Protect

First Results [3] – Queue 0's Occupancy



AmLight ExP
Americas Lightpaths Express & Protect

19

Challenges being addressed

- *How many high-priority flows can be handled in real-time by a typical network server (INT Collector)?*
- *What is the impact caused by INT in a complex network such as AmLight-ExP (MTU, delay)?*
- ~~*How to dynamically enable INT monitoring of specific flows?*~~
- *What is the definition of real-time for AmLight and LSST?*
- *How to store and process multiple Gbps of telemetry data per switch?*

AmLight ExP
Americas Lightpaths Express & Protect

TNC / 2019

Thank You!

AmLight-INT: In-band Network Telemetry to support big data applications

Jeronimo Bezerra

AmLight Chief Network Engineer

Tallinn, Estonia