AmLight EXP
Americas Lightpaths Express & Protect

Internet2 Technology Exchange

Dec 10th, 2019

AmLight-INT: In-band Network Telemetry @ AmLight

Jeronimo Bezerra - IT Associate Director/FIU

Arun Paneri – Director of Engineering/NoviFlow

Outline

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

Challenges to [Academic] Network Operators

- [Academic] network monitoring and operation have never been so diverse:
  - Big data applications, dynamic circuits, clouds, SDN/SDX, compute/storage & network integration, network security, optical spectrum sharing, federation, …
  - Sometimes, network applications require real-time SLA-driven performance
  - Sometimes, inter-domain
- For network operators, we are monitoring and measuring network utilization and performance using tools designed for a completely different scenario
  - Such tools can collect counters based on samples or on-demand (SNMP, NetFlow...)
  - Port-mirror and network taps impose huge challenges for scalability
- Any network performance assessment performed in such environment is extremely complex and time-consuming
  - Especially complex when real-time answers are needed

Introduction to *Network Telemetry*

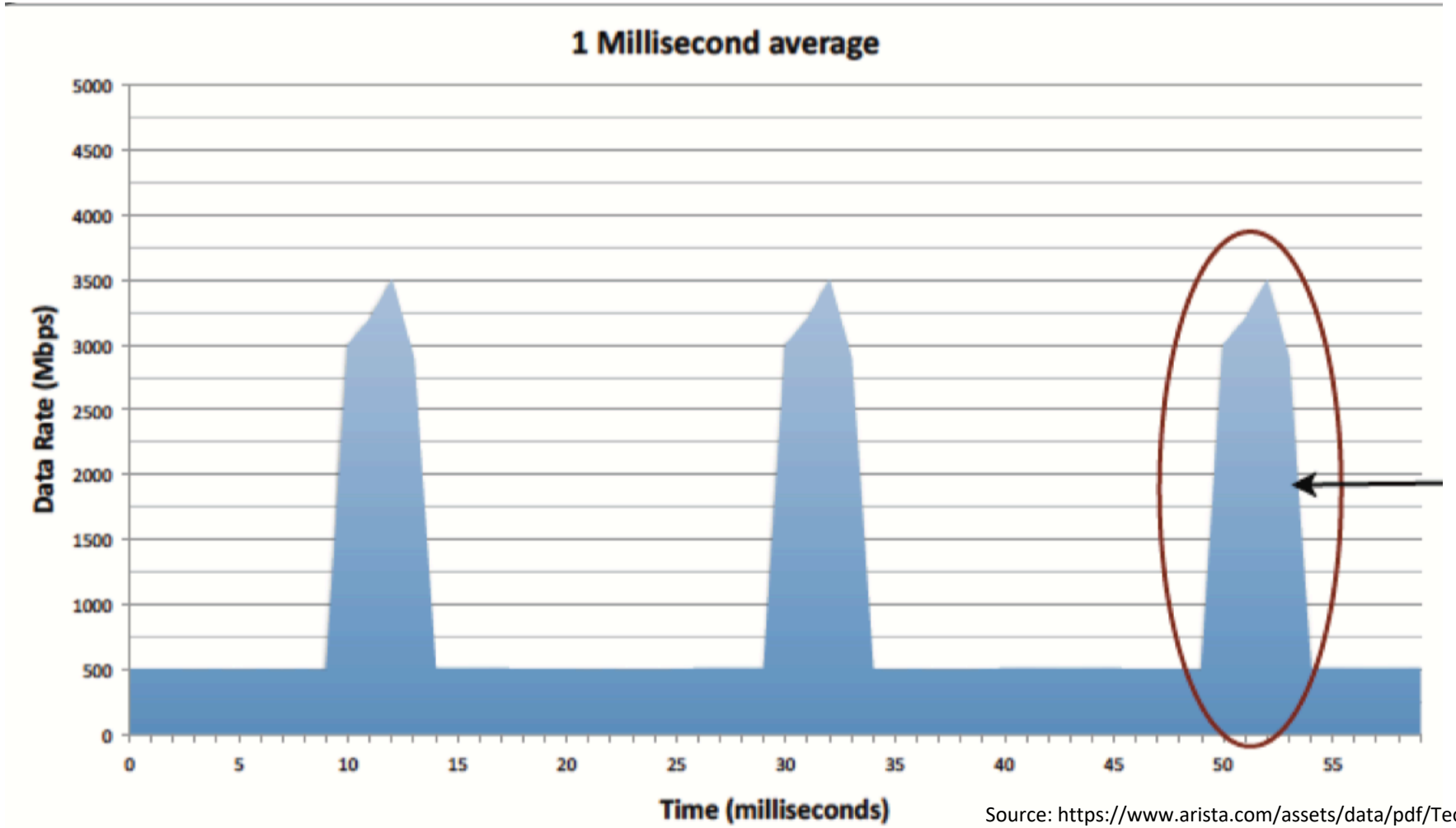
- *Network telemetry is the extension of network reporting to higher granularities and sample rates combined with actionable metrics and alerting [1]*

- 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

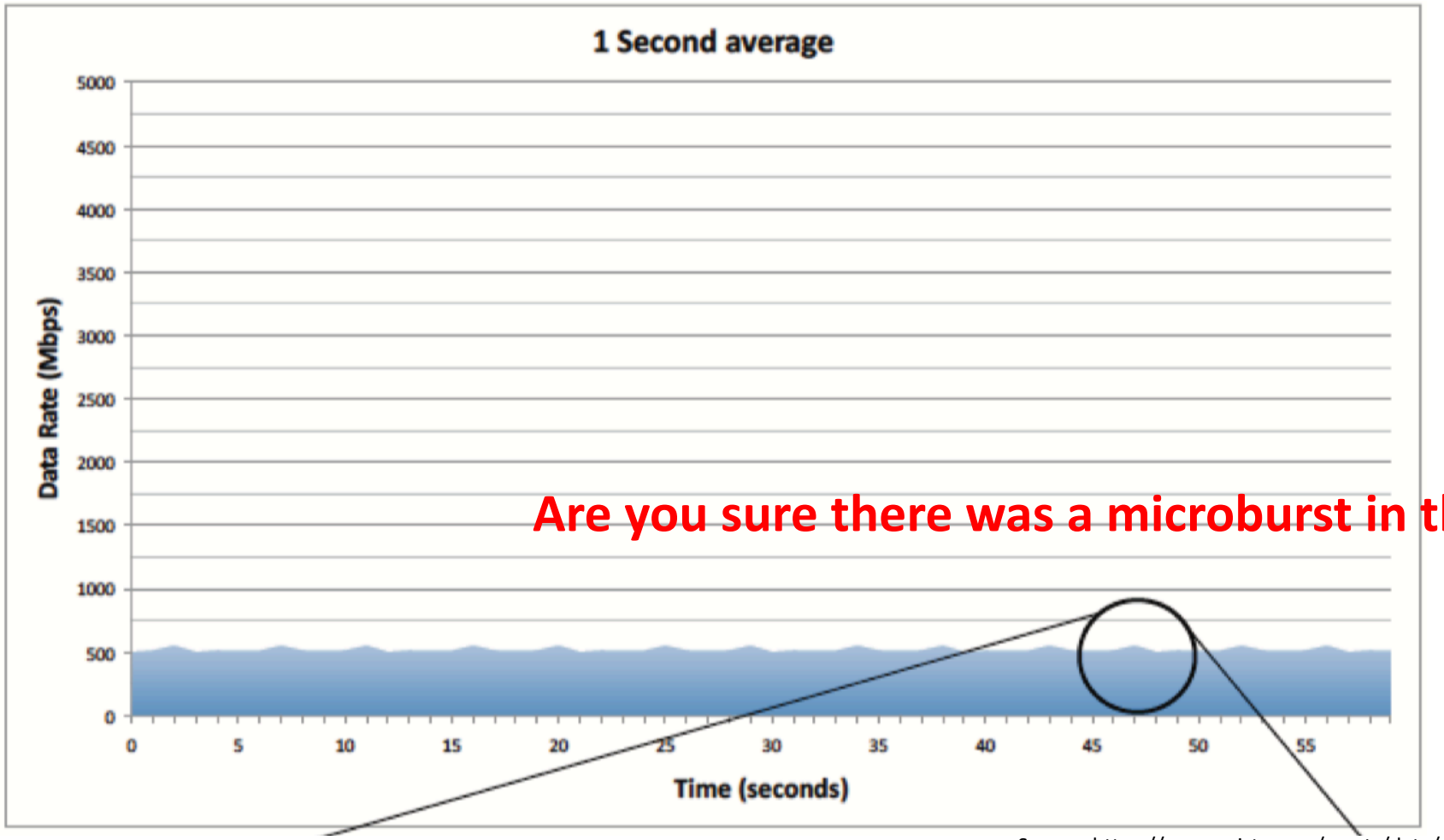
- With all historic network state, forensic troubleshooting is enabled

Example: Microbursts vs. Telemetry



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

Example: Microbursts vs. Legacy Monitoring



Are you sure there was a microburst in the last minute?

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

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)

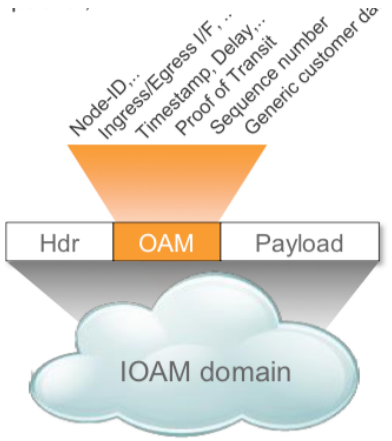
- IOAM, In-situ OAM, In-band OAM, INT, In-band Network Telemetry are used interchangeably in *this* presentation.

In-band Network Telemetry (INT)

- 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.*



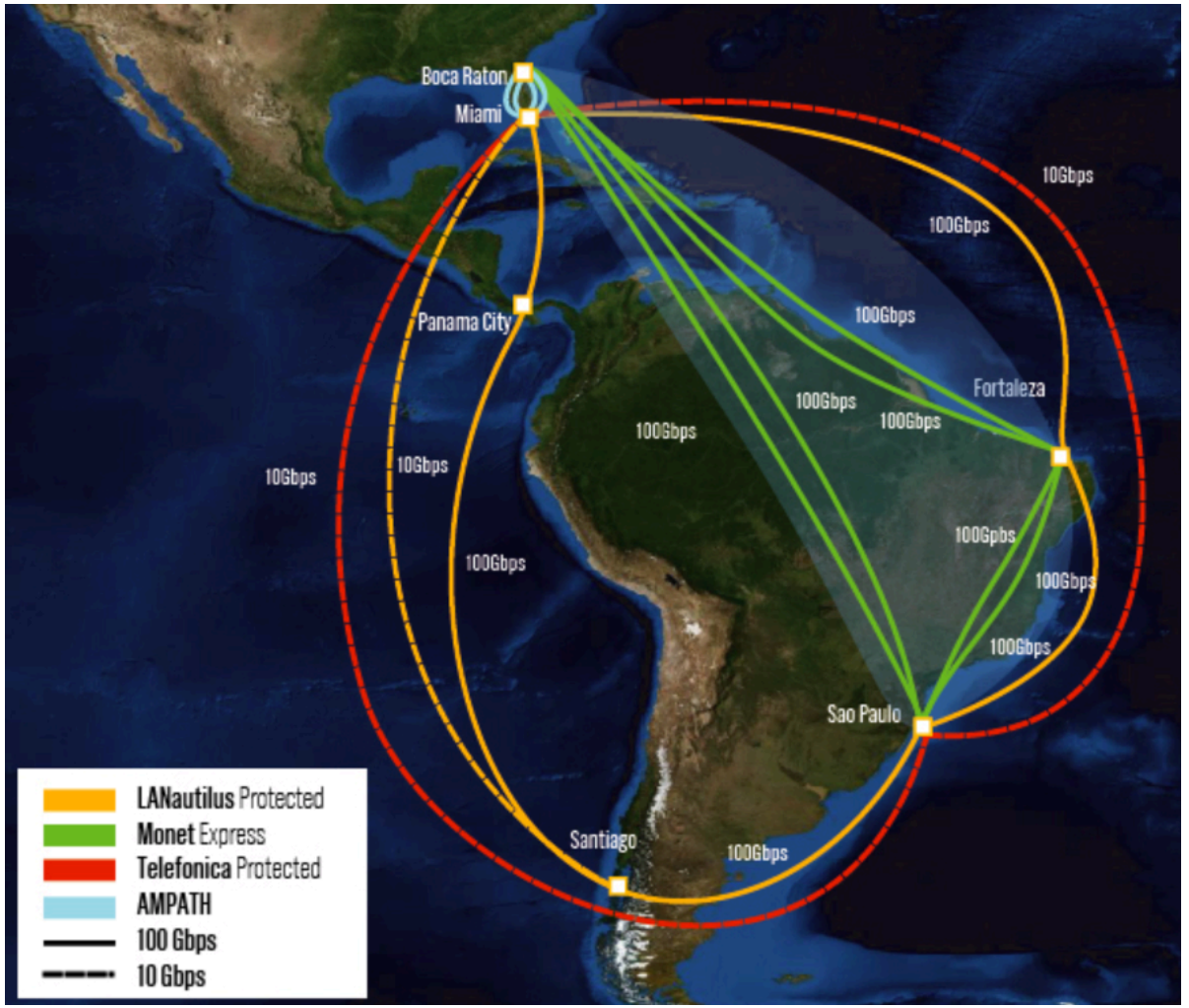
Issues addressed with In-band Network Telemetry

- Per Interface's Buffer Occupancy
  - Identifying sources of delay and jitter
  - Packet drop mitigation
- Proof-of-transit
- Instant bandwidth availability and micro-bursts detection
- Packet loss mitigation
- Rogue traffic identification
- TCP performance mitigation

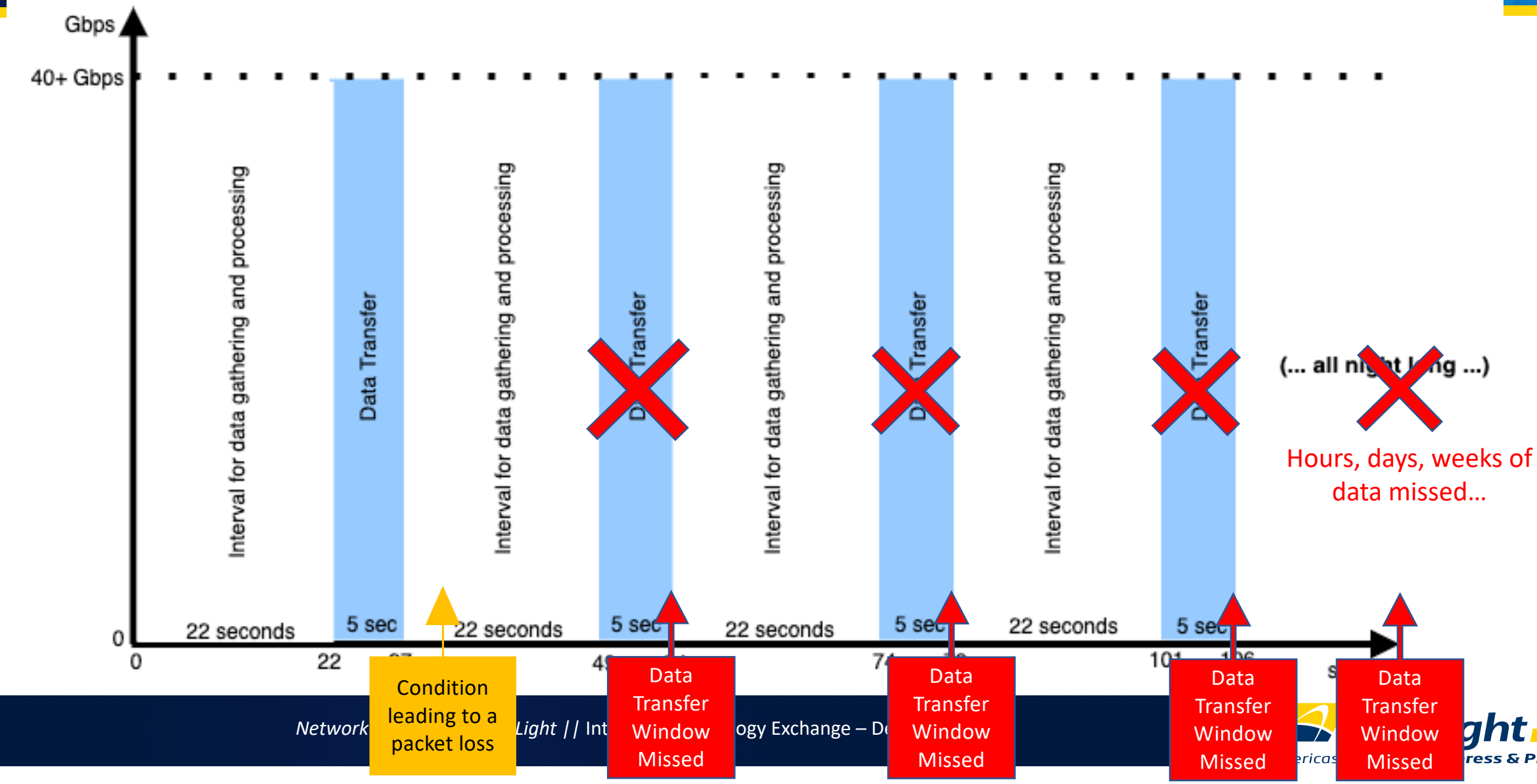
In real-time and at line-rate!

Introduction to AmLight

- AmLight Express and Protect (AmLight-ExP) (NSF International Research Network Connections (IRNC) Award #1451018)
- 680Gbps of upstream capacity between the U.S. and Latin America
- Production SDN Infrastructure since 2014
- NAPs: Florida(2), Brazil(2), Chile, Puerto Rico, and Panama
- Carries Academic and Commercial traffic
- Control Plane: OpenFlow 1.0 and 1.3
- Inter-domain Provisioning with NSI
- A consortium involving FIU, NSF, RNP, ANSP, CLARA, REUNA, and AURA.



The Use Case: Large Synoptic Survey Telescope (LSST)



Telemetry at AmLight: LSST Use Case [2]

- 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?

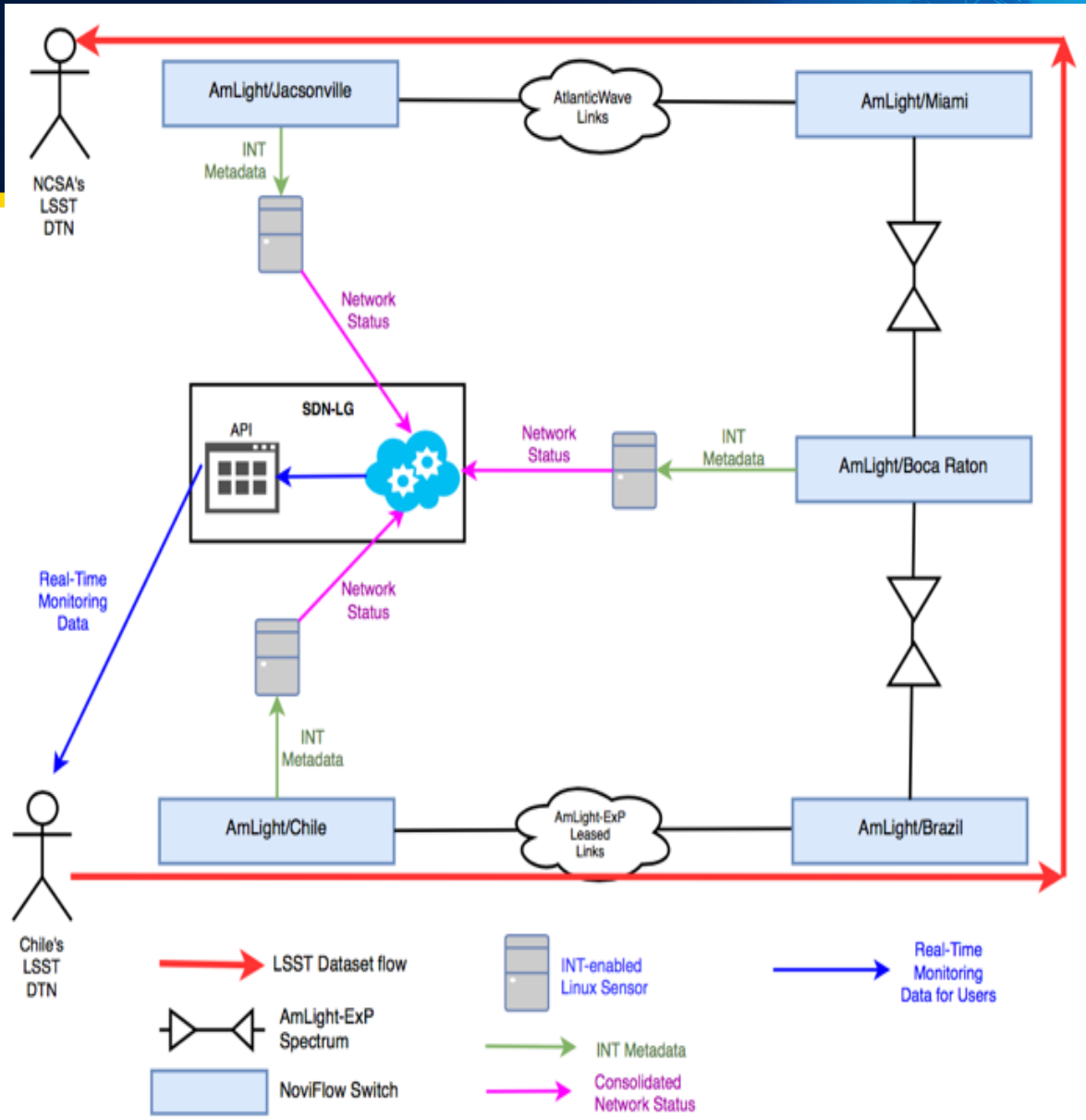
- If the data transfer window is missed, will AmLight engineering team be able to fix whatever it is happening before the next data transfer window (in less than 22 seconds)?

- How many windows are we going to miss if we troubleshoot it manually?

AmLight-INT Project

- NSF IRNC: Backbone: AmLight In-band Network Telemetry (AmLight-INT), Award# OAC-1848746
  - Started in Nov 2018
  - Collaboration with NoviFlow

- 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

- Collaboration between FIU and NoviFlow to expand AmLight SDN network towards an INT-capable domain

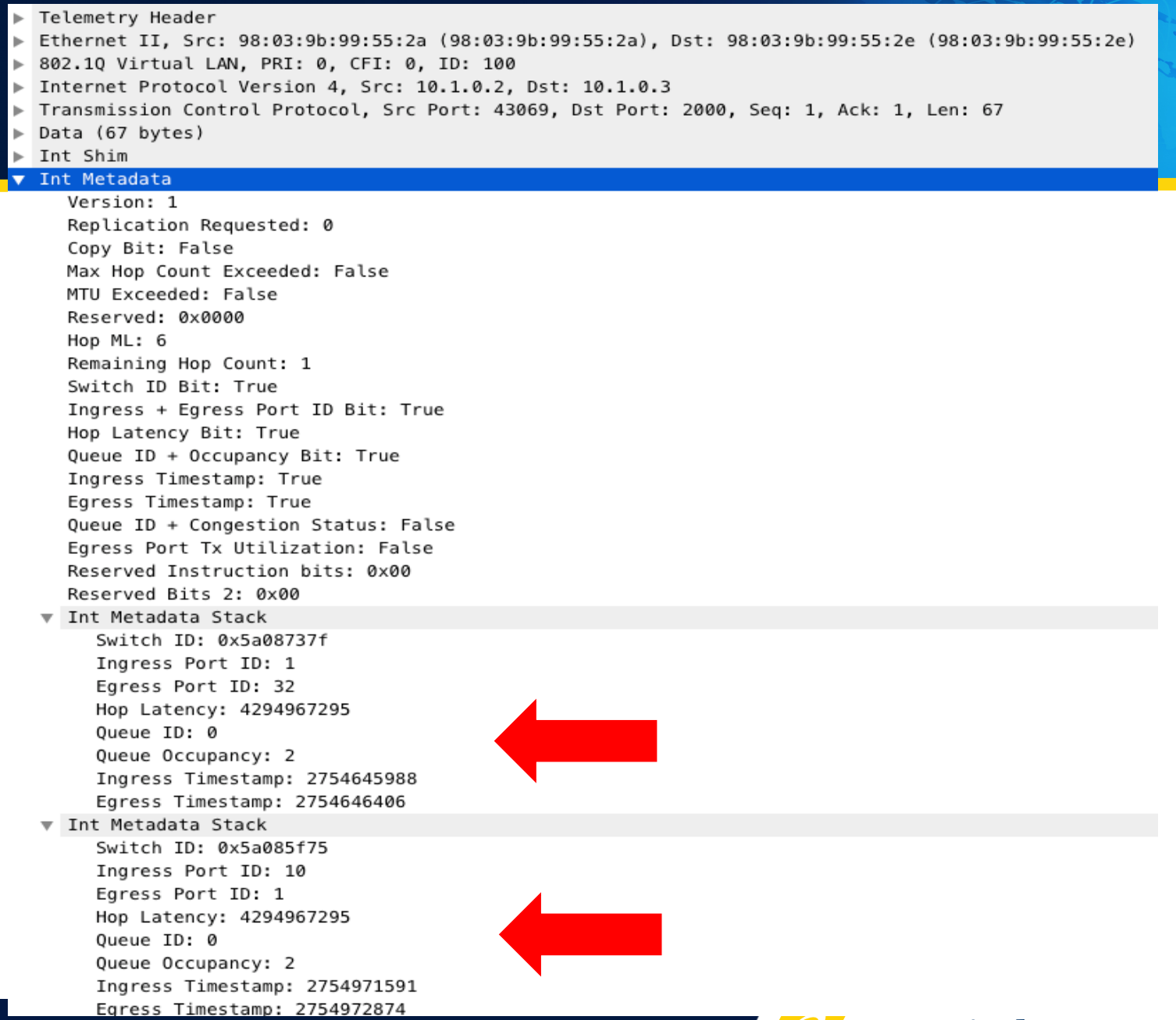
- Characteristics of the NoviFlow switches in use at AmLight:
  - Barefoot Tofino chip:
    - Provides a software-based SDN evolution path to P4-Runtime
    - 32 x 100G (high throughput: 3.2 Tbps)
  - NoviWare supports OpenFlow 1.3 (also 1.4 and 1.5) with BFD and LAG

- NoviFlow has already released five NOS versions to enable INT
  - P4/INT specification being followed
    - Nothing is proprietary or strictly created to support the LSST project

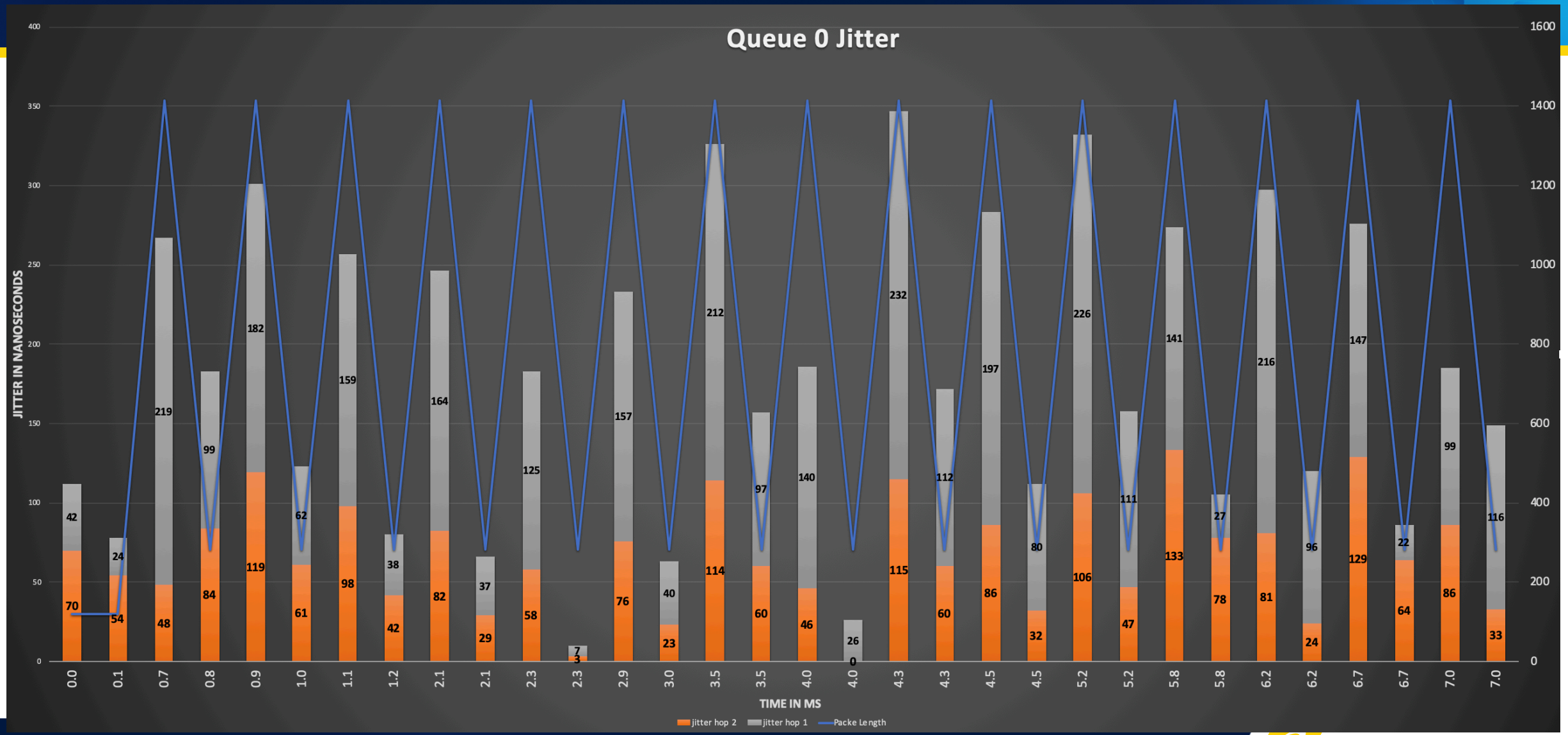
Some Results

- Wireshark Dissector created by NoviFlow (figure)

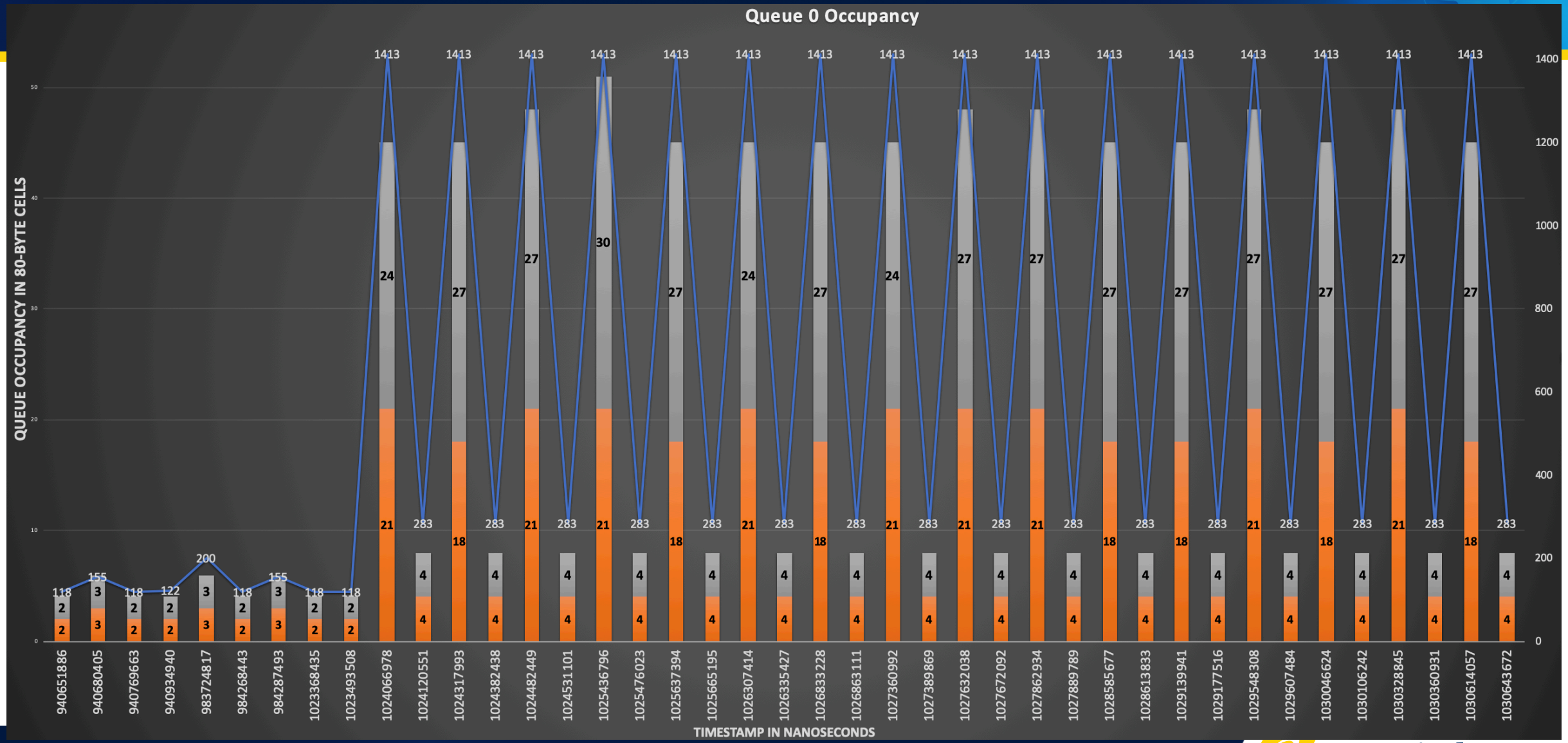
- AmLight-INT Collector v0.1:
  - Developed using Python 3.7
  - Receives Telemetry Reports from switches
  - Parses and sends to a RabbitMQ queue to be consumed
  - Saves Telemetry Reports to disk



Queue 0's Jitter



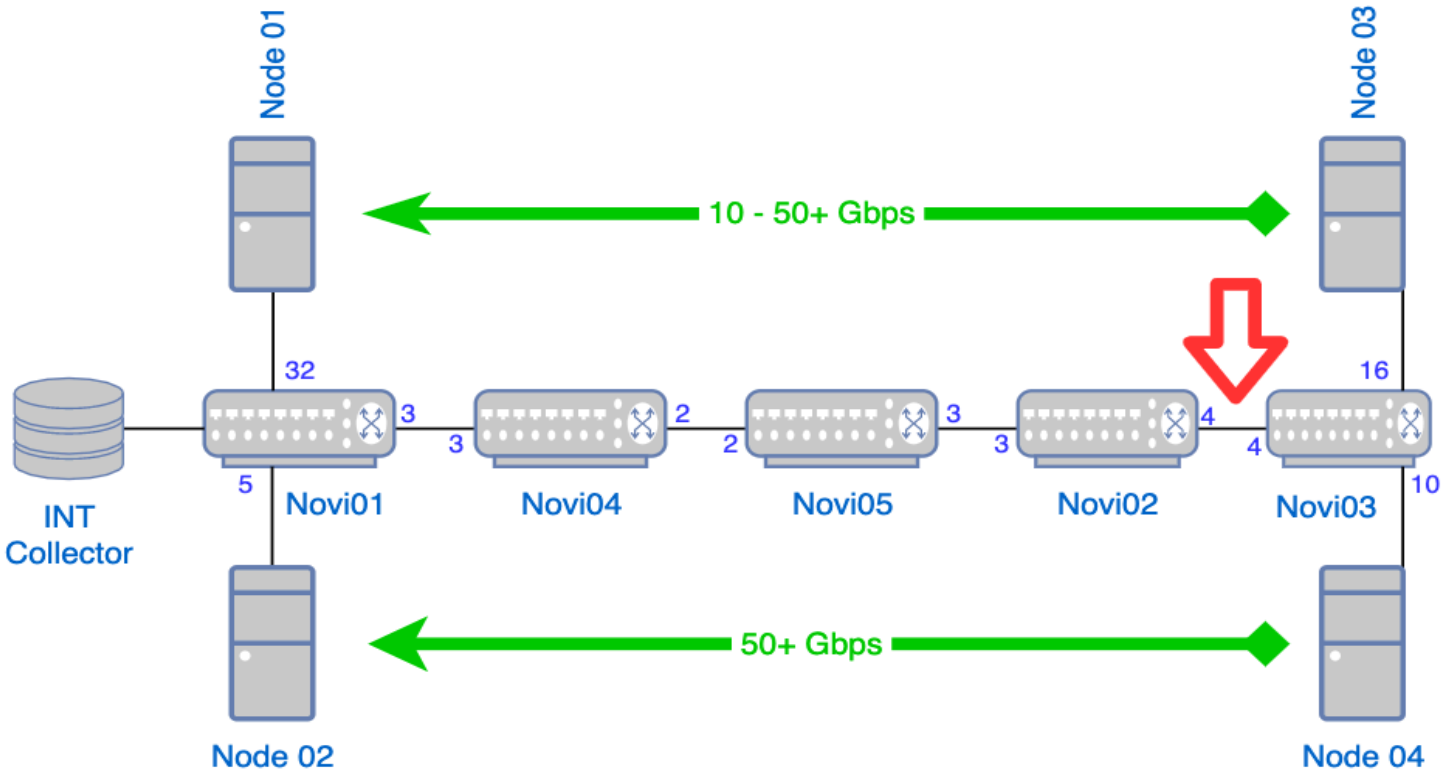
Queue 0's Occupancy



AmLight-INT QueueTop

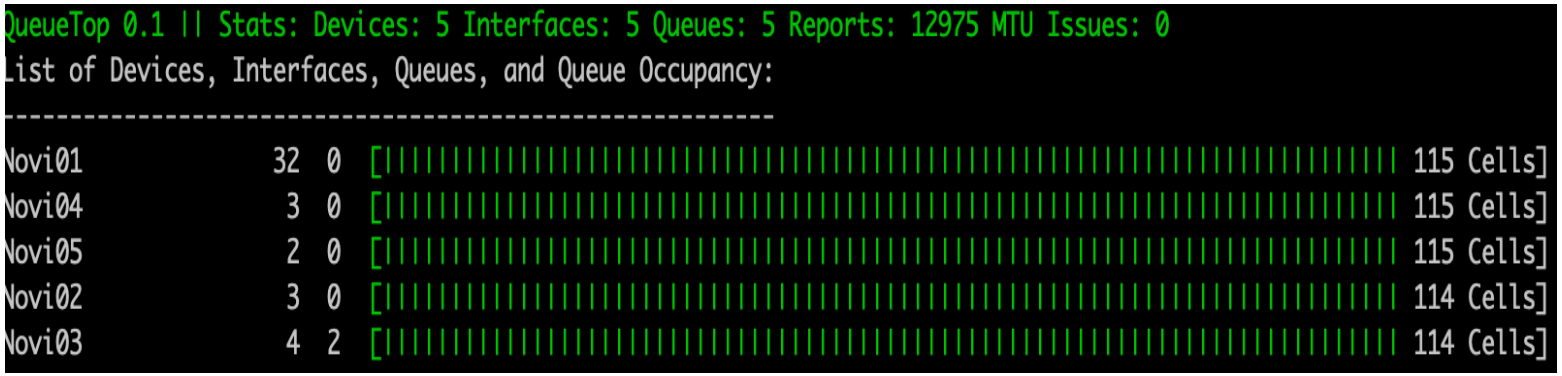
- AmLight-INT Collector's QueueTop application consumes INT data and display realtime monitoring of the network's queues

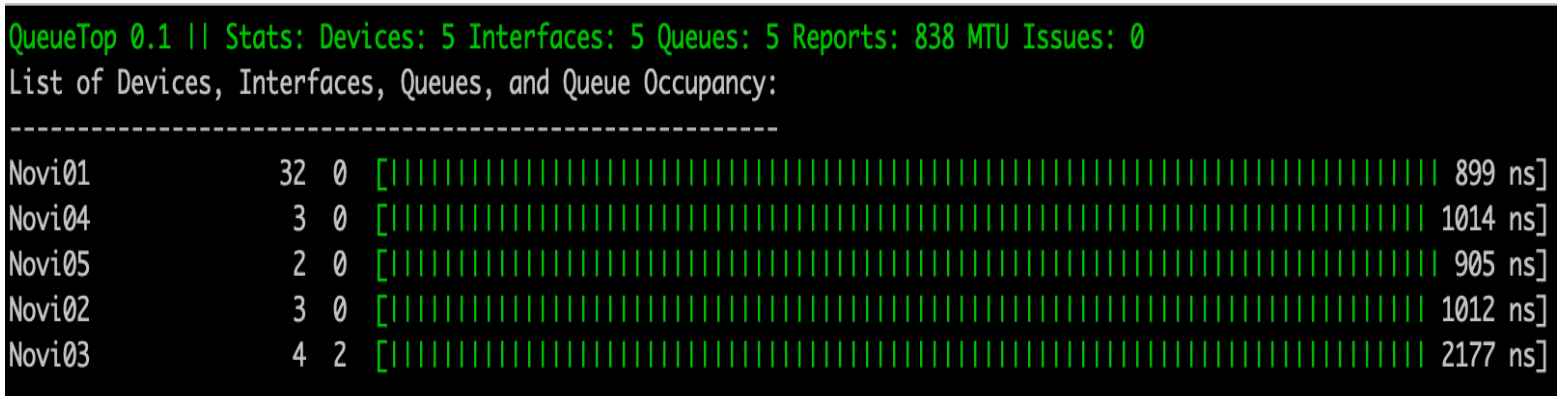
- Topology on the right created to enable experimentation
  - All links and devices are 100G
  - Novi03 switch port 04 has a bottleneck: Node 03 and Node 04 are sending data to their peers.
  - Let's see what happens next...



One Source - One Destination - TCP - ~50Gbps

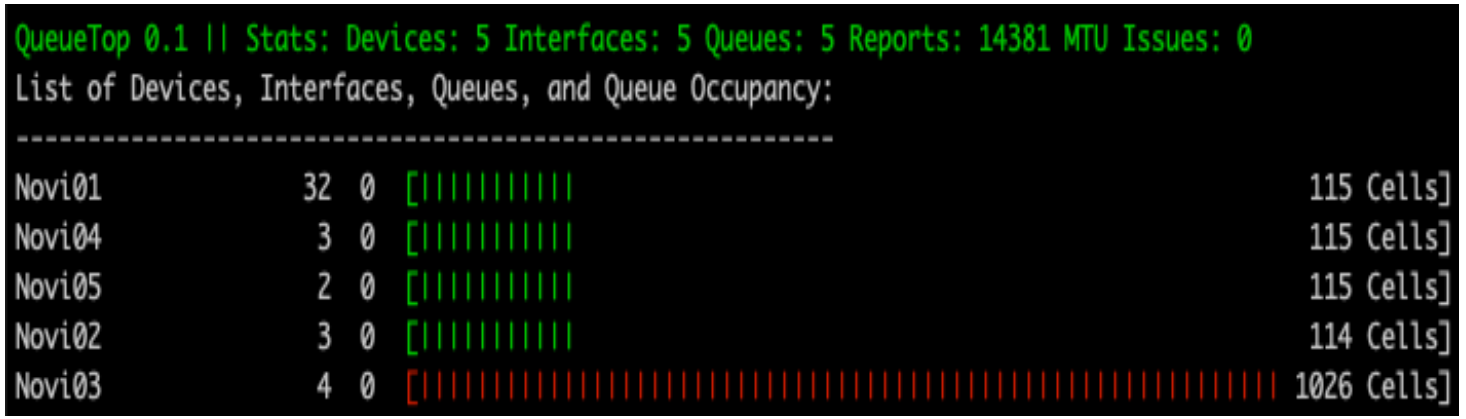
- Node 04 sending data using TCP to Node 02 at ~50Gbps
- No other traffic
- Top:
  - All Queues are using 114-115 cells (or 9K bytes)
- Bottom:
  - Hop Delay around 1 microsecond (except for Novi03 that ADDs INT header)





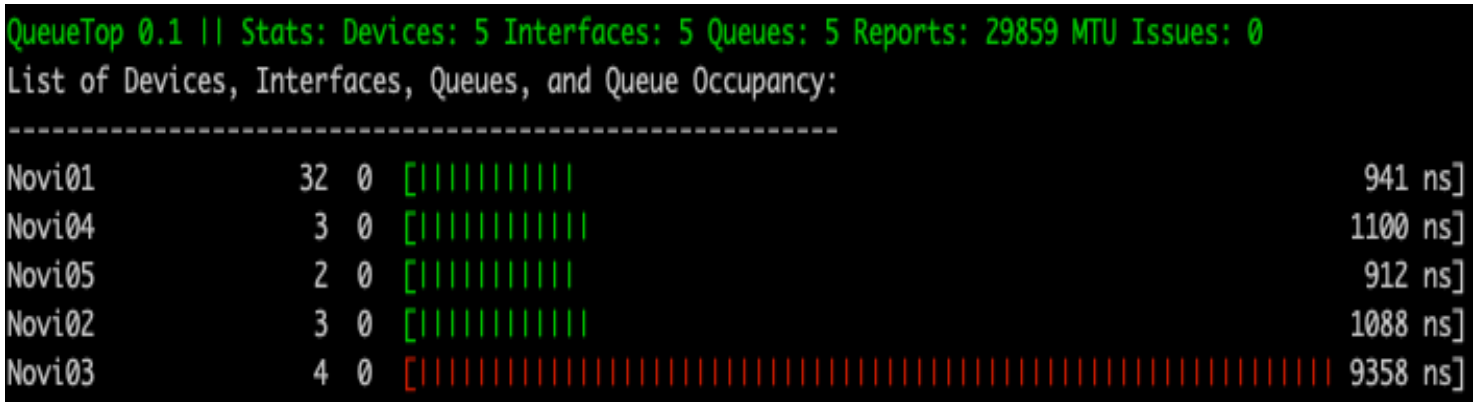
Two Sources - Two Destinations - TCP - ~80Gbps

- Node 04 sending data using TCP to Node 02 at ~50Gbps
- Node 03 sending data using TCP to Node 01 at ~25Gbps
- Shared interface/queue on Novi03 port 4
- Top:
  - Now Novi03 uses 1026 cells
- Bottom:
  - Hop Delay at Nov03 around 9 microsend (add_int_metadata and queueing)



← Hop Queue Occupancy increasing 9x

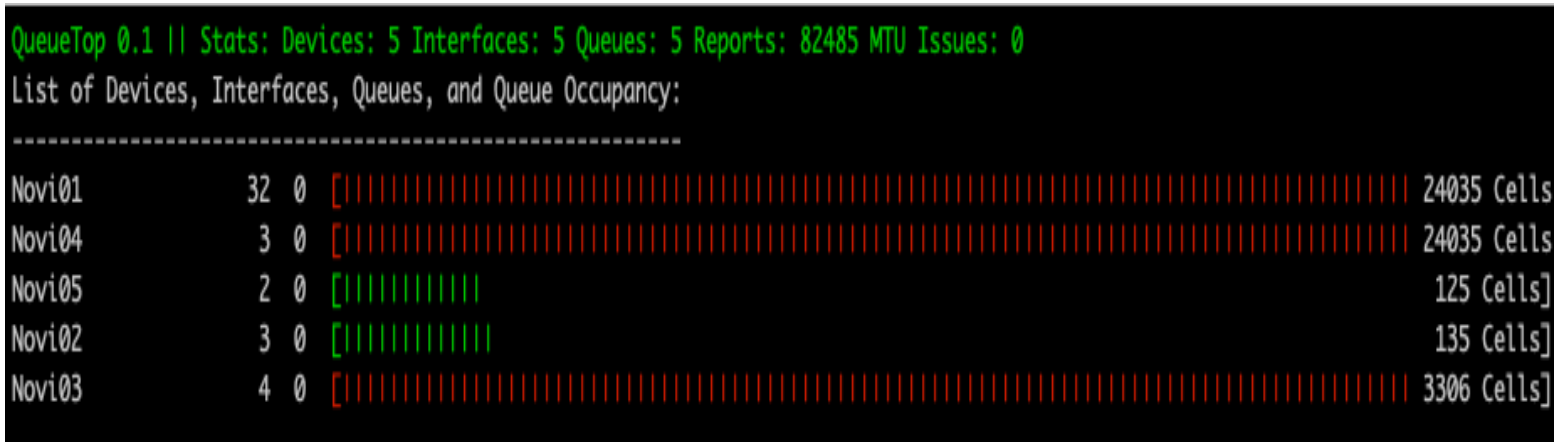
Hop Delay increasing 4x →

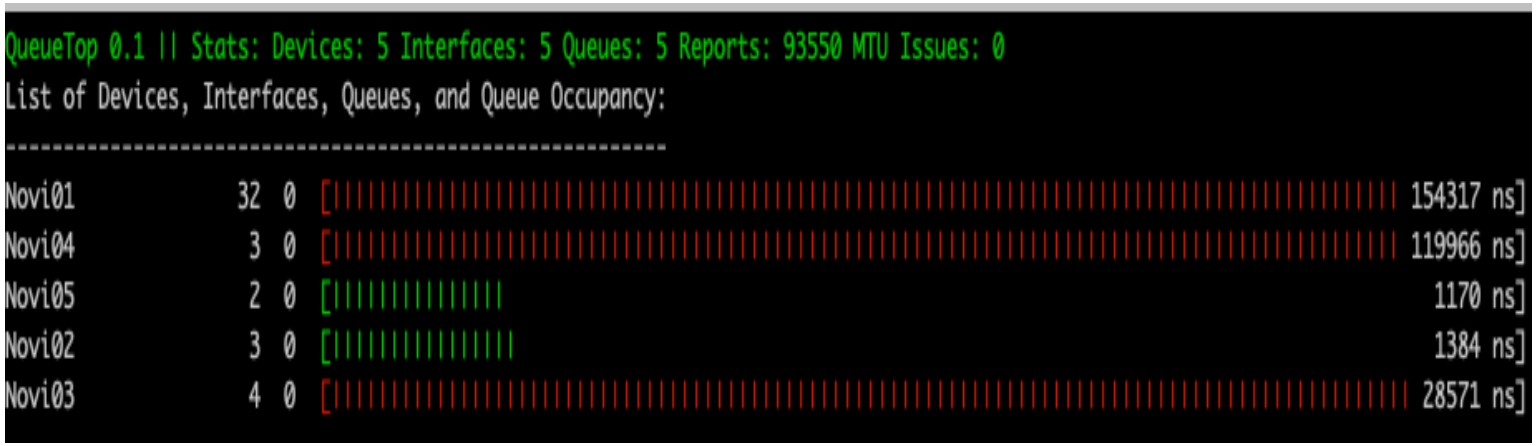


Two Sources - Two Destinations - TCP – 100% output utilization

- Node 04 trying to send as much data using TCP as possible to Node 02
- Node 03 trying to send as much data using TCP as possible to Node 01
- Shared interface/queue on Novi03 port 4
- Top:
  - Now Novi03 uses 3306 cells (or 264KB)
- Bottom:
  - Hop Delay at Nov03 around 28 microsends (add_int_metadata and queueing)

Question: What has happened to Novi01 and Novi04 queues???? Under investigation.





Next Steps

- Understanding the behavior seen so far:
  - With the current tools, we will test some theories, such as buffer sizing.
- Improve INT Collector's performance:
  - Currently, a 100Gbps flow with 9000 Bytes packets generates around 3-5 Gbps of telemetry.
  - Using Netronome P4 NICs at the INT Collectors
- Next tools:
  - Integration with InfluxDB and Elastic for network visualization/historical data.
  - All tools will be available as Open Source code through the AmLight Github account soon:
    - http://github.com/amlight

AmLight EXP
Americas Lightpaths Express & Protect

Thank You! / Questions? / Comments?

AmLight-INT: In-band Network Telemetry @ AmLight

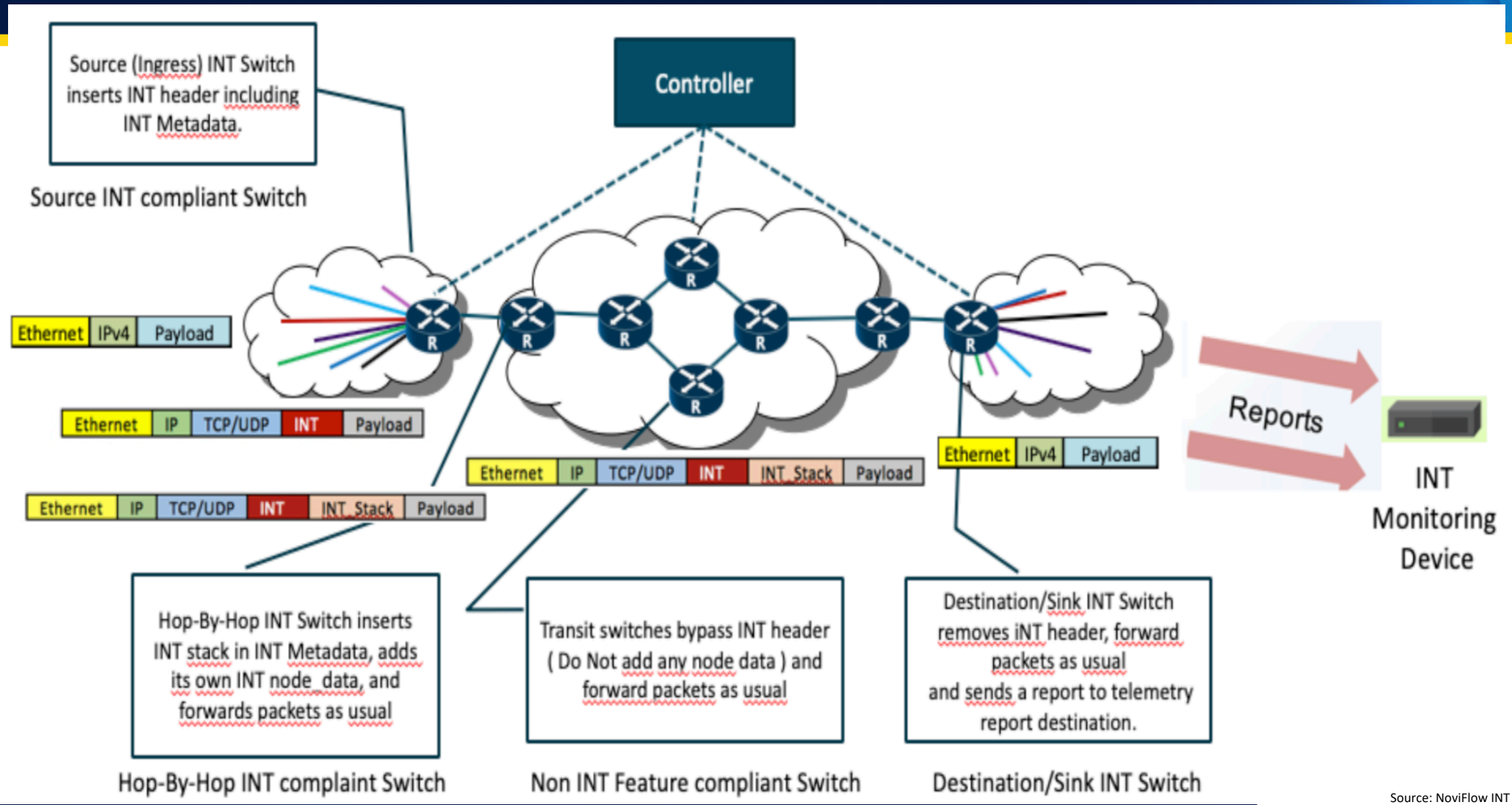
Jeronimo Bezerra - <jbezerra@fiu.edu>

Arun Paneri – Director of Engineering/NoviFlow

References

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

INT: How does it work?



Source: NoviFlow INT SetUp Guide