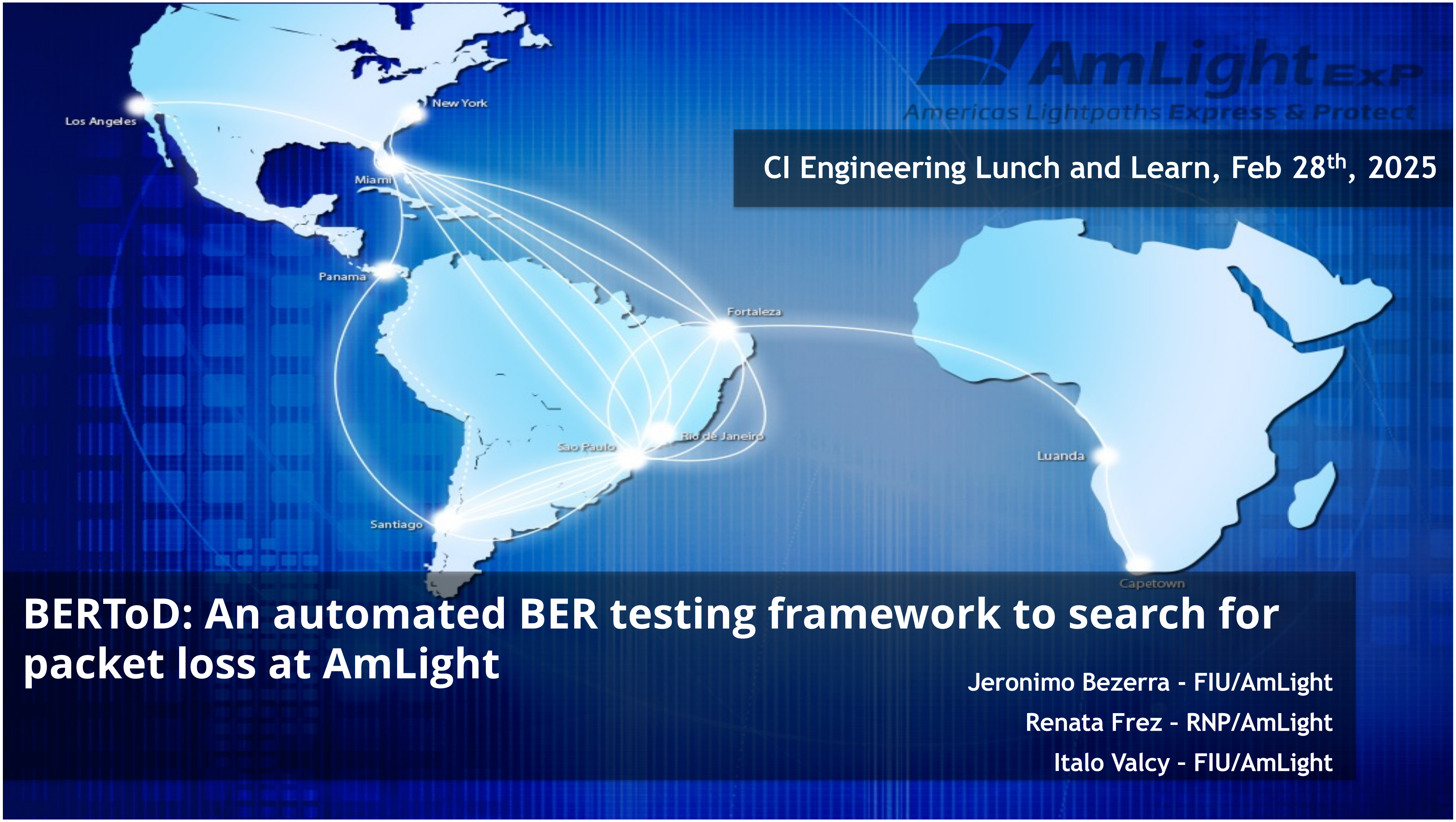
CI Engineering Lunch and Learn, Feb 28th, 2025



BERToD: An automated BER testing framework to search for packet loss at AmLight

Jeronimo Bezerra - FIU/AmLight

Renata Frez - RNP/AmLight

Italo Valcy - FIU/AmLight

Outline

- Motivation
- How BERToD works
- Lessons Learned
- Next Steps
- Conclusion

Disclaimer

- Packet vs Frame
  - Interchangeable in this presentation

- Packet Loss vs Packet Drop:
  - Drop: We drop our packets
    - Tail drop, QoS, blocking topologies, traffic engineering, small buffers
  - Loss: Someone/thing loses/corrupts our packets
    - Fiber cuts, power outages, damage components

- For this talk, our focus is on packet loss!

Motivation

- ***Data transfers over long-haul links suffer extra with packet loss***
  - 125 ms round-trip time (RTT) from Chile or Brazil to Jacksonville, FL
  - 131 ms RTT from Chile or Brazil to Atlanta
  - *A packet loss rate of $1x10^{-3}$ is enough to disrupt data movement workflows over 100+ ms RTTs.*
    - Reference: July 7th, 2023, CI Engineering Lunch and Learn *Handling Microbursts @ AmLight – Part 2 of 2*
    - *https://www.es.net/science-engagement/ci-engineering-lunch-and-learn-series/*

- Science applications are expecting better network performance
  - SLA-driven science drivers are demanding more granular loss detection measurement (1x10-9 or *1 out of 1,000,000,000* packets)
  - Such granularity is hard to achieve by just using standard hardware and software

- AmLight has grown in complexity in the last 5 years
  - Next slide

- Current solutions for packet loss detection have scalability, accuracy, or granularity limitations

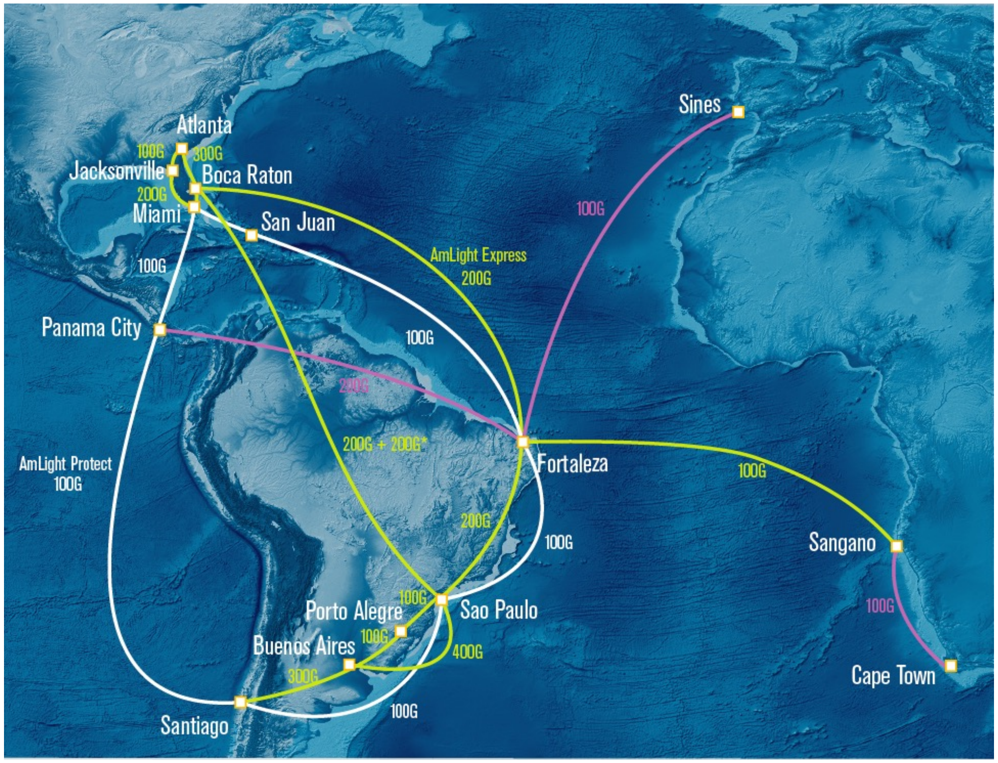
NSF IRNC: AmLight Network – 2020-2025

- A distributed academic exchange point built to enable collaboration among Latin America, Africa, and the U.S.
- Supported by NSF, OAC, and the IRNC program under award # OAC-2029283 for the 2021-2025
- Partnerships with R&E networks in the U.S., Latin America, Caribbean and Africa, built upon layers of trust and openness by sharing:
  - Infrastructure resources
  - Human resources

(NSF Award # OAC-2029283)

NSF IRNC: AmLight Network – 2020-2025

- 39x 100G links:
  - 2.1+ Tbps of international connectivity
  - AmLight will reach 5+ Tbps of total capacity^{2025}
  - Dark fiber, spectrum, waves, and lit services
- 9x Sites / 19x racks:
  - Miami, Boca Raton, Jacksonville, Sao Paulo, Fortaleza, Santiago, San Juan, Panama City, Cape Town, Atlanta, and Buenos Aires
- Network and Monitoring Devices:
  - 20x programmable switches and 5x Juniper routers
  - 10x 10G perfSonar nodes
  - 4x 100G servers
  - 4x In-band Network Telemetry (INT) collectors
    - ~10Mpps & 96TB of telemetry data per day



AmLight ExP
Americas Lightpaths Express & Protect

Current approaches to detect/isolate packet loss

- AmLight monitoring:
  - SDN counters, ICMP, SNMP, traceroute, optical telemetry metrics (polling-based), INT and JTI reports (streaming-based), perfSONAR measurements, and dedicated 100G nodes

- Even with all of those, fault isolation and mitigation are still challenging and have a high OPEX:
  - Evaluate results, correlate data, run extra tests, send field technicians to clean/replace suspicious components, steer traffic, and run again with different outcomes
    - Days/Weeks of work

- Existing packet generators/network testers are used in an ad-hoc fashion
  - Hardware-based granularity,
  - But manual configuration and with a learning curve to configure and read results

BERToD - Bit Error Rate Testing on Demand

- An automated packet loss detection framework that uses granular per-packet network telemetry (INT), SDN, a customized networking pipeline, and hardware-based packet generator to detect bit error rates as low as $1x10^{-12}$

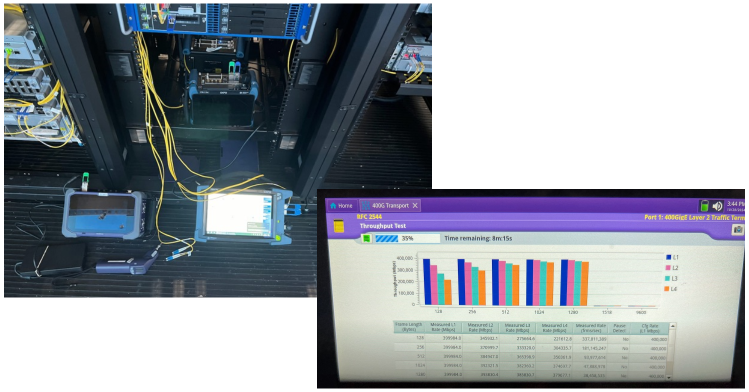
- BERToD leverages recent developments at AmLight:
  - Flexible forwarding rules provided by the SDN switches
  - Link and buffer utilization monitoring provided by In-band Network Telemetry (INT)
  - Topological data and dynamic service instantiation provided by the Kytos-ng SDN Controller

- *Near* deterministic results due to specialized network hardware being used end-to-end:
  - Highly accurate with granular results

Before we go any deeper, let’s go through some concepts and technologies

Traffic Generators/Network Testers

- Also known as network performance testers, they are appliances with specialized application and hardware focused on benchmarking network performance and reliability, as well as protocols, devices, and applications.
- Traffic Generators perform packet creation and processing entirely on specialized ASIC/FPGA to achieve deterministic results.
- Highly flexible in terms of packet creation: types of packets, size, headers, number of packets, packet rate/bandwidth, and even customizable payloads.
- Traffic generators have APIs to support remote integration.
- During SC23 and SC24, SCinet had access to EXFO, Viavi, and KeySight solutions to test the WAN links.



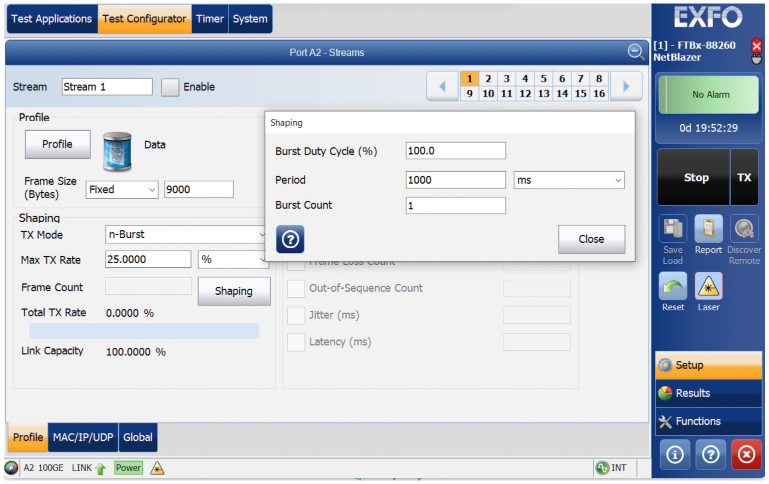
Frame Length (Bytes)	Measured L1 Rate (Mbps)	Measured L2 Rate (Mbps)	Measured L3 Rate (Mbps)	Measured L4 Rate (Mbps)	Measured Rate (frms/sec)	Pause Detect	Cfg Rate (L1 Mbps)
128	399984.0	345932.1	275664.6	221612.8	337,811,389	No	400,000
256	399984.0	370999.7	333320.0	304335.7	181,145,247	No	400,000
512	399984.0	384947.0	365398.9	350361.9	93,977,614	No	400,000
1024	399984.0	392321.5	382360.2	374697.7	47,888,978	No	400,000
1280	399984.0	393830.4	385830.7	379677.1	38,458,535	No	400,000

Traffic Generators/Network Testers

- AmLight has an EXFO FTB-1 NetBlazer with 4x100G interfaces
  - One 2x100G module for experimentation/testbed
  - One 2x100G module for BERToD/production

- AmLight created Python wrapper to use EXFO's SCPI API
  - SCPI (Standard Commands for Programmable Instruments) is no fun!

- BERToD uses two applications: EtherBERT and MonGen.
  - For EtherBERT, PRBS31 is supported (high accuracy)
  - Others are supported, such as RFC2544 and RFC6349



[1] - FTBx-88260 NetBlazer

No Alarm

0d 19:52:29

Stream: Stream 1 □ Enable

Profile

Frame Size (Bytes): Fixed 9000

Shaping

TX Mode: n-Burst

Max TX Rate: 25.0000 %

Frame Count:

Total TX Rate: 0.0000 %

Link Capacity: 100.0000 %

□ Frame Loss Count

□ Out-of-Sequence Count

□ Jitter (ms)

□ Latency (ms)

Shaping

Burst Duty Cycle (%): 100.0

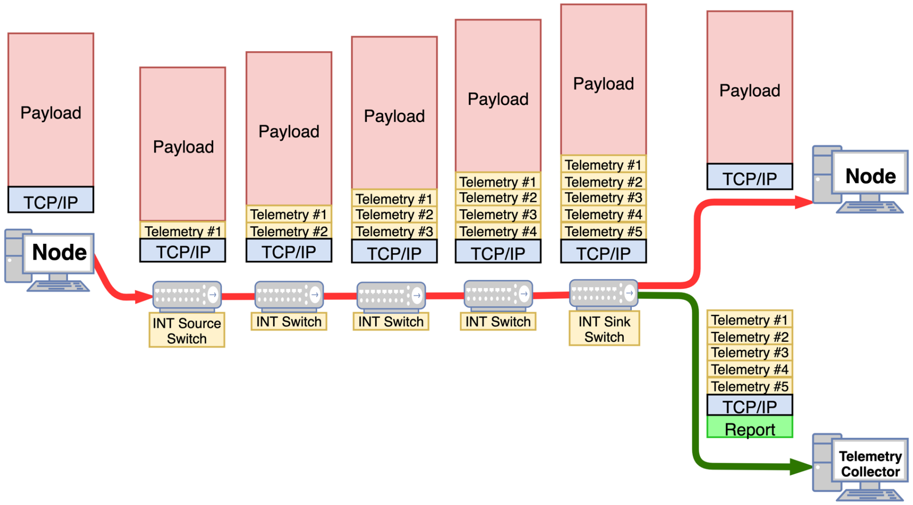
Period: 1000 ms

Burst Count: 1

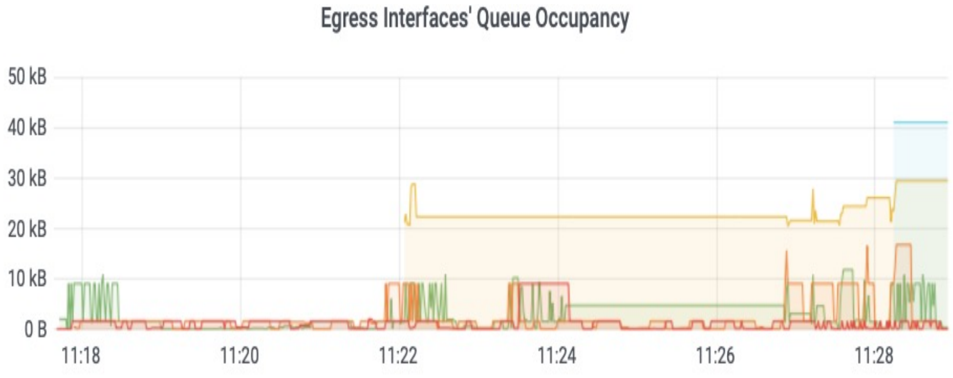
In-band Network Telemetry (INT)

- INT is a streaming telemetry solution based on P4 that records network telemetry data in the packet while the packet traverses a path between two points in the network

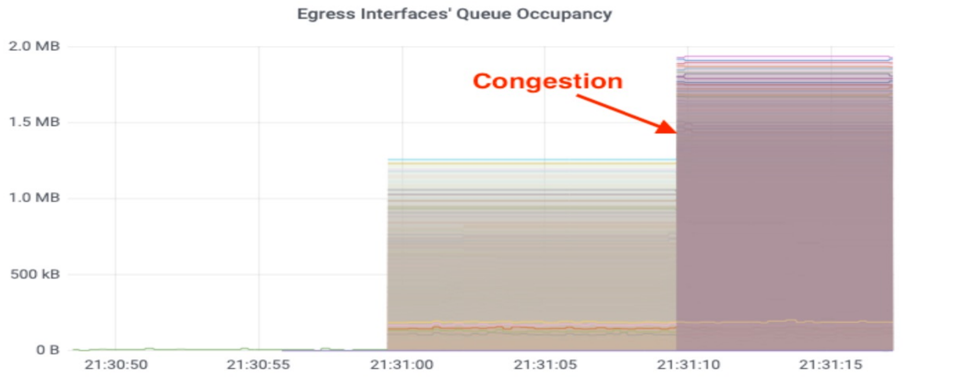
- Telemetry is exported directly from the Data Plane and the Control Plane is not affected:
  - Translation: you can track/monitor/evaluate EVERY single packet at line rate and in real time.



In-band Network Telemetry (INT)

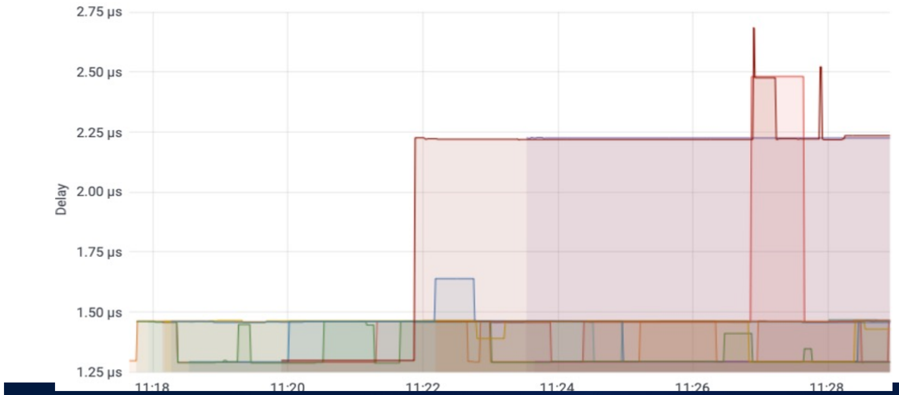


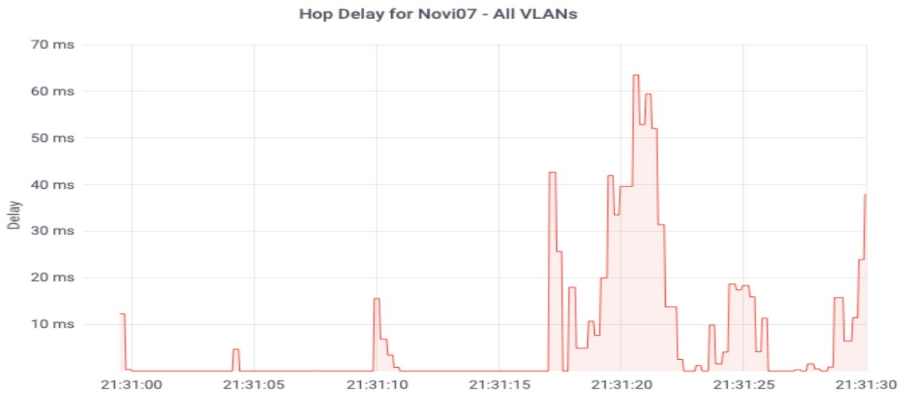
Egress Interfaces' Queue Occupancy



Egress Interfaces' Queue Occupancy

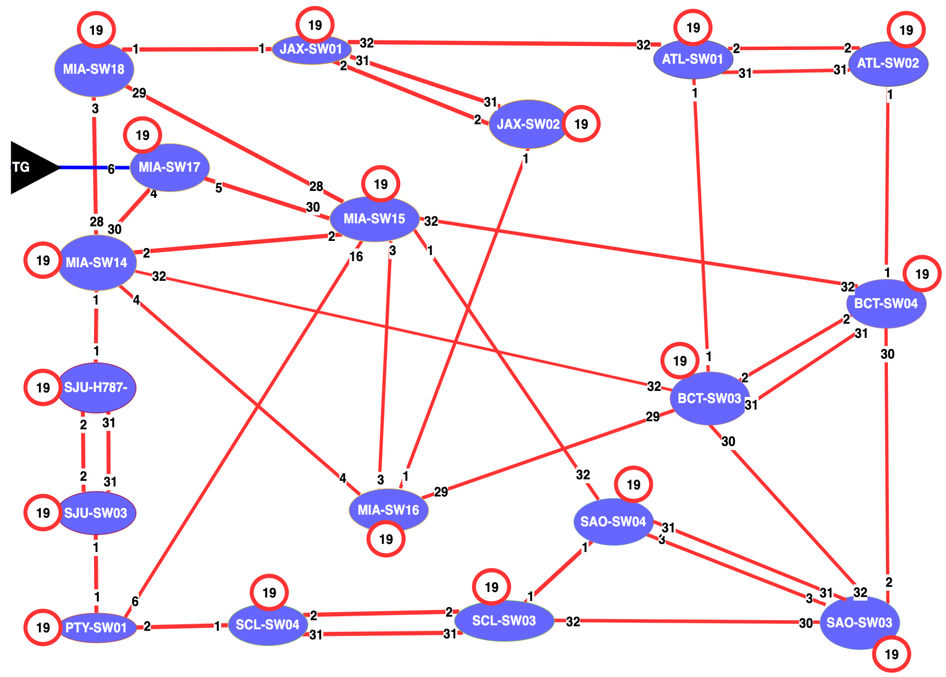
Congestion

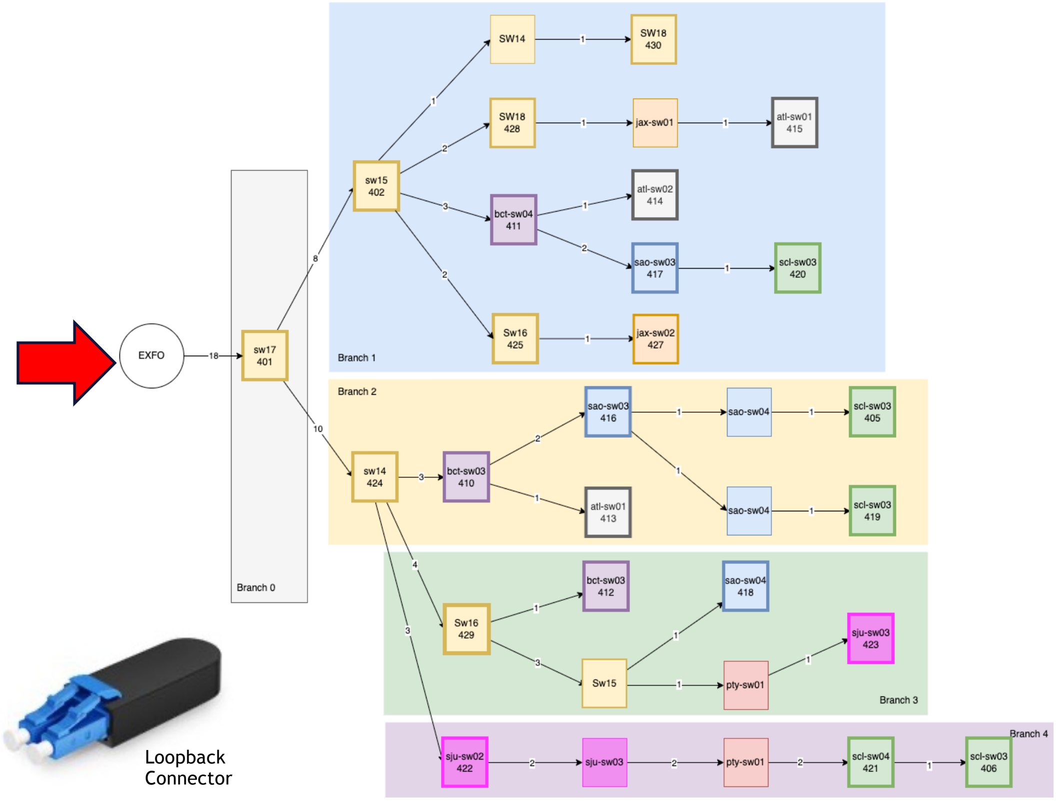




Hop Delay for Novi07 - All VLANs

SDN: Building paths over AmLight





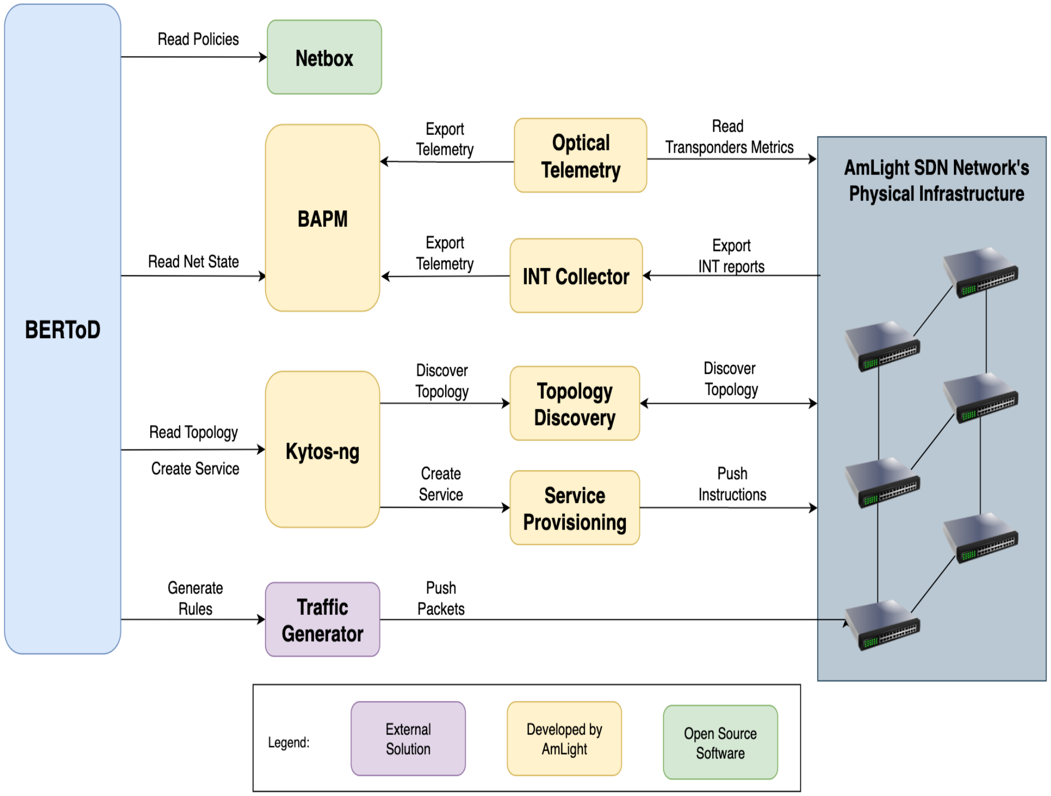
Loopback Connector

~~Before we go deeper, let’s go through some concepts and technologies~~

Ok, let’s move on.

BERToD – Components Explained

- Reads testing policies
  - Test duration, application, packet length, TCP/IP headers, number of packets, maximum bandwidth, scheduling, etc.
- Kytos-ng:
  - Generates the current network topology and instantiates testing paths over the network, from the packet generator to remote loops.
- BAPM (Behavior, Anomaly, and Performance Manager)
  - Generates the network state based on the telemetry sources available. For instance:
    - Identify the bandwidth available and buffer utilization for each interface based on the last 30 seconds of utilization.
- Packet Generator
  - Sends packets based on the policy (next slide)



BERToD

Read Policies → Netbox

Read Net State → BAPM

Optical Telemetry: Export Telemetry → BAPM; Read Transponders Metrics → AmLight SDN Network's Physical Infrastructure

INT Collector: Export Telemetry → BAPM; Export INT reports ← AmLight SDN Network's Physical Infrastructure

Read Topology / Create Service → Kytos-ng

Kytos-ng: Discover Topology → Topology Discovery; Create Service → Service Provisioning

Topology Discovery: Discover Topology ↔ AmLight SDN Network's Physical Infrastructure

Service Provisioning: Push Instructions → AmLight SDN Network's Physical Infrastructure

Generate Rules → Traffic Generator

Traffic Generator: Push Packets → AmLight SDN Network's Physical Infrastructure

AmLight SDN Network's Physical Infrastructure

Legend: External Solution | Developed by AmLight | Open Source Software

BERToD - Bit Error Rate Test on Demand [2]

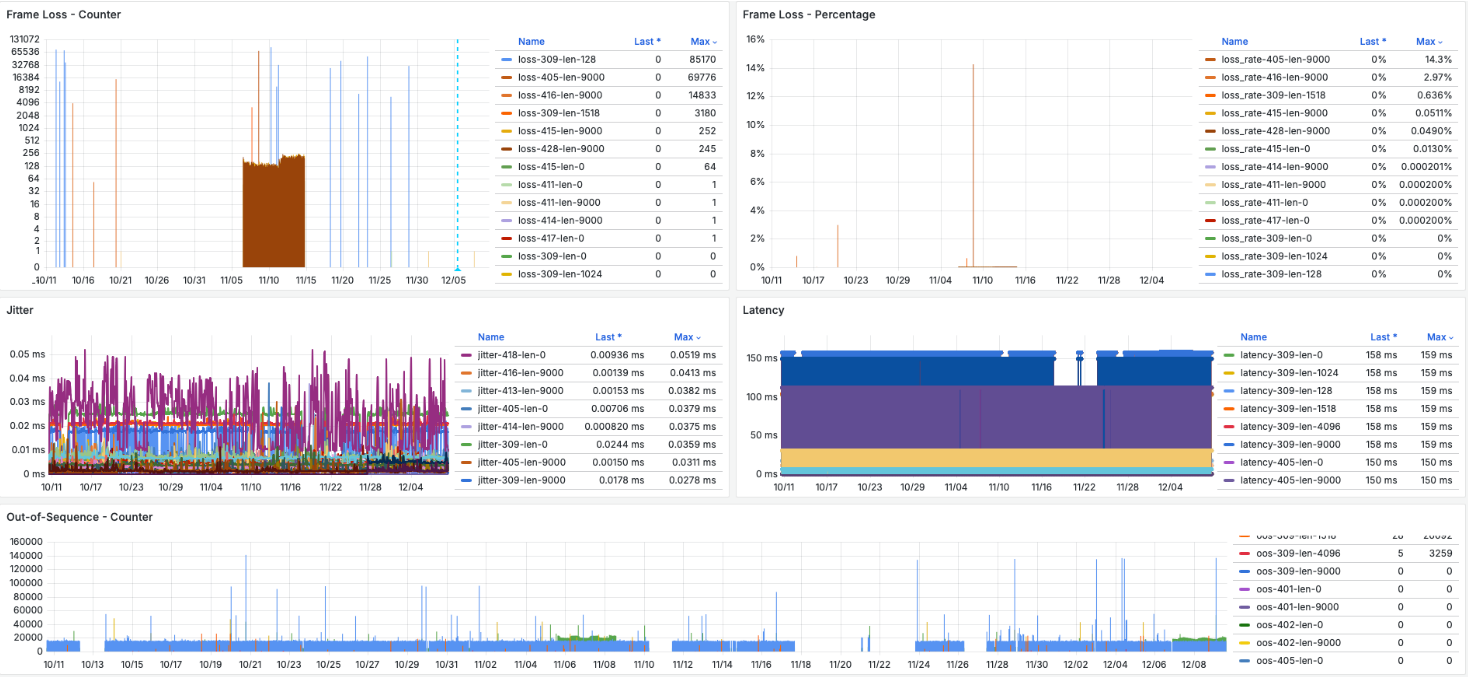
- Test every possible link every 30 min:
  - Latency, jitter, frame loss, and out-of-sequence tests
  - Multiple frame sizes: 68, 256, 512, 1024, 1518, 9000 bytes
  - Each test runs for up to 10 seconds, and we send up to 500,000 frames
    - In case a test fails, run again with a multiplier metric (for instance, 3)
  - Choice for max bandwidth comes from BAPM
    - Up to 50% of the available bandwidth based on the last 30 seconds (and up to 40 Gbps)

- Displaying results:
  - Last hour, Last 7 days, heatmap, and text outputs

- Grafana Annotations are used to document known topology events and actions to help correlate events.

BERToD - Granular Individual Results

- Using Grafana to plot each test’s loss, jitter, latency, and out-of-sequence
- Great way to understand the last 24 hours
- Filters available to visualize test results based on frame size and individual paths
- Not great for correlating events



Frame Loss - Counter

Name	Last *	Max ⌄
loss-309-len-128	0	85170
loss-405-len-9000	0	69776
loss-416-len-9000	0	14833
loss-309-len-1518	0	3180
loss-415-len-9000	0	252
loss-428-len-9000	0	245
loss-415-len-0	0	64
loss-411-len-0	0	1
loss-411-len-9000	0	1
loss-414-len-9000	0	1
loss-417-len-0	0	1
loss-309-len-0	0	0
loss-309-len-1024	0	0

Frame Loss - Percentage

Name	Last *	Max ⌄
loss_rate-405-len-9000	0%	14.3%
loss_rate-416-len-9000	0%	2.97%
loss_rate-309-len-1518	0%	0.636%
loss_rate-415-len-9000	0%	0.0511%
loss_rate-428-len-9000	0%	0.0490%
loss_rate-415-len-0	0%	0.0130%
loss_rate-414-len-9000	0%	0.000201%
loss_rate-411-len-9000	0%	0.000200%
loss_rate-411-len-0	0%	0.000200%
loss_rate-417-len-0	0%	0.000200%
loss_rate-309-len-0	0%	0%
loss_rate-309-len-1024	0%	0%
loss_rate-309-len-128	0%	0%

Jitter

Name	Last *	Max ⌄
jitter-418-len-0	0.00936 ms	0.0519 ms
jitter-416-len-9000	0.00139 ms	0.0413 ms
jitter-413-len-9000	0.00153 ms	0.0382 ms
jitter-405-len-0	0.00706 ms	0.0379 ms
jitter-414-len-9000	0.000820 ms	0.0375 ms
jitter-309-len-0	0.0244 ms	0.0359 ms
jitter-405-len-9000	0.00150 ms	0.0311 ms
jitter-309-len-9000	0.0178 ms	0.0278 ms

Latency

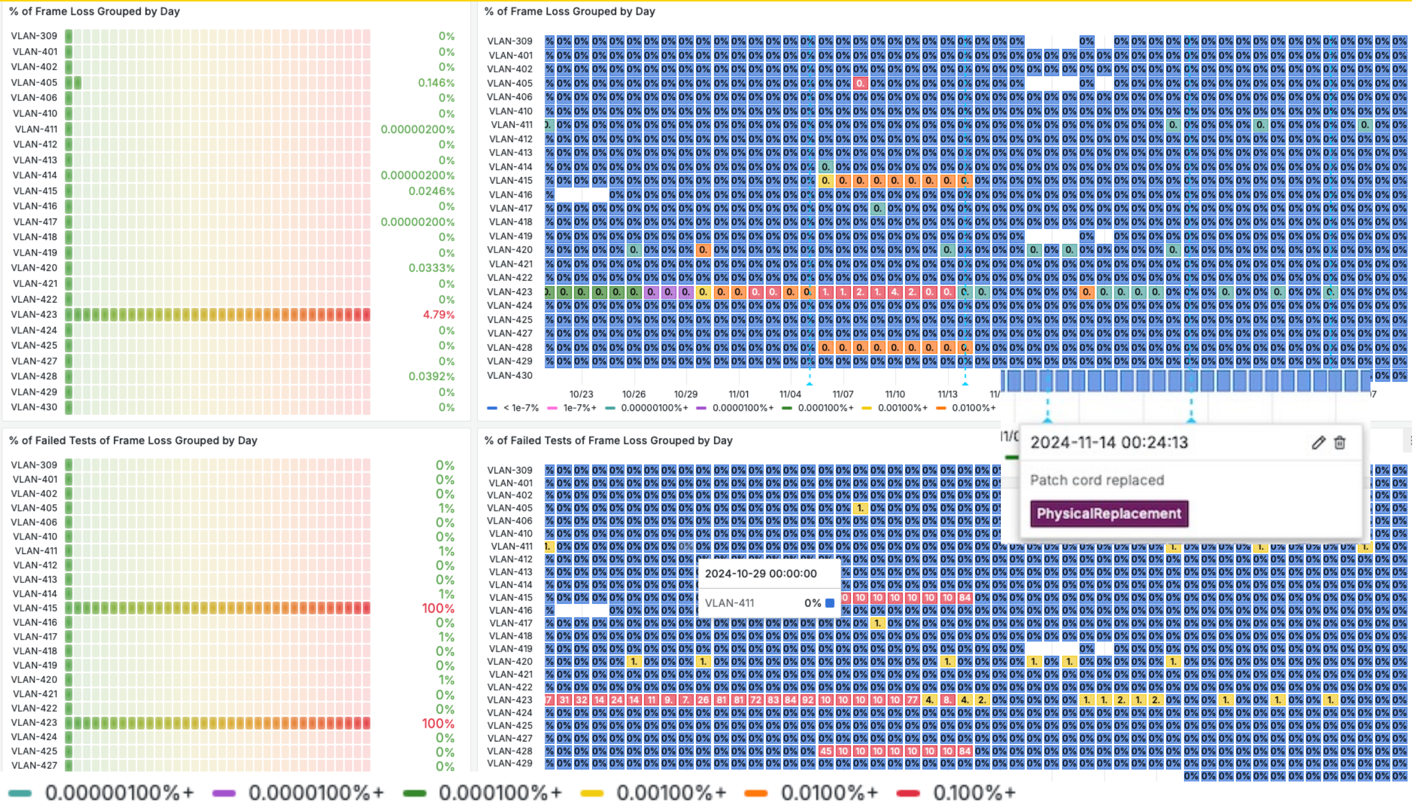
Name	Last *	Max ⌄
latency-309-len-0	158 ms	159 ms
latency-309-len-1024	158 ms	159 ms
latency-309-len-128	158 ms	159 ms
latency-309-len-1518	158 ms	159 ms
latency-309-len-4096	158 ms	159 ms
latency-309-len-9000	158 ms	159 ms
latency-405-len-0	150 ms	150 ms
latency-405-len-9000	150 ms	150 ms

Out-of-Sequence - Counter

oos-309-len-1518	28	20092
oos-309-len-4096	5	3259
oos-309-len-9000	0	0
oos-401-len-0	0	0
oos-401-len-9000	0	0
oos-402-len-0	0	0
oos-402-len-9000	0	0
oos-405-len-0	0	0

BERToD - Historical Results

- Using Grafana to plot each test's loss per day
- Great way to correlate events and identify patterns
- Filters available to visualize test results based on frame size and individual paths
- Used with annotations to add context



Lessons Learned

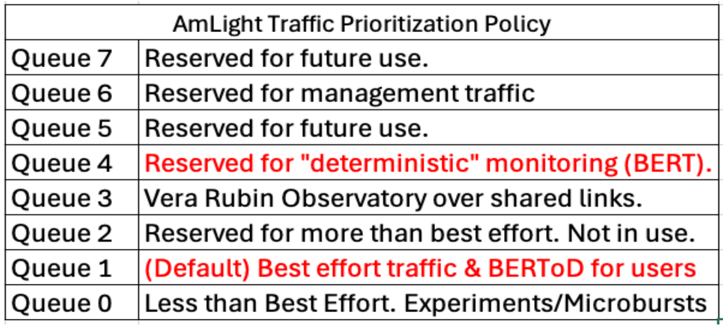
- Testing infrastructure vs testing user experience
  - To achieve deterministic results, network resources must be fully available.
  - Even with buffer occupancy monitoring, traffic engineering at AmLight had to be enhanced to avoid new drops ➔

- How to monitor the actual user experience?
  - Using perfSONAR and BERToD in the same queue as users

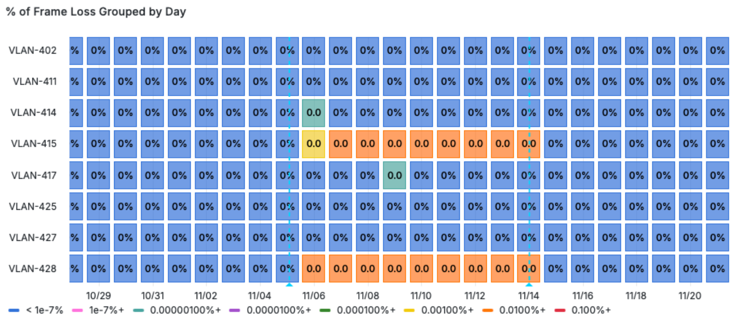
- “*There is no such innocent maintenance at the datacenter*”.
  - Mishandled patch cords are the main reason for sudden spikes of errors ➔

- Dirty fiber/connector is the main reason for discrete errors (<0.0001%)

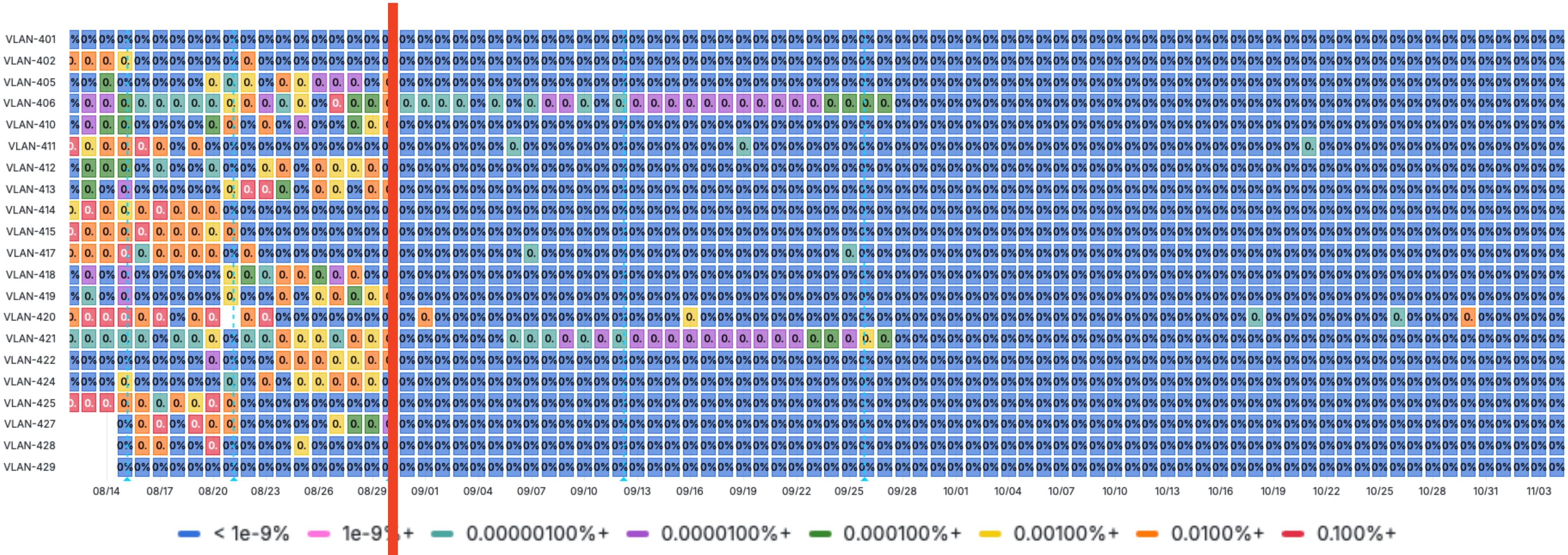
- Some vendors have weird policies, and small frames are delivered out of sequence (under investigation)



AmLight Traffic Prioritization Policy	
Queue 7	Reserved for future use.
Queue 6	Reserved for management traffic
Queue 5	Reserved for future use.
Queue 4	Reserved for "deterministic" monitoring (BERT).
Queue 3	Vera Rubin Observatory over shared links.
Queue 2	Reserved for more than best effort. Not in use.
Queue 1	(Default) Best effort traffic & BERToD for users
Queue 0	Less than Best Effort. Experiments/Microbursts



Lessons Learned – Testing Infra vs User Experience



Tests using Priority Queue BE

Tests using Priority Queue 4

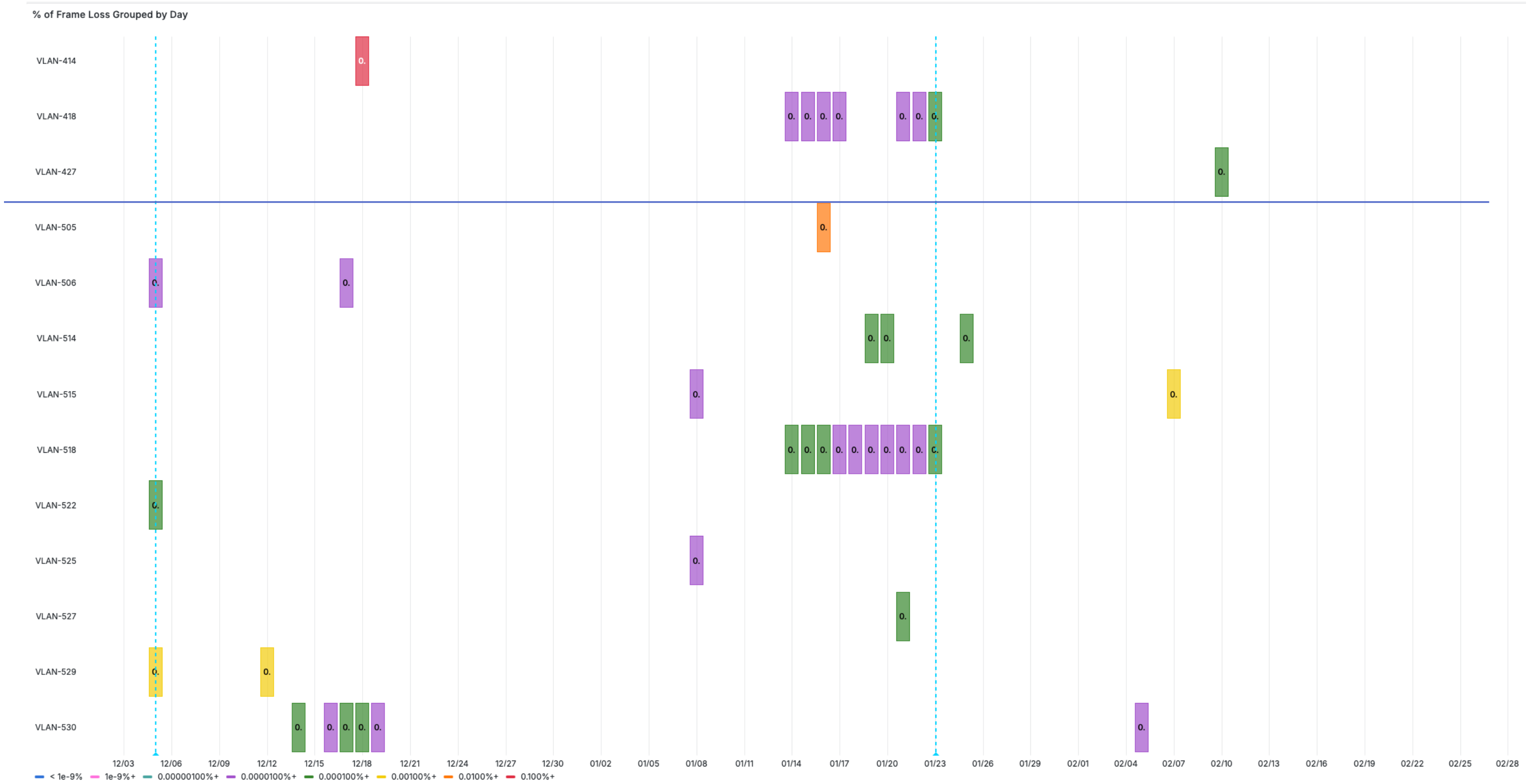
Next steps: Re-Testing after a link flap

- Integration with Kytos-ng SDN Controller to test links after **each link flap**:
  - The goal is to evaluate if the link is clean after a maintenance/repair before using it again!
    - After a link flaps, the SDN controller waits up to 2 min to confirm the link is stable and then initiates the quarantine mode
    - BERToD is notified of the quarantine and starts a 5 min test
    - If results are clean, BERToD sets the link as operational/ready
    - The SDN controller then makes the link available to all applications

Next steps: Automate fault isolation

Automate the fault isolation process using all data sources available

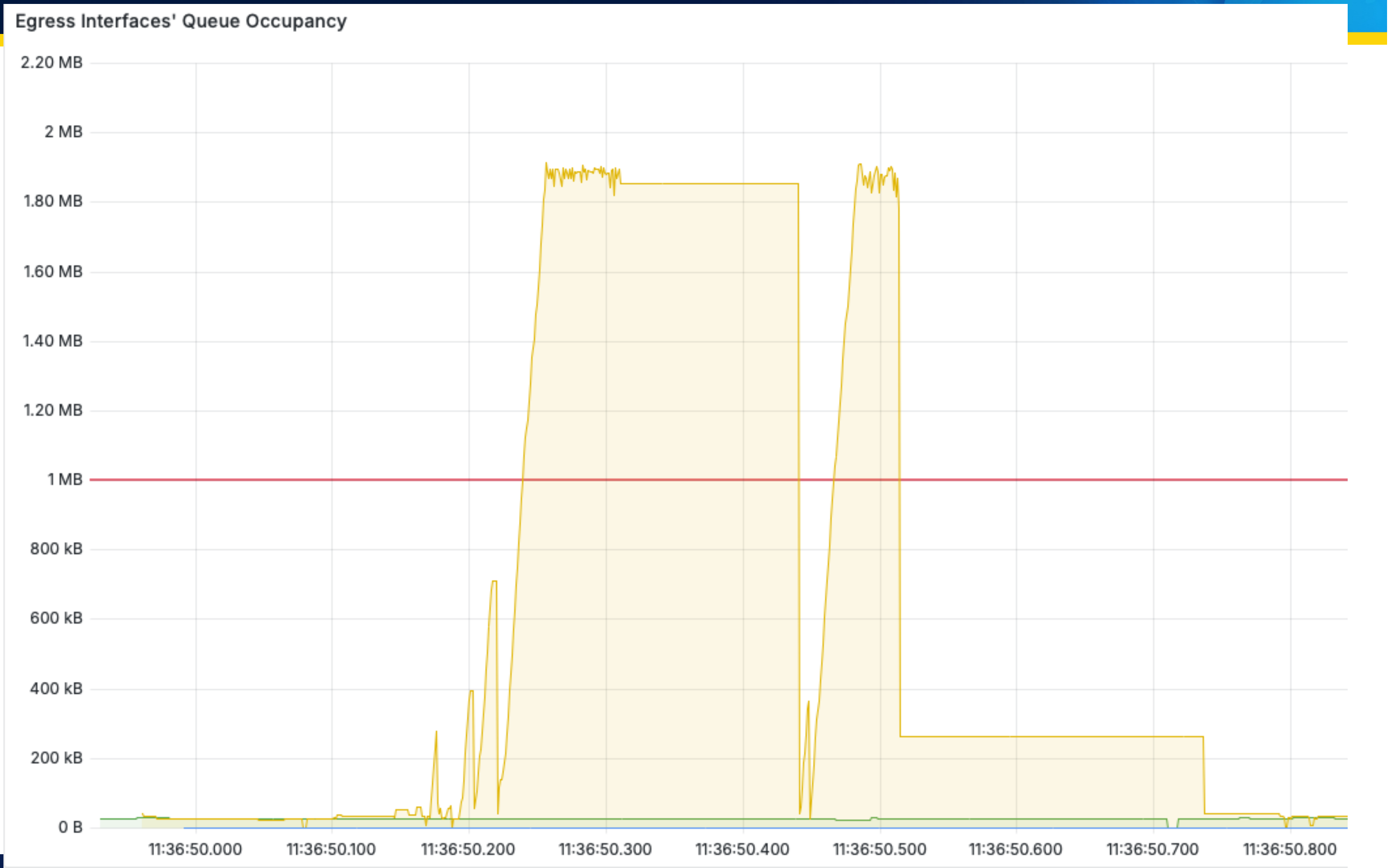
- SDN logs, topology changes, EVC optimizations, events/demos, optical monitoring, and visits to the data center.
- An AI/ML researcher is playing with our testing dataset and log entries to isolate issues.



AmLight ExP
Americas Lightpaths Express & Protect

Next steps: Automate fault isolation [2]

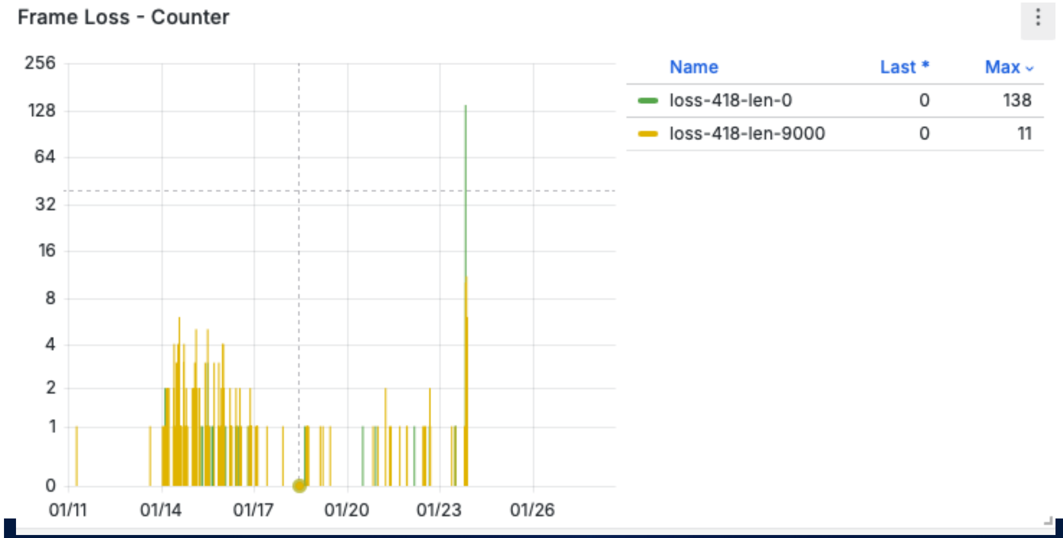
How to create a correlation between a, for instance, high utilization queue and the BERToD test results?



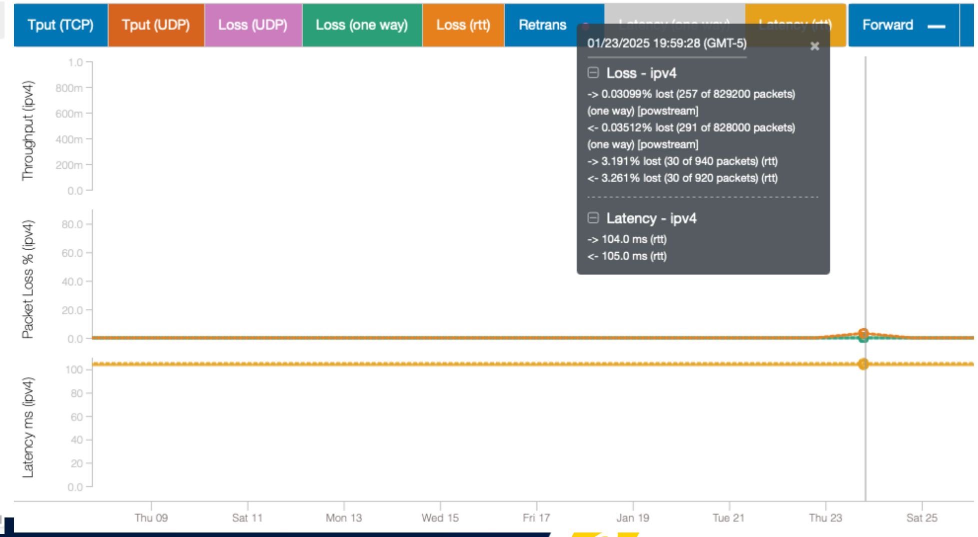
AmLight ExP
Americas Lightpaths Express & Protect

Next Steps: Comparing results with perfSONAR

- We are still learning how to compare the results provided by BERToD and perfSONAR.
- An example to share:
  - BERToD started reporting consistent errors when testing the leased link between Fortaleza, Brazil and Miami.
  - The issue started on Jan 14th, 2025. After a maintenance on the Jan 24th, the errors disappeared (unknown cause).
  - Although the errors reported were around 0.0002%, we believe perfSONAR should have detected something before the 24th.



Name	Last *	Max
loss-418-len-0	0	138
loss-418-len-9000	0	11



01/23/2025 19:59:28 (GMT-5)

Loss - ipv4
-> 0.03099% lost (257 of 829200 packets) (one way) [powstream]
<- 0.03512% lost (291 of 828000 packets) (one way) [powstream]
-> 3.191% lost (30 of 940 packets) (rtt)
<- 3.261% lost (30 of 920 packets) (rtt)

Latency - ipv4
-> 104.0 ms (rtt)
<- 105.0 ms (rtt)

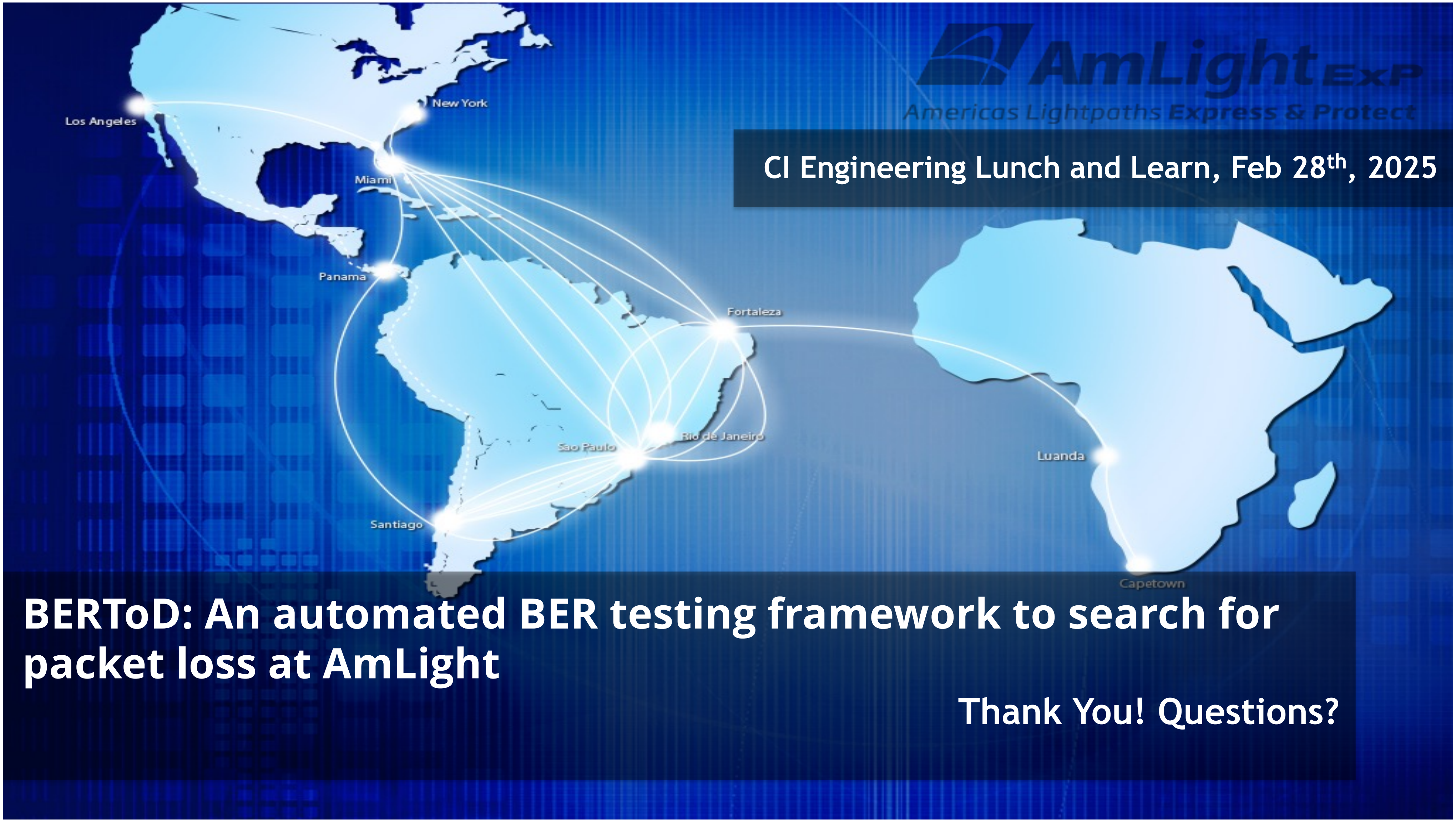
AmLight ExP
Americas Lightpaths Express & Protect

Conclusion

- BERToD is a fantastic addition to the network monitoring portfolio thanks to the hardware-based traffic generator and enhanced network telemetry provided by the AmLight SDN solution.
  - Production since September 2024. Used daily by AmLight OPS.

- Having a hardware-based traffic generator enables quick testing with extreme accuracy
  - Helps us follow the demands of our SLA-driven science drivers

- BERToD is a great complement to perfSONAR @ AmLight.
  - While perfSONAR allows AmLight to test applications and protocols with excellent per-direction visibility, BERToD provides extreme performance visibility for applications over ultra-long paths where any packet loss causes damage.



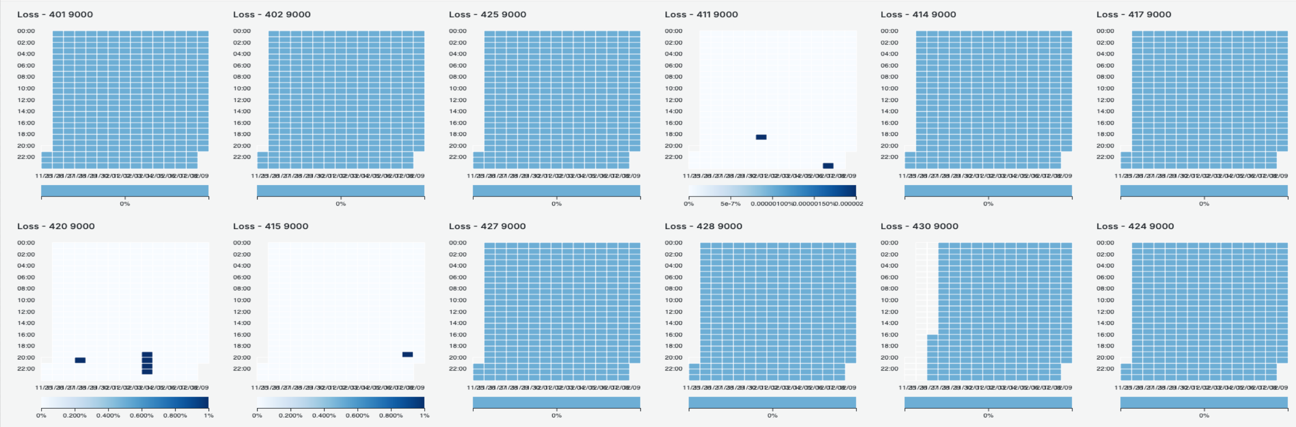
CI Engineering Lunch and Learn, Feb 28th, 2025

BERToD: An automated BER testing framework to search for packet loss at AmLight

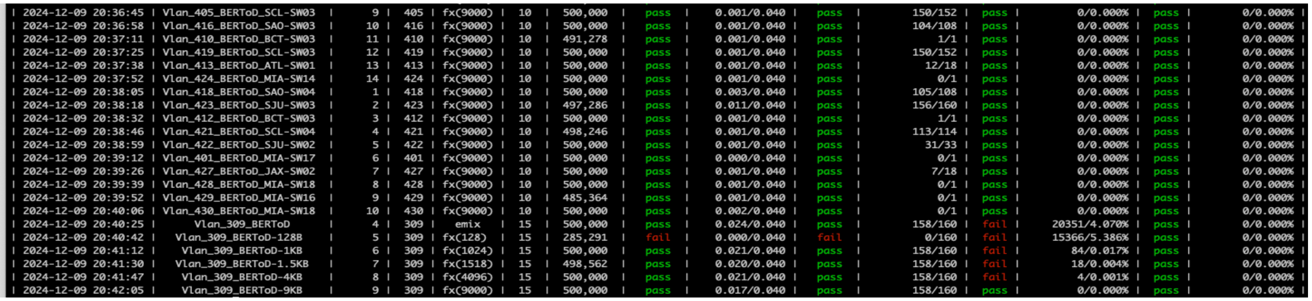
Thank You! Questions?

BERToD - Bit Error Rate Test on Demand [5]

- Per-Hour Heatmap visualization created to help identify patterns across tests



- Command-line to access full results and test configuration



```
| 2024-12-09 20:36:45 | Vlan_405_BERToD_SCL-SW03 |      9 | 405 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |     150/152 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:36:58 | Vlan_416_BERToD_SAO-SW03 |     10 | 416 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |     104/108 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:37:11 | Vlan_410_BERToD_BCT-SW03 |     11 | 410 | fx(9000) | 10 | 491,278 |  pass |  0.001/0.040 |  pass |         1/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:37:25 | Vlan_419_BERToD_SCL-SW03 |     12 | 419 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |     150/152 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:37:38 | Vlan_413_BERToD_ATL-SW01 |     13 | 413 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |       12/18 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:37:52 | Vlan_424_BERToD_MIA-SW14 |     14 | 424 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |         0/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:38:05 | Vlan_418_BERToD_SAO-SW04 |      1 | 418 | fx(9000) | 10 | 500,000 |  pass |  0.003/0.040 |  pass |     105/108 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:38:18 | Vlan_423_BERToD_SJU-SW03 |      2 | 423 | fx(9000) | 10 | 497,286 |  pass |  0.011/0.040 |  pass |     156/160 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:38:32 | Vlan_412_BERToD_BCT-SW03 |      3 | 412 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |         1/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:38:46 | Vlan_421_BERToD_SCL-SW04 |      4 | 421 | fx(9000) | 10 | 498,246 |  pass |  0.001/0.040 |  pass |     113/114 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:38:59 | Vlan_422_BERToD_SJU-SW02 |      5 | 422 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |       31/33 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:39:12 | Vlan_401_BERToD_MIA-SW17 |      6 | 401 | fx(9000) | 10 | 500,000 |  pass |  0.000/0.040 |  pass |         0/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:39:26 | Vlan_427_BERToD_JAX-SW02 |      7 | 427 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |        7/18 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:39:39 | Vlan_428_BERToD_MIA-SW18 |      8 | 428 | fx(9000) | 10 | 500,000 |  pass |  0.001/0.040 |  pass |         0/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:39:52 | Vlan_429_BERToD_MIA-SW16 |      9 | 429 | fx(9000) | 10 | 485,364 |  pass |  0.001/0.040 |  pass |         0/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:40:06 | Vlan_430_BERToD_MIA-SW18 |     10 | 430 | fx(9000) | 10 | 500,000 |  pass |  0.002/0.040 |  pass |         0/1 | pass |      0/0.000% | pass |      0/0.000% |
| 2024-12-09 20:40:25 |     Vlan_309_BERToD      |      4 | 309 |   emix   | 15 | 500,000 |  pass |  0.024/0.040 |  pass |     158/160 | fail |  20351/4.070% | pass |      0/0.000% |
| 2024-12-09 20:40:42 |   Vlan_309_BERToD-128B   |      5 | 309 | fx(128)  | 15 | 285,291 |  fail |  0.000/0.040 |  fail |       0/160 | fail |  15366/5.386% | pass |      0/0.000% |
| 2024-12-09 20:41:12 |    Vlan_309_BERToD-1KB   |      6 | 309 | fx(1024) | 15 | 500,000 |  pass |  0.021/0.040 |  pass |     158/160 | fail |     84/0.017% | pass |      0/0.000% |
| 2024-12-09 20:41:30 |   Vlan_309_BERToD-1.5KB  |      7 | 309 | fx(1518) | 15 | 498,562 |  pass |  0.020/0.040 |  pass |     158/160 | fail |     18/0.004% | pass |      0/0.000% |
| 2024-12-09 20:41:47 |    Vlan_309_BERToD-4KB   |      8 | 309 | fx(4096) | 15 | 500,000 |  pass |  0.021/0.040 |  pass |     158/160 | fail |      4/0.001% | pass |      0/0.000% |
| 2024-12-09 20:42:05 |    Vlan_309_BERToD-9KB   |      9 | 309 | fx(9000) | 15 | 500,000 |  pass |  0.017/0.040 |  pass |     158/160 | pass |      0/0.000% | pass |      0/0.000% |
```