New York
Los Angeles
Miami
Panama
Fortaleza
Rio de Janeiro
Sao Paulo
Santiago

AmLight ExP
Americas Lightpaths Express & Protect

Innovating the Network for Data-Intensive Science (INDIS) Workshop

Nov 15th, 2021

Deploying per-packet telemetry in a long-haul network: the AmLight use case

Jeronimo Bezerra <jab@amlight.net>

Outline

- Why monitoring every packet?
- What is In-band Network Telemetry (INT)?
- How is AmLight using INT?
- Moving INT to production at AmLight
- Some Results
- Conclusion

Why is AmLight interested in monitoring every packet?

- The need:
  - Monitoring real-time SLA-driven applications
  - Detecting [micro] bursts and events impacting AmLight’s network functions and applications

- The challenge:
  - SNMP/OpenFlow counters: polling data is not recommended in a sub-15s interval.
  - Sampling technologies: export data after a few seconds.
  - Packet sniffing at 100G: high CAPEX and OPEX costs.
  - Streaming telemetry solutions: share summaries every second[s].

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, the Control Plane is not affected:
  - Translating: *you can track/monitor/evaluate EVERY single packet at line rate and in real time.*

- Examples of telemetry information added:
  - Timestamp, ingress port, egress port, queue buffer utilization, sequence #, and many others

How does In-band Network Telemetry (INT) work?

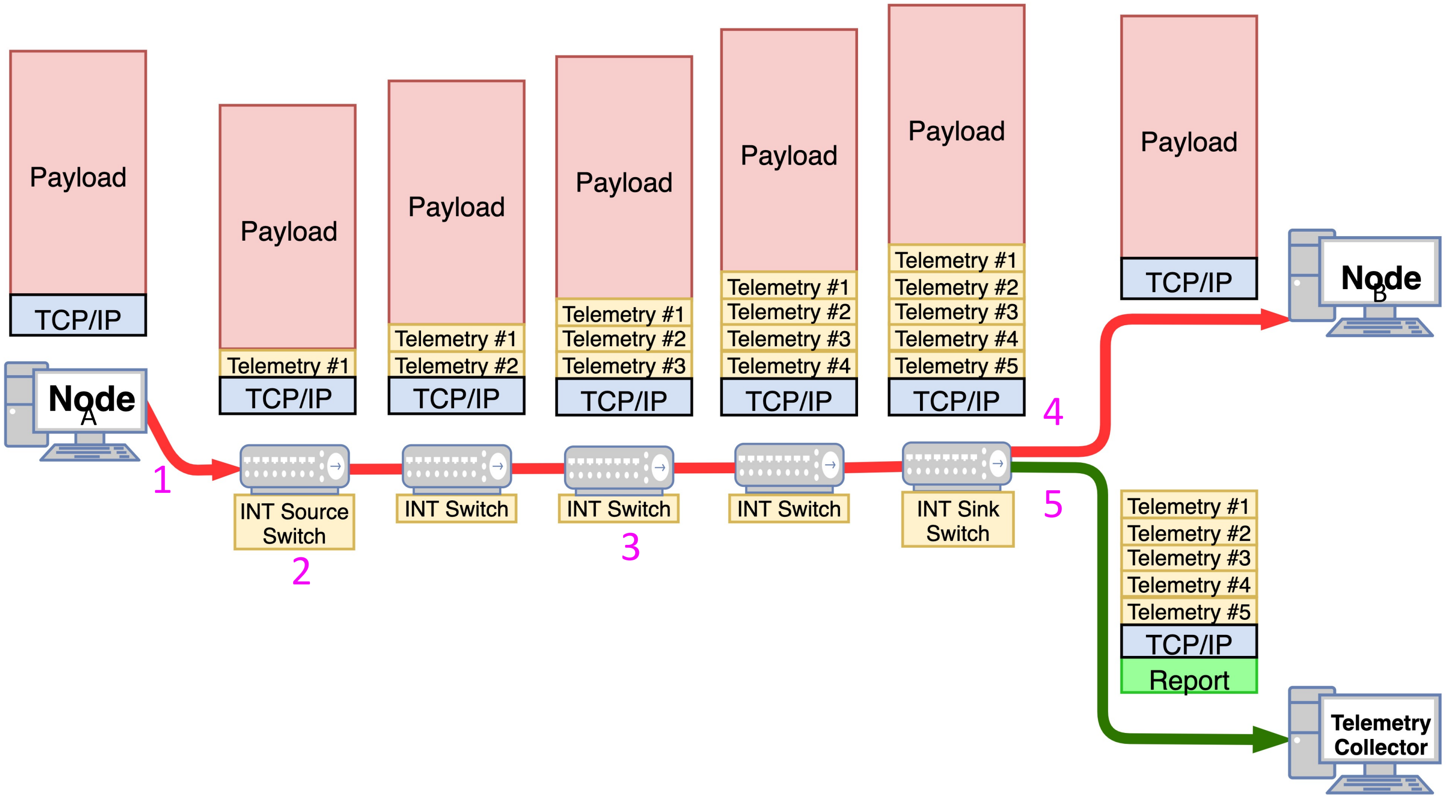
1 – User sends a TCP or UDP packet unaware of INT

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

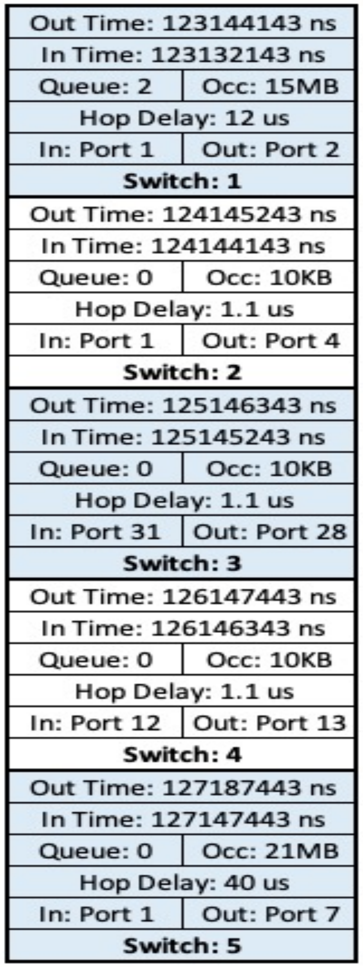
5 – Last switch (INT Sink Switch) forwards the 1:1 telemetry report to the Telemetry Collector



AmLight ExP
Americas Lightpaths Express & Protect

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

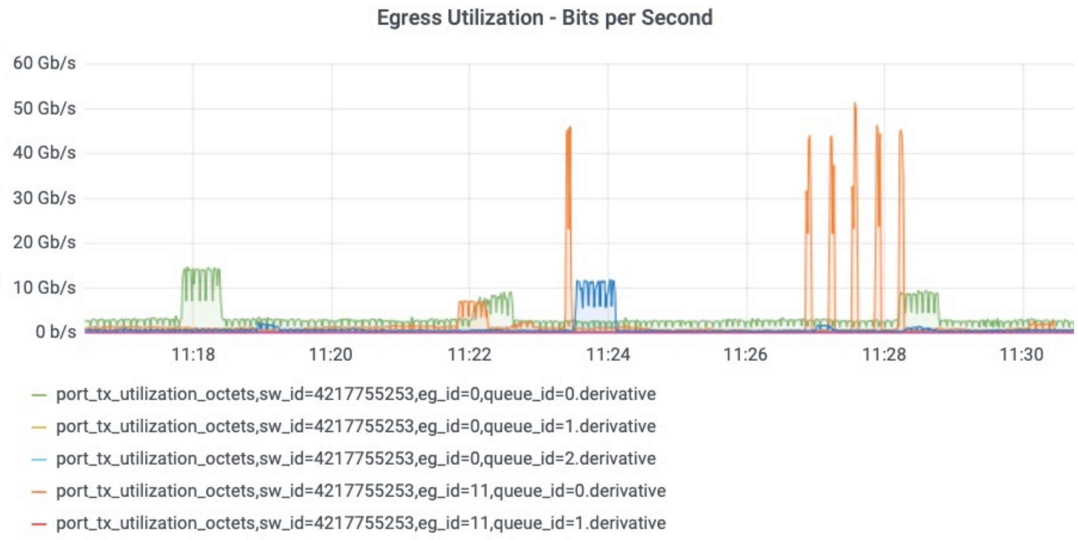
- The AmLight INT switches leverage the Tofino chip to collect:
  - Per switch:
    - Switch ID
    - Ingress port
    - Egress port
    - Ingress timestamp
    - Egress timestamp
    - Egress queue ID
    - Egress queue occupancy
  - Per telemetry report:
    - Report timestamp
    - Report sequence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



Out Time: 123144143 ns	
In Time: 123132143 ns	
Queue: 2	Occ: 15MB
Hop Delay: 12 us	
In: Port 1	Out: Port 2
Switch: 1	
Out Time: 124145243 ns	
In Time: 124144143 ns	
Queue: 0	Occ: 10KB
Hop Delay: 1.1 us	
In: Port 1	Out: Port 4
Switch: 2	
Out Time: 125146343 ns	
In Time: 125145243 ns	
Queue: 0	Occ: 10KB
Hop Delay: 1.1 us	
In: Port 31	Out: Port 28
Switch: 3	
Out Time: 126147443 ns	
In Time: 126146343 ns	
Queue: 0	Occ: 10KB
Hop Delay: 1.1 us	
In: Port 12	Out: Port 13
Switch: 4	
Out Time: 127187443 ns	
In Time: 127147443 ns	
Queue: 0	Occ: 21MB
Hop Delay: 40 us	
In: Port 1	Out: Port 7
Switch: 5	

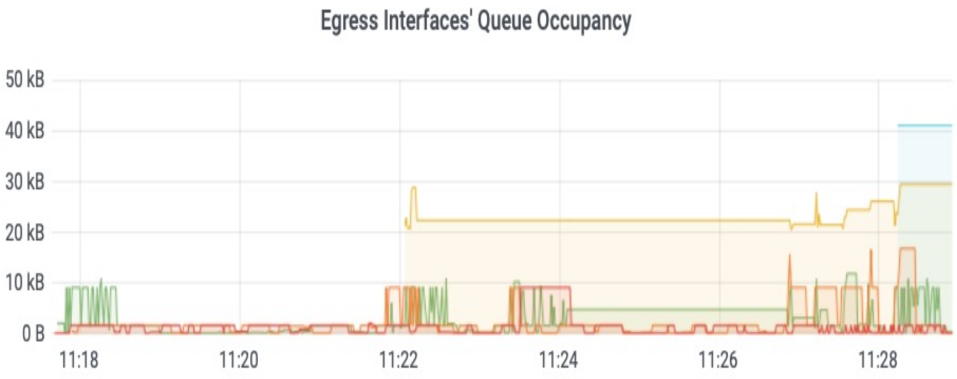
What INT metadata is being used and how? [2]

- Instantaneous Ingress and Egress Interface utilization
  - 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

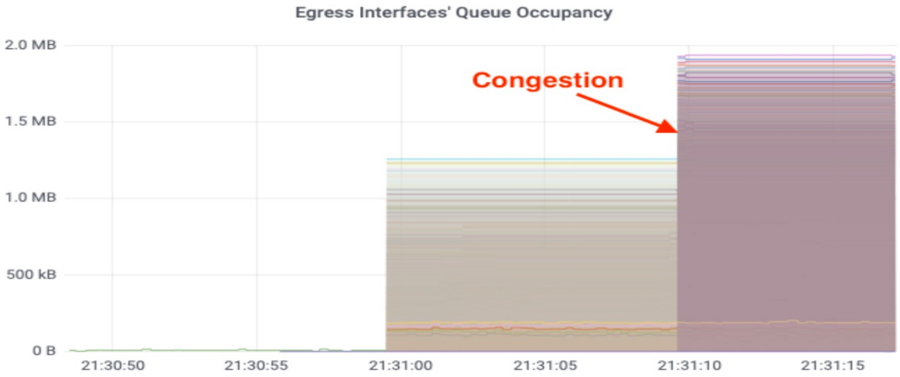


What INT metadata is being used and how? [3]

- Instantaneous Egress Interface Queue utilization (or buffer)
  - AmLight monitors every queue of every interface of every switch:
    - Useful for evaluating QoS policies
    - Useful for detecting sources of packet drops



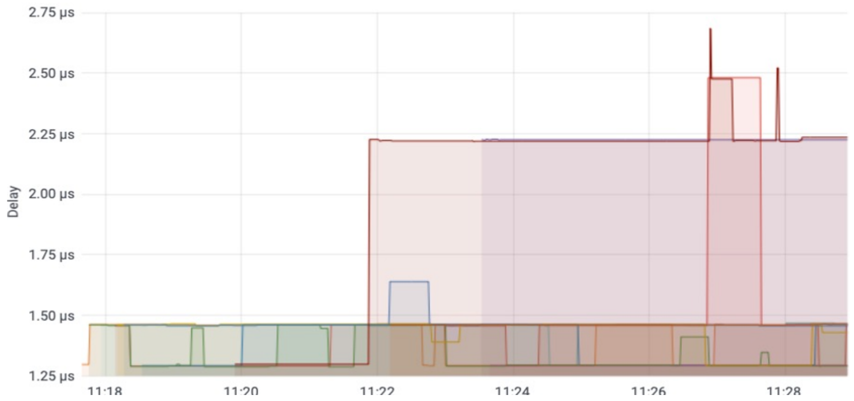
Average Buffer Utilization



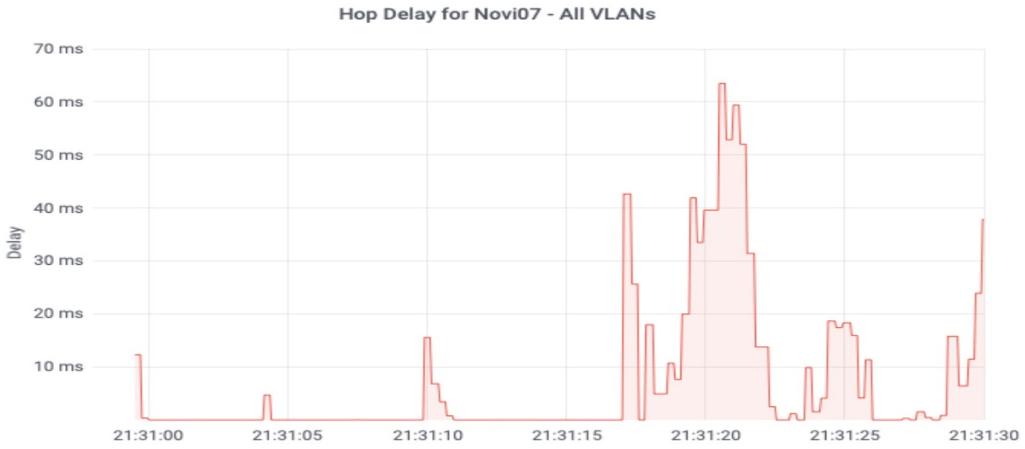
Under-Congestion Buffers

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

- Sources of jitter
  - 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)



Average Hop Delay (in microseconds)



Hop Delay under congestion (in milliseconds)

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

- **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

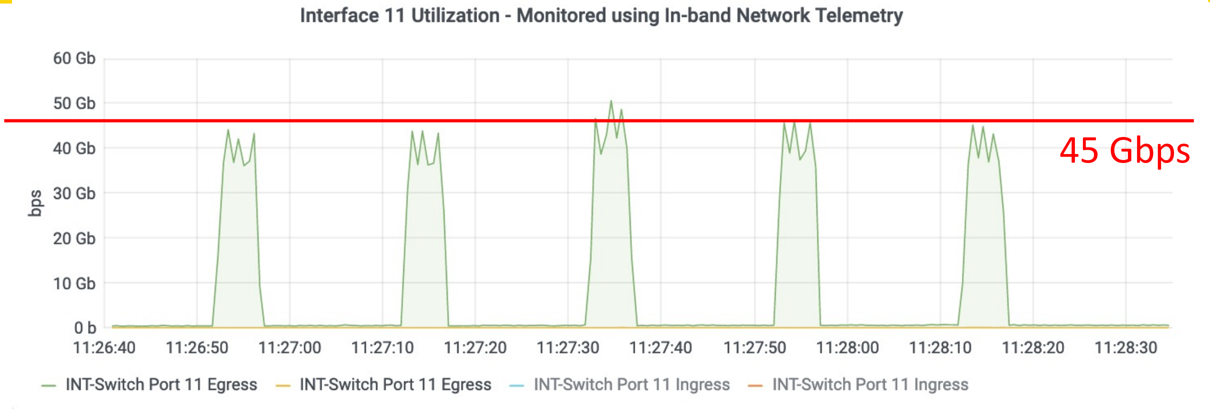
Use Case: Mitigating [malicious] [micro] bursts

- 5 data transfers/bursts of 40-50Gbps for 5 seconds.

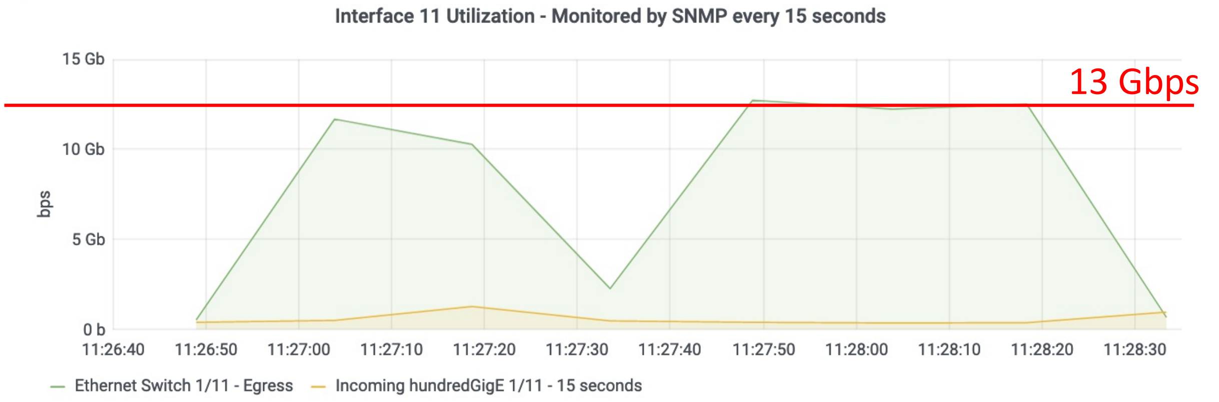
- Top: INT metadata exported in real time, per packet

- 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.*



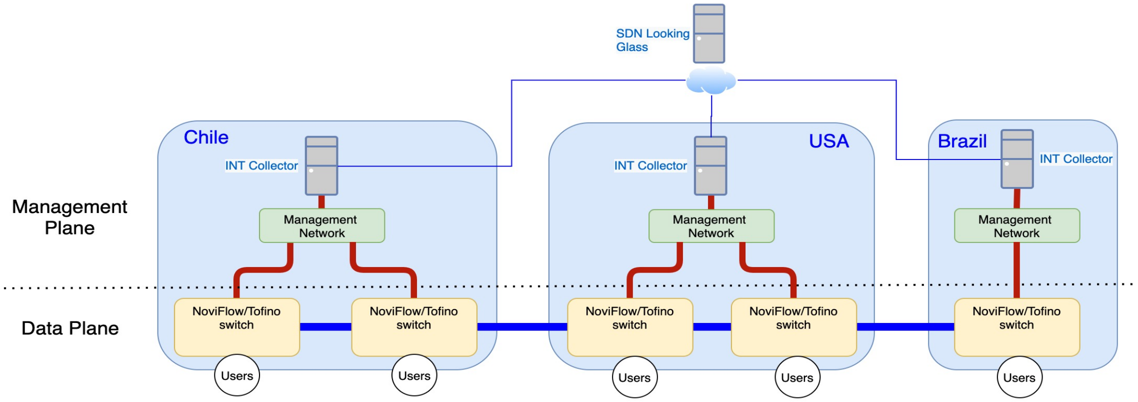
Interface 11 Utilization - Monitored using In-band Network Telemetry

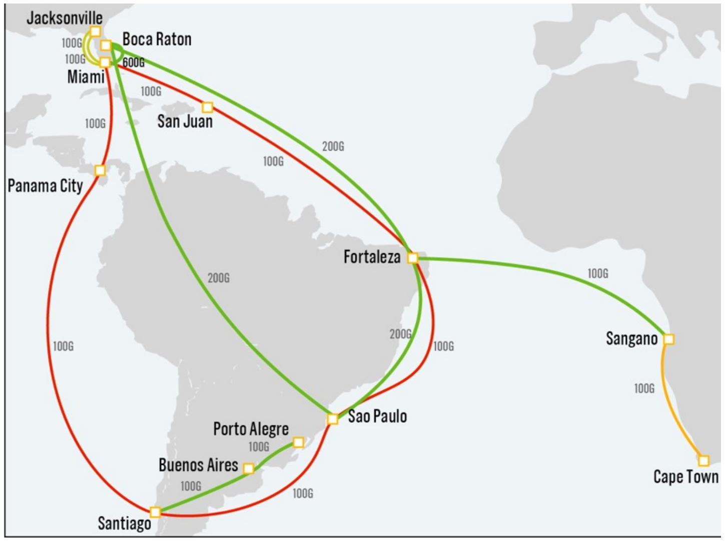


Interface 11 Utilization - Monitored by SNMP every 15 seconds

Ongoing Deployment at AmLight [1]

- At each AmLight site, P4 switches are replacing the current data plane
- Each pop has a Telemetry Collector parsing Mpps 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.*

- *How to dynamically enable INT monitoring of specific flows?*
    - *AmLight SDN orchestrator is very flexible when selecting what to monitor (per-source, per-destination, TCP and UDP, per port, etc.)*

Conclusion & Future Work

- 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 enables 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.*

Demo! Demo! Demo! Demo!

- We will be showcasing our environment in a more interactive approach
  - Challenges, benefits, some screens, our setup, and future.

- Zoom link:
  - https://go.fiu.edu/sc21_demo

- Tomorrow, at 10:30 AM EST.

Join us!

Los Angeles
New York
Miami
Panama
Fortaleza
Rio de Janeiro
Sao Paulo
Santiago

AmLight ExP
Americas Lightpaths Express & Protect

Thank You! Questions?

Jeronimo Bezerra, Italo Brito, Arturo Quintana,
Julio Ibarra, & Vasilka Chergarova/FIU

Renata Frez/RNP

Heidi Morgan/USC

Marc LeClerc and Arun Paneri/NoviFlow

<sdn@amlight.net>

Deploying per-packet telemetry in a long-haul network:
the AmLight use case