Automating Triaging of Network Circuit Flaps and Port Failures

Chris Berry, Qing Wang – Optical Engineers
Chet Powers, Harshit Gulati, Clement Cheung, Sam Pham – Software Engineers
Meta Platforms, Inc.

Connect. Collaborate. Accelerate.

OPEN Compute Project®

Connect. Collaborate. Accelerate.

Agenda

- Motivation
- Approach
- Examples
- Software Implementation
- Areas for Improvement
- Areas for Engagement

Motivation

- When deploying very large numbers of switches, optics, and fiber in a dynamic environment, ports go down and flap constantly
- With many deployed circuits, most possible failures will eventually be seen and need to be triaged and repaired.
- Things fail

– Software:	Configuration, initialization, bugs, etc.
– Chassis Hardware:	Components fail, solder, defects, etc.
– Optics:	Lasers, packaging, firmware, etc.
– Fiber:	Breaks, bends, etc.
– Removable Interfaces:	Contamination, contact, ESD, etc.
– Operational:	Installation, accidental disturbance, etc.

Motivation

- Network circuits involve complex systems connected together
- Many circuits that flap involve swapping components, many of which are diagnosed as No Trouble Found (NTF), and the circuit continues to flap
- Determining root cause of failures can be difficult for someone on the floor, particularly for transient, flapping interfaces
- Proper triaging needs to be able to look across devices during the outage to determine root cause
- To manage triaging and dealing with failures in a systematic way, we're developing an ever improving system to automatically triage failures

Approach: Check Software

- A network outage (either port goes down or flaps) triggers an automated investigation

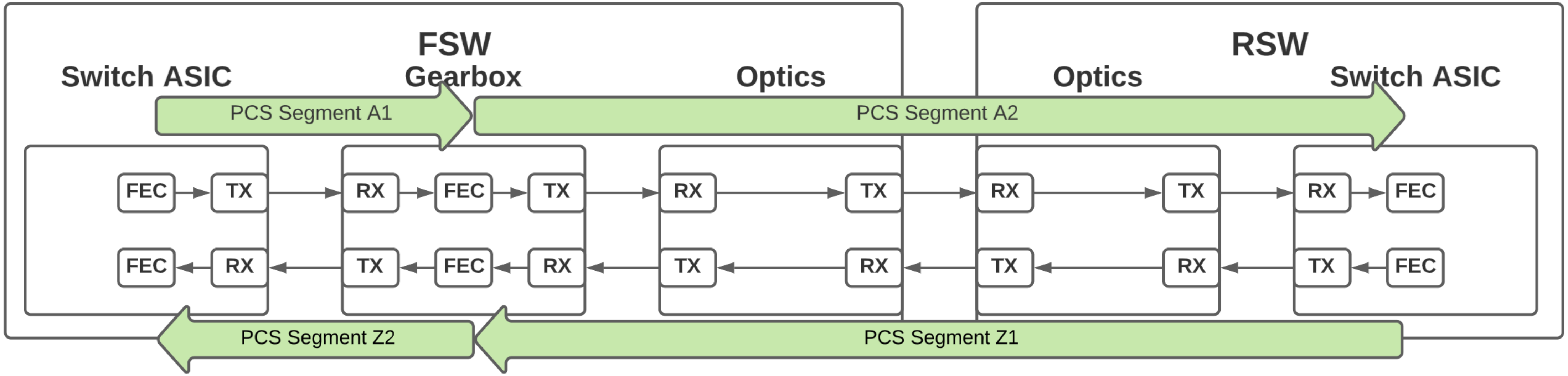
1. Look for potential software (SW) issues. Retrieve logs from both ends of the circuit including data from before, during, and soon after the event
   - Was SW healthy?
   - Did SW recently upgrade or reboot?
   - Was all the hardware on the switch configured / operational?
   - Were the optics powered up and configured?

Approach: Check Hardware

2. Look for detectable system hardware issues
   - Did the system restart?
   - Was a pluggable blade or optic reseated?
   - Are system power supplies in normal range?
   - Are components installed compatible?
   - Are temperatures in range?

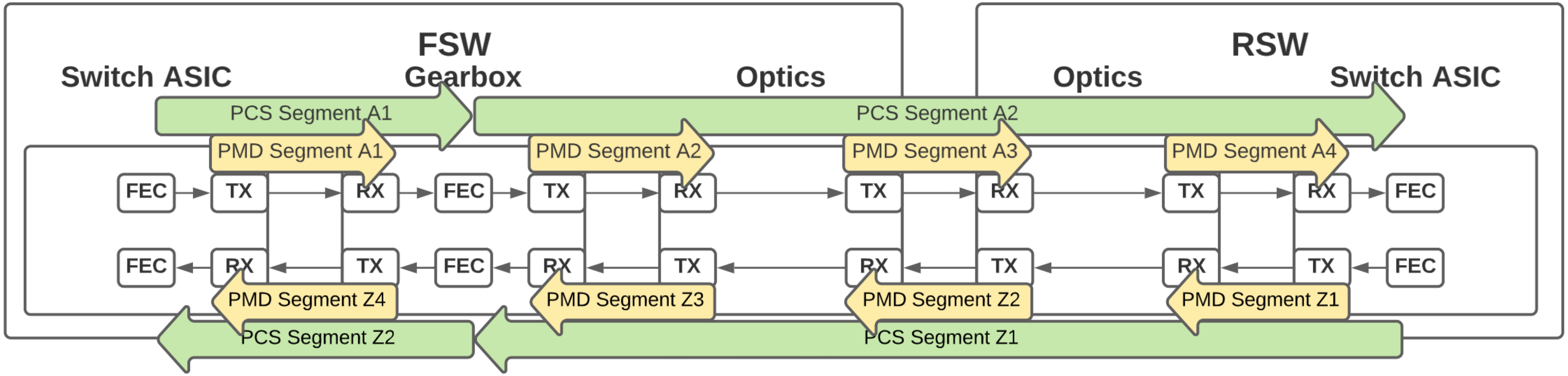
Approach: Locate PCS Segment

3. Isolate issue to failing FEC/PCS segment(s) across circuit
   - Move bidirectionally from Tx to Rx looking for uncorrectable FEC or PCS errors anywhere that FEC is decoded to isolate issue
   - Failing segment implicates components and interfaces included



Approach: Locate PMD Segment

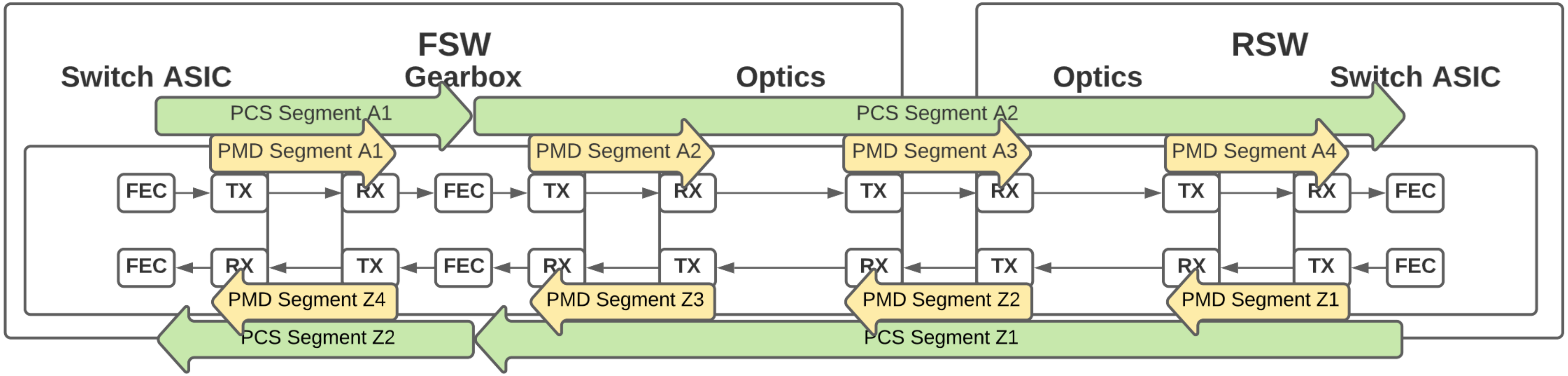
3. Check PMD/PMA segment(s) within failing PCS segments for errors
   - Booleans: Device faults, signal detects, loss of lock, alarms
   - Analog metrics: SNR, BER monitors, signal levels and stability



Connect. Collaborate. Accelerate.

Approach: Combine Directions

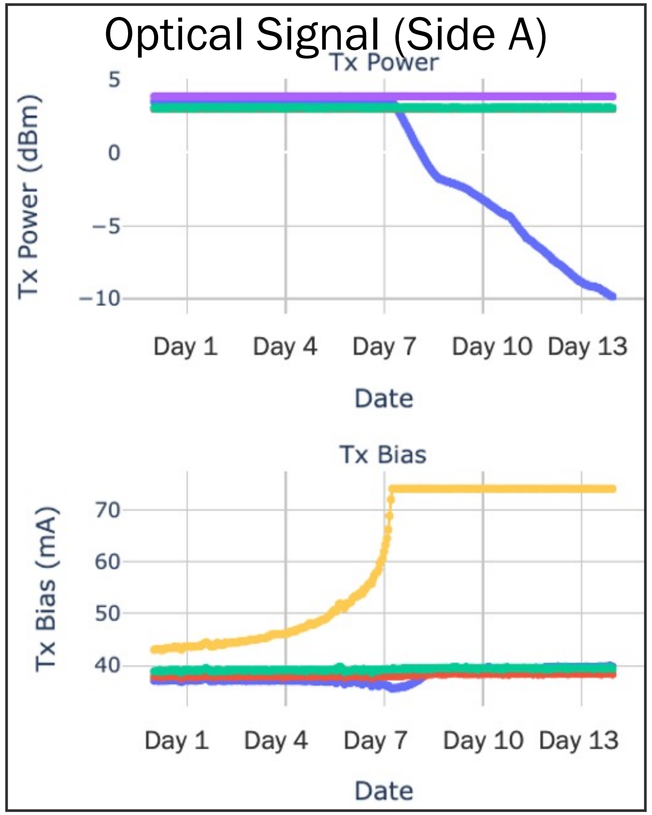
4. Combine results from both directions to finalize decision
   - Events such as fiber unplug/pinch may cause errors in PCS/PMD segments including fiber in both directions to lose Rx Optical Power simultaneously



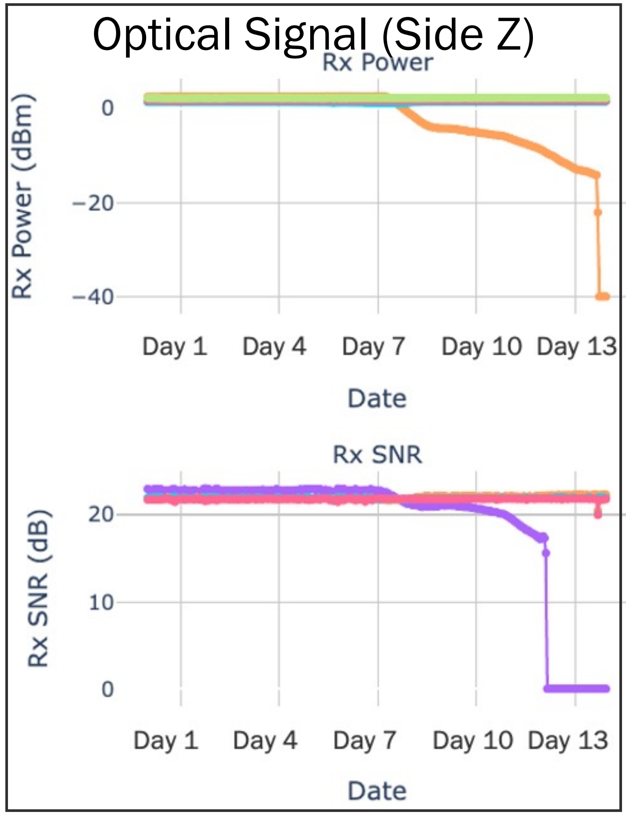
Connect. Collaborate. Accelerate.

Example: Laser Failing

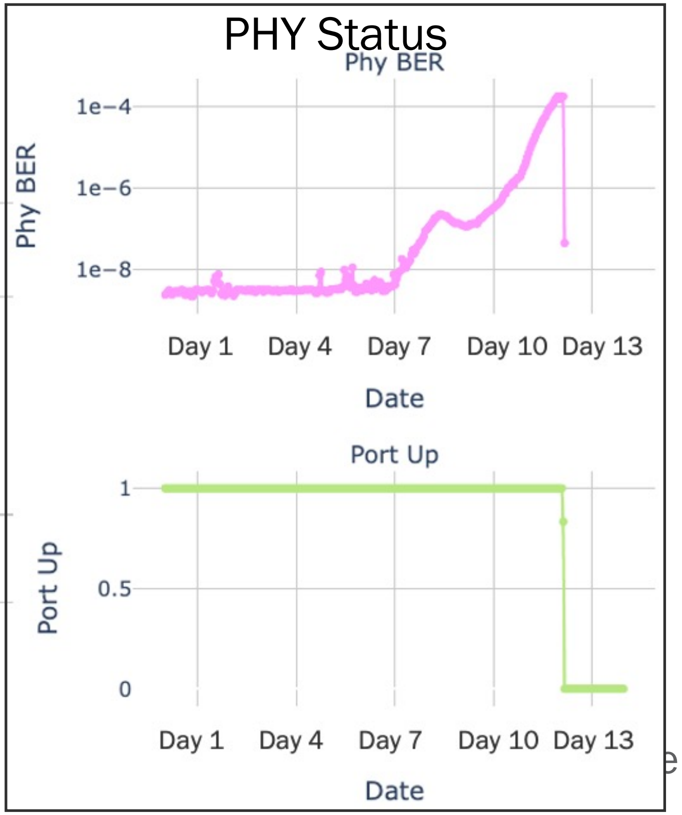
- Lasers failing are commonly cited as a primary circuit failure
- They are easy to diagnose as the failure is permanent and could be diagnosed without a system detecting transient conditions



Optical Signal (Side A)



Optical Signal (Side Z)

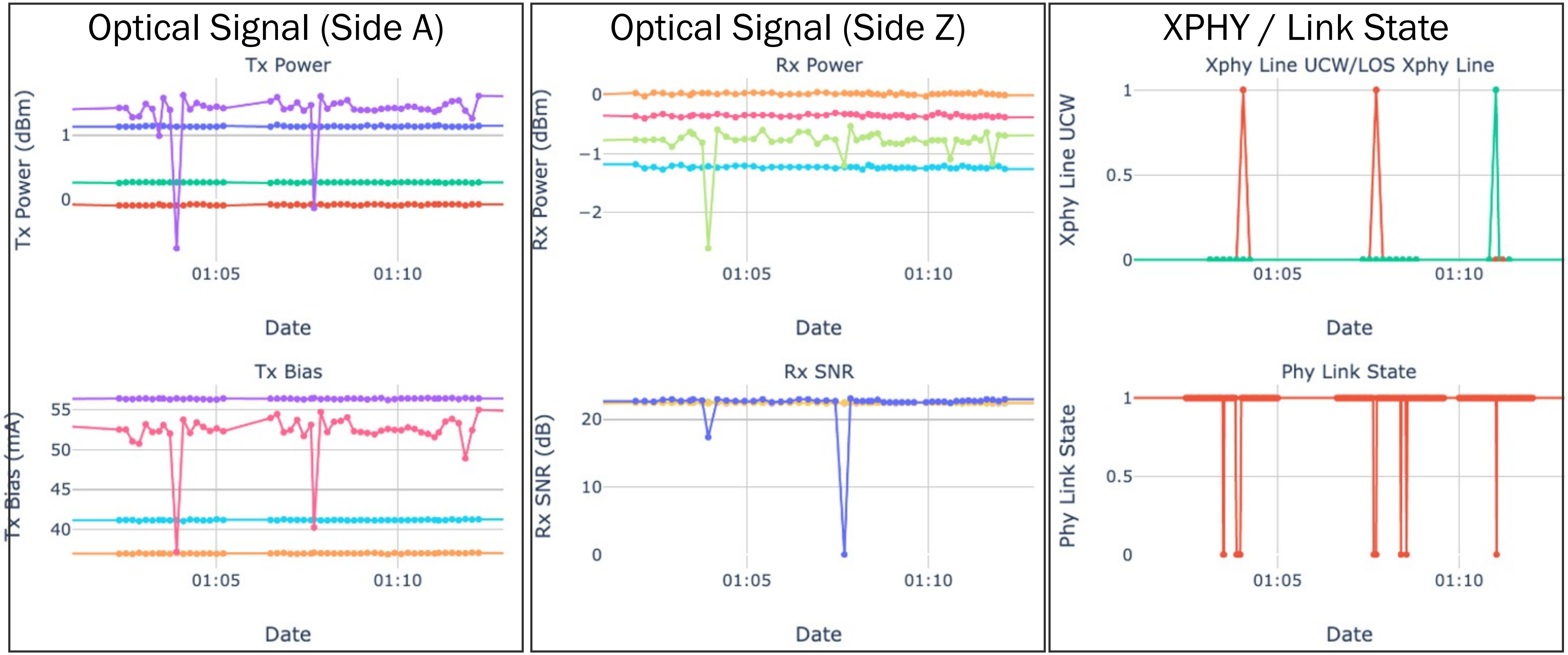


PHY Status

erate.

Example: Unstable Transmitters

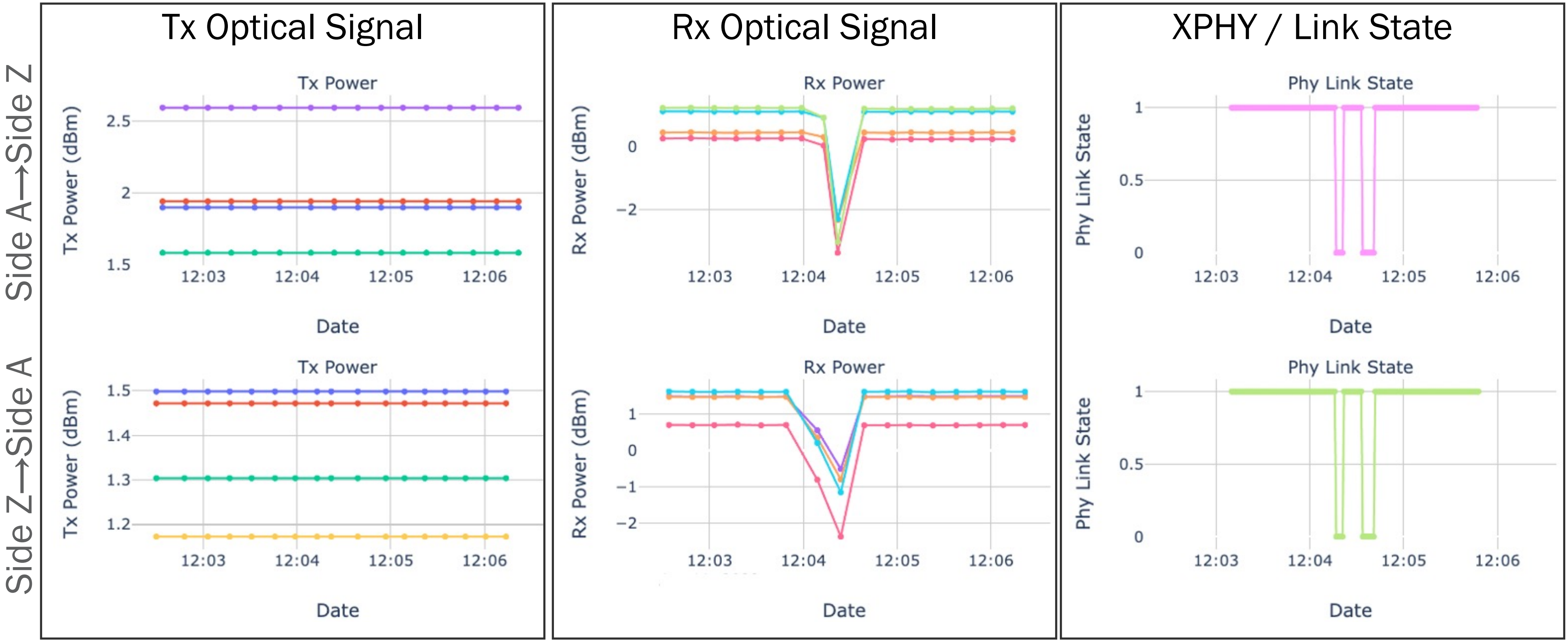
- Transient issues / link flaps are more difficult to diagnose as issue can be intermittent and missed in a subsequent capture
- Case below diagnosed root cause is laser instability



erate.

Example: Fiber Disruption

- Transient issues / link flaps are more difficult to diagnose
- Case below diagnosed as fiber issue and likely due to operational disturbance



erate.

Software Implementation: Counters

- SW/HW checks
  - Counters for service restarts, module swaps, module resets, etc.
  - wedge_agent.unclean_exits, wedge_agent.uptime, module.present, module.remediationCount

- Link-level checks
  - Periodically collect diagnostic data for iphy + xphy + optics in “snapshots” and store in memory
    - Snapshot format is generic and supports all platforms
  - On link events, publish recent + future snapshots to persistent storage

Snapshots

```
union LinkSnapshot {
  1: transceiver.TransceiverInfo transceiverInfo;
  2: PhyInfo phyInfo;
}
```

```
struct TransceiverInfo {
  …
  20: optional list<MediaLaneSignals> mediaLaneSignals;
}
```

```
struct PhyInfo {
  …
  10: optional PhySideInfo system;
  11: PhySideInfo line;
}
```

```
struct PhySideInfo {
  2: optional PcsInfo pcs;
  3: PmdInfo pmd;
  4: optional RsInfo rs; // Reconciliation sub-layer
}
```

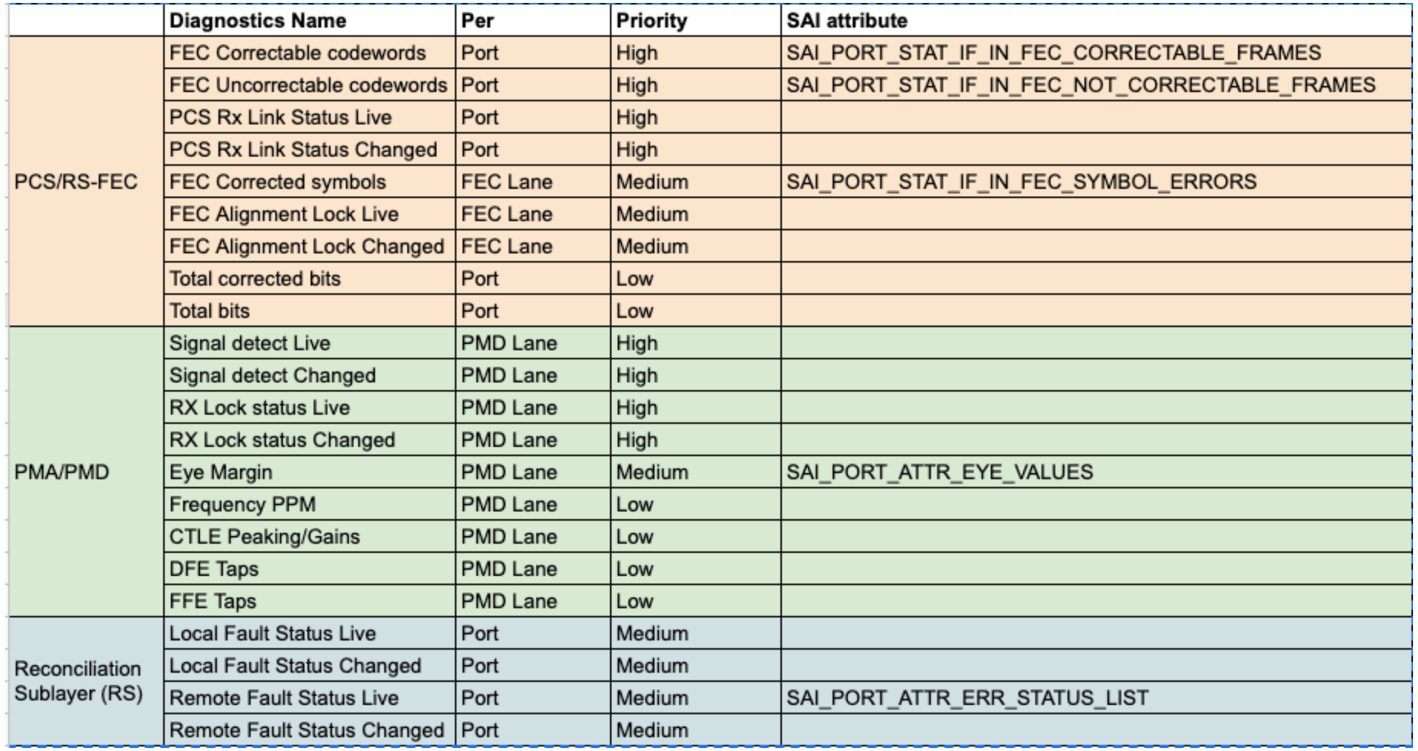
Software Implementation: Workflow

- Determine an issue_start_time based on when a link event was detected
- Check counters for any SW/HW issues around issue_start_time
- Collect all “snapshots” around issue_start_time. Look through snapshots in order to see if they match failure heuristics

Software Implementation: SAI

- We are largely focusing on migrating to SAI (Switch Abstraction Interface) for controlling our switching hardware

- Most phy parameters are currently not supported by any SAI interface. We are working to enumerate which phy parameters should be added to SAI



	Diagnostics Name	Per	Priority	SAI attribute
PCS/RS-FEC	FEC Correctable codewords	Port	High	SAI_PORT_STAT_IF_IN_FEC_CORRECTABLE_FRAMES
	FEC Uncorrectable codewords	Port	High	SAI_PORT_STAT_IF_IN_FEC_NOT_CORRECTABLE_FRAMES
	PCS Rx Link Status Live	Port	High	
	PCS Rx Link Status Changed	Port	High	
	FEC Corrected symbols	FEC Lane	Medium	SAI_PORT_STAT_IF_IN_FEC_SYMBOL_ERRORS
	FEC Alignment Lock Live	FEC Lane	Medium	
	FEC Alignment Lock Changed	FEC Lane	Medium	
	Total corrected bits	Port	Low	
	Total bits	Port	Low	
PMA/PMD	Signal detect Live	PMD Lane	High	
	Signal detect Changed	PMD Lane	High	
	RX Lock status Live	PMD Lane	High	
	RX Lock status Changed	PMD Lane	High	
	Eye Margin	PMD Lane	Medium	SAI_PORT_ATTR_EYE_VALUES
	Frequency PPM	PMD Lane	Low	
	CTLE Peaking/Gains	PMD Lane	Low	
	DFE Taps	PMD Lane	Low	
	FFE Taps	PMD Lane	Low	
Reconciliation Sublayer (RS)	Local Fault Status Live	Port	Medium	
	Local Fault Status Changed	Port	Medium	
	Remote Fault Status Live	Port	Medium	SAI_PORT_ATTR_ERR_STATUS_LIST
	Remote Fault Status Changed	Port	Medium	

Areas for Improvement: Optical Diagnostics

- Adding additional diagnostics would help detect more issues
  - Host side diagnostics
    - SNR and BER are being added in newer modules.
    - Additional signal quality metrics would be helpful for detecting poor electrical contact, ESD, equalization/ initialization issues
  - Media side diagnostics
    - Reflection is hard to diagnose without sending a tech. Would be very helpful to have MPI detector in DSP
    - Internal optical loopback is rare, but very useful

Areas for Improvement: PHY/XPHY Diagnostics

- PHY / XPHY diagnostics are not consistently defined and implemented
  - Scaling, availability, and details of interpretation are often hardware dependent and limited on older hardware
  - Behavior of counters often varies
  - Adding more standardized latched / sticky bits for signal detect, lock status, and errors
  - Speed of reading diagnostics through API widely varies
  - SAI interface to all PHYs / XPHYs critical to continue scaling

Areas for Engagement

- We would love to hear ideas from Industry on automating triaging from network operators and suppliers alike
- We will try to continue advocating for better diagnostics through standards such as the Common Management Interface Specification (CMIS) for optics and through SAI for PHYs/XPHYs

Thank you

Contributors

Chet Powers

Sam Pham

Clement Cheung

Harshit Gulati

Qing Wang

Chris Berry

Connect. Collaborate. Accelerate.

OPEN Compute Project®