Production Monitoring of Optics in Meta Datacenters

Anju John
Hardware Insights Engineer

Susu He
Optical Engineer

Contributors: Absar Ulhassan, Qing Wang, Chet Powers

Meta

Agenda

Background

Optics Telemetry at Scale

Fleet Metrics & Insights

Automated Triage Process

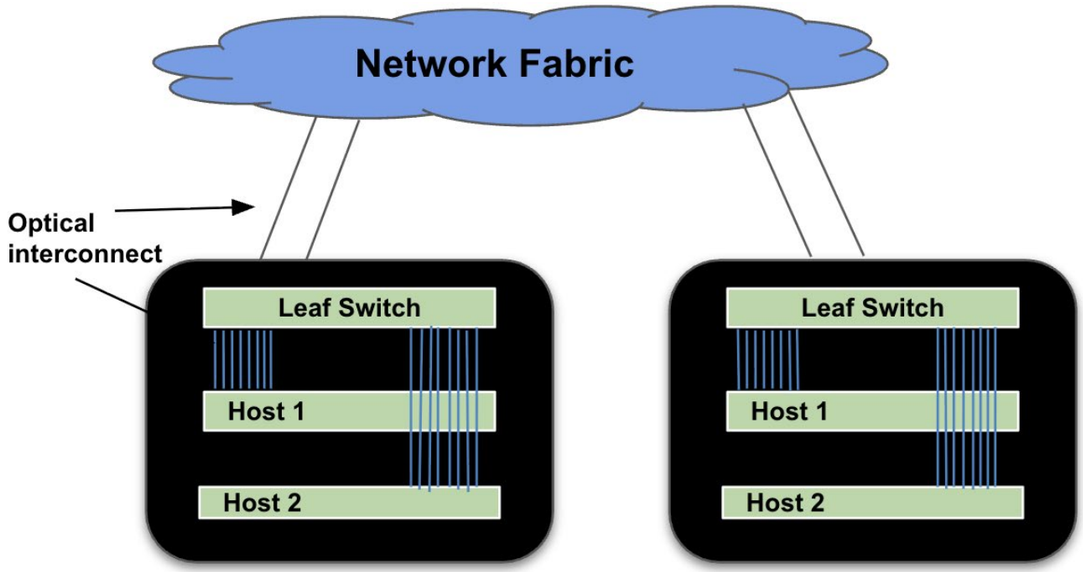
Conclusion

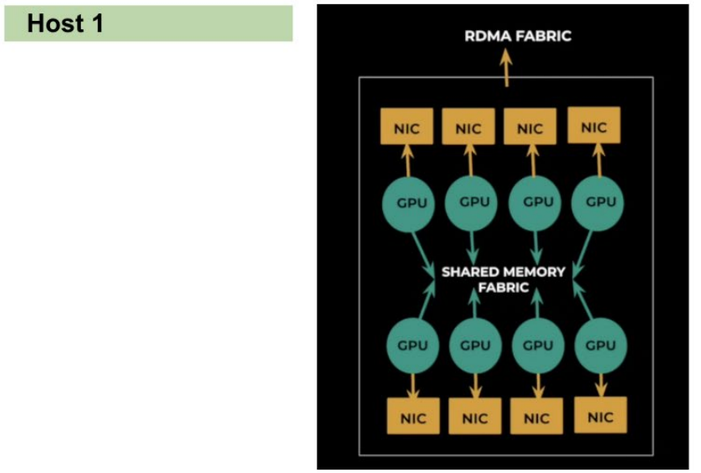
01. BACKGROUND

- Meta’s data center footprint is rapidly expanding, driven by the rise in AI/ML workloads and deployment of large-scale clusters.
- Demand for high-speed optical interconnects is increasing to support these advanced workloads.
- Efficient deployment of massive interconnect networks requires robust monitoring capabilities throughout the full life cycle of optical modules.
- Proactive monitoring and management of optical interconnects are critical to maintaining reliability and performance at scale.

Optical Interconnect at Meta DC

- Optical Interconnects are used in leaf and spine switch layers, backbone networks.
- Optics are directly connected to NICs in the GPU servers with switching chassis for large training clusters
- Optical link failures directly lead to a decrease in GPU node availability making job completion times longer.





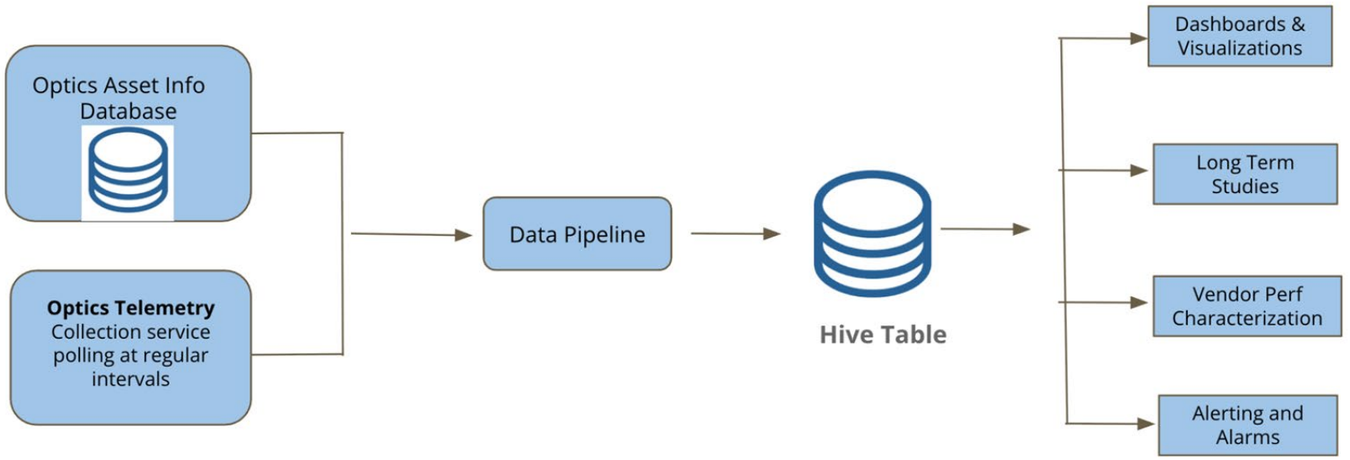
End-to-end link metrics in Optical Modules

- Active components in end-to-end optical link

- Sample metrics available on Host ASIC and optics side for monitoring
  - Transmit Power and Receive Power
  - Transmit Bias Current
  - Module Case Temperature
  - Signal-to Noise Ratios
  - FEC uncorrectable errors

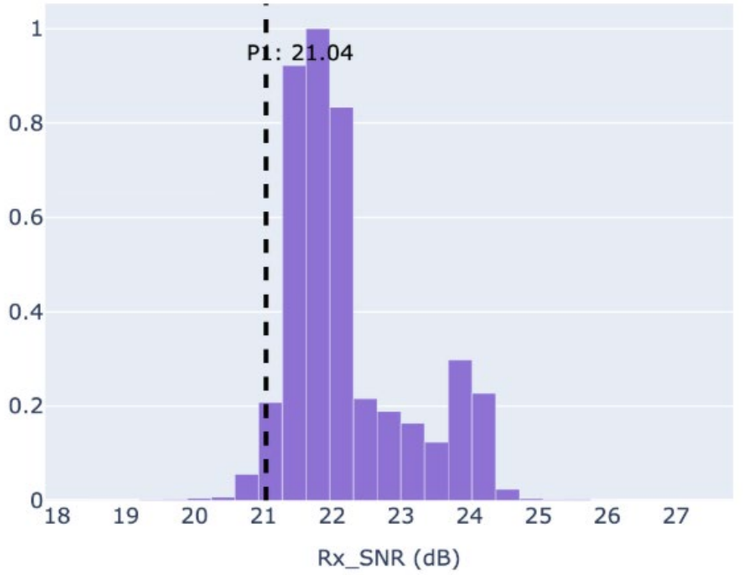
02. OPTIC TELEMETRY & INFRASTRUCTURE AT SCALE

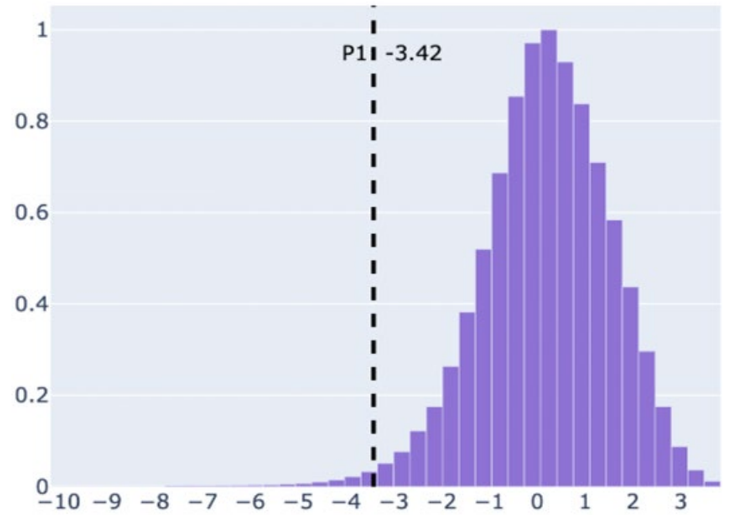
- DOM tracks critical transceiver metrics.
- QSFP service on production switches enables telemetry collection on regular intervals.
- Time-series data and static asset details are processed using pipeline.
- Results stored in hive tables.
- Provides a comprehensive view from switch-switch or switch-server links.



03. FLEET METRICS & INSIGHTS

- High-level metrics for optics reliability are collected to analyze trends and margins
- Margins can be deduced by finding the lowest 1% of the distribution of data for a parameter - representing the 'lowest' performing links.
- The difference between the P1 and spec value can be a reasonable indicator of the parameter margin
- Figs here depict the distribution of minimum Rx optical power and Rx SNR over a day.





Link Interruption Rate

- Another important metric to monitor in the evolving AI infrastructure is the interruption rate and the attribution of link related failures to the overall job reliability.

- Traditionally, metrics such as MTBF (Mean Time Between Failure), AFR (Annual Failure Rate), and ASR (Annual Swap Rate) have been used to quantify failure levels.

- AI systems require a comprehensive approach to evaluating system-level reliability due to their growing complexity and interconnectedness which includes the first hop network

- Link Interruption metric and the factors attributing to link failures can give meaningful insights on job reliability.

04. AUTOMATED TRIAGE PROCESS

What is Automated Triage?

- Automated triage is Meta’s system for detecting, diagnosing, and resolving optics -related faults in data center networks.
- Integrates with telemetry infrastructure to continuously monitor optical module health.
- Enables rapid identification and localization of failures, minimizing downtime.

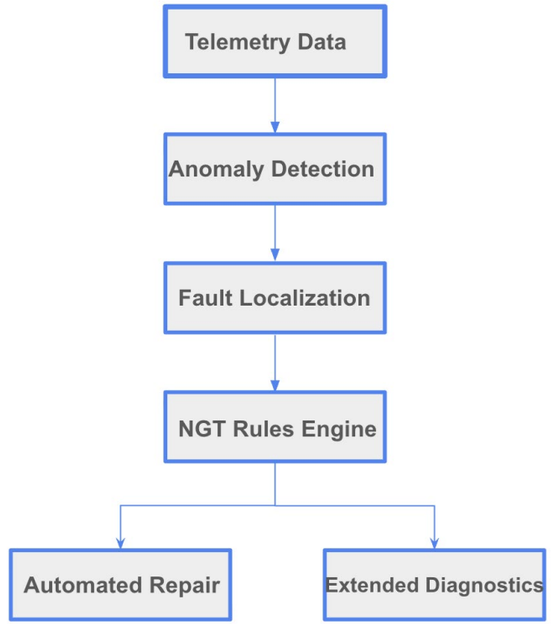
Key Features

- Monitors telemetry data for abnormal signals.
- Localizes faults to the correct component (optics, switch port, or upstream system).
- Triggers automated or operator -guided remediation workflows.

How Does It Work?

- Anomaly Detection
  - Compares telemetry against static thresholds and dynamic fleet baselines.
  - Flags outliers and persistent link instability.
- Fault Localization
  - Correlates data from both ends of the link, nearby ports, event logs, and swap history.
  - Uses Meta's **N**ext **G**eneration **T**riage (NGT) rules engine to infer root causes.
- Remediation
  - Automated repair workflow:
    - schedules module replacement, issues work orders, flags for operator follow up.
  - Extended diagnostics:
    - Lookbacks, PRBS test etc.

-

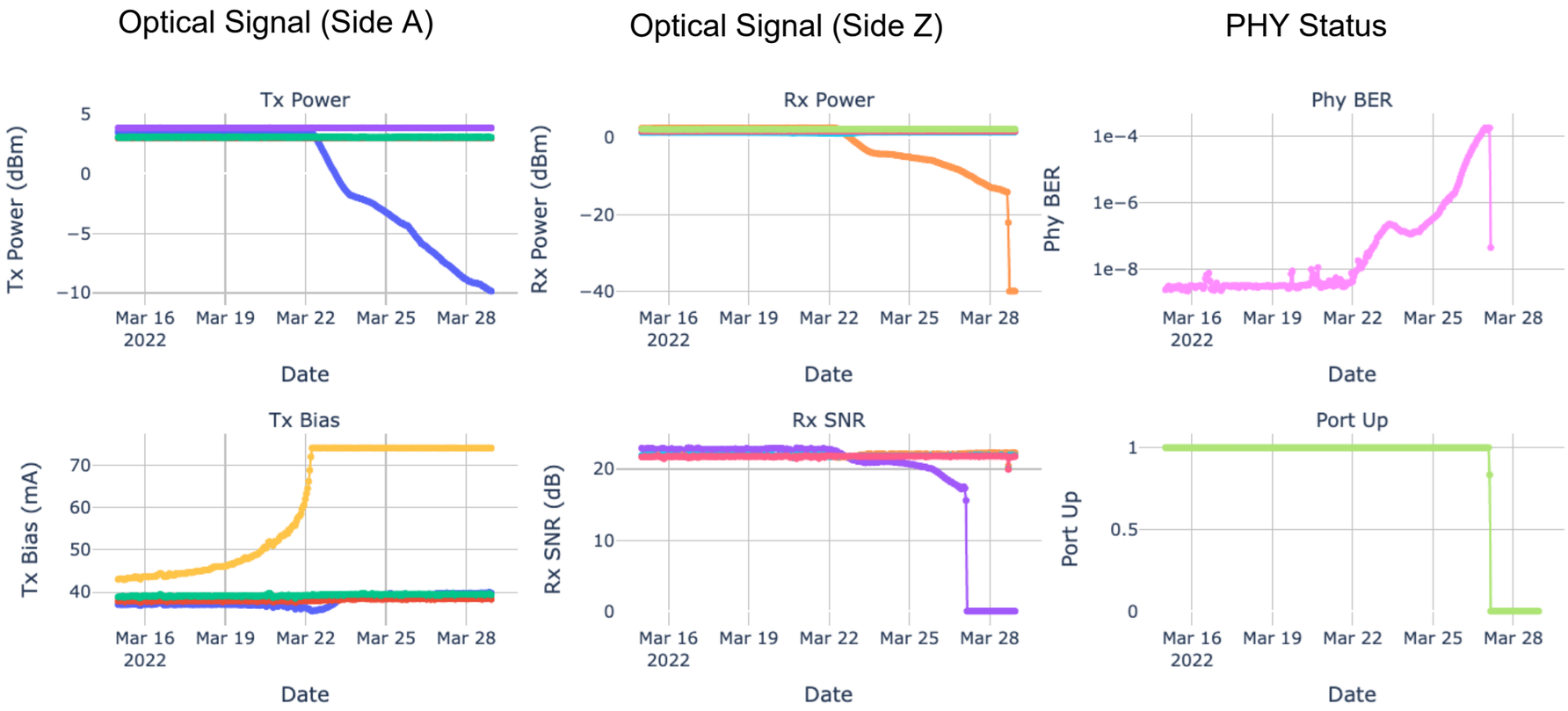


Impact

- Reduces Mean Time to Detect (MTTD) and Mean Time to Repair (MTTR).
- Lowers unnecessary part swaps ("No Trouble Found" cases).
- Improves reliability and uptime for AI/ML workloads.

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.



05. CONCLUSION

- Robust optics monitoring is essential for Meta’s data center reliability and performance.
- End-to-end telemetry and automated triage enable failure detection and rapid resolution of issues.
- Proactive management minimizes downtime and supports large-scale AI/ML workloads.


Meta