Production Monitoring of Optics in Meta Datacenters

Anju John, Susu He, Absar Ulhassan, Qing Wang, Chet Powers
Meta HQ 1 Hacker Wy, Menlo Park, CA 94025, Ph: 832-965-8601

I. ABSTRACT

As Meta continues to expand its data center footprint - particularly with the recent rise in AI/ML workloads and the deployment of large scale clusters - the demand for high-speed optical interconnects is growing rapidly. Efficient deployment of such a massive interconnect network requires robust monitoring capabilities across the full life cycle of the optical modules. With model training taking place over large clusters where multiple accelerators are communicating with each other, the consequences of a single link failure become increasingly severe; potentially impacting job completion times and cluster performance. This paper outlines Meta's strategy for monitoring optical modules, beginning at the NPI stage and extending through production and sustaining throughout the full lifecycle of the optical modules. We highlight the infrastructure capabilities required to support both unit-level and fleet-scale monitoring and maintenance. Additionally, we explore key trends and challenges in optics performance and the debugging process as optical networks scale in complexity and importance.

Keywords— Optics, Interconnect, Datacenters, AI Infrastructure

II. INTRODUCTION

As our network hardware fleet continues to expand, along with the introduction of new optics variants in NPI, it becomes increasingly critical to monitor optics hardware performance metrics and automate the debugging process. This ensures the reliability of training jobs and facilitates early identification of failure trends. The demand for greater bandwidth (BW) per network switch port and the availability of next-gen optics technologies enabled us to go from 100Gb/s per port to 200Gb/s and on to 400Gb/s. Efficient maintenance of such a massive network interconnect fabric with a heterogeneous fleet, calls for appropriate fleet-wide monitoring capabilities and troubleshooting to gauge the health of these links and reducing downtimes.. Furthermore, with the deployment of optics in the server domain for scale-out of large scale training clusters, optics will directly connect NICs in the GPU servers with switching chassis. (Fig. 1)

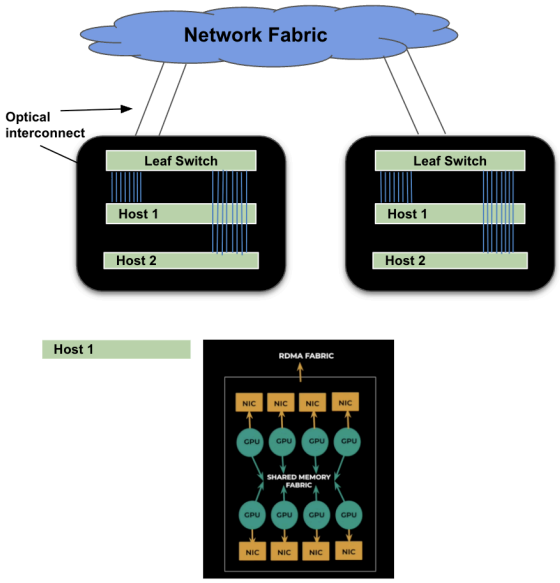

Fig. 1: AI Backend Network

For AI training [1-4] workloads, the potential impact of optics failures becomes higher since link downtime would directly lead to a decrease in GPU node availability making job completion times longer. Given the massive scale of deployed optics and the critical nature of connectivity in AI zones, it is crucial to have real-time end-to-end link monitoring in place that allows us to assess the performance and

reliability of the full link. Besides these main goals, we also hope to gain some predictive power to identify low-performing links/parts ahead of time based on data trends and stats, i.e. by looking for signs of degradation in health metrics.

To this end, we use aggregate data over a day, e.g. max value of a parameter (such as optics case temperature) and monitor if a large swath of optics in a region are getting close to thresholds. This helps us pro-actively take actions to prevent large-scale impact on network performance.

A large-scale monitoring system coupled with an automated troubleshooting process that identifies the failing endpoint and suggests appropriate repair actions enables the network operations team to reduce the mean time to repair (MTTR), streamline the repair workflow and reduce overall downtime. In this paper, we explore the large-scale optics telemetry monitoring infrastructure and the automated failure detection workflows deployed across Meta's data centers.

III. OPTICS TELEMETRY INFRASTRUCTURE

With the rapidly growing number of optics deployments across Meta data centers, managing and maintaining the health of such a vast network interconnect fabric necessitates robust, fleet-wide monitoring solutions to ensure operational efficiency and reliability. This section provides an overview of the monitoring framework and the optics hardware telemetry metrics collected.

Digital Optics Monitoring (DOM) is an industry-standard digital interface designed to track critical performance metrics of transceivers, facilitating effective monitoring and diagnostics in optical communication systems. A service implemented on our production switches, called QSFP service, enables the collection of these metrics on regular time intervals. The resulting time-series optics health data, combined with asset-related information are processed using a data processing pipeline and exported to a data warehouse for long-term data storage. The optics asset information database serves as the central repository for managing optics assets, and is updated whenever a new serial number is added or changed in any given port of a switch. The optics time-series telemetry data is collected per optics on the switch port, per channel wherever applicable for both endpoints of the links. This will provide a comprehensive view of the entire link from switch-switch or server-switch connections. (Fig. 2)

A major challenge in telemetry collection is managing the vast volume of data generated by millions of transceivers, which necessitates a trade-off between collection frequency and the range of metrics gathered. To optimize the data collection, aggregated statistical metrics (min, max, avg etc) are derived from the data and are computed on a daily basis and visualized through dashboards. This approach not only allows us to evaluate the real-time performance of the optics fleet but also enables analysis of historical trends, providing valuable insights over time.

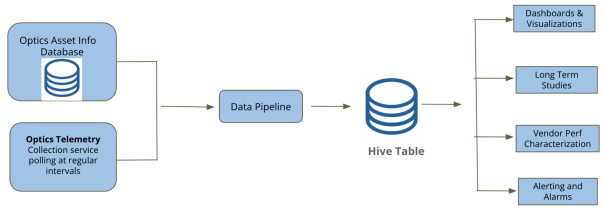

Fig. 2: Optics telemetry infrastructure

IV. METRICS & INSIGHTS

The key factor indicating optical transceiver performance is link availability over its lifespan. To quantify such a metric, one might consider as a candidate: the frequency of replacement of

optics modules for restoring an end-to-end connection between hosts. While an annual swap rate (ASR) seems a simple metric, it may not accurately represent true link failures due to various factors - mainly the existence of misdiagnosed or NTF (no trouble found) cases. Besides this, a system's margin relative to absolute limits (the spec) stands out as a prominent measure of operational performance. While transceiver diagnostics provide a wide range of metrics, collecting all of them is often impractical. Therefore, we focus on a set of high-level, essential metrics that are critical for ensuring optics reliability [5-6] and detecting emerging trends. Along with static asset information to identify the link e.g., host, port and optics vendor, we track and store several link parameters using the monitoring frameworks to measure the quality of optics at scale.

An appropriate method of deducing such a margin is by looking at the lowest 1% of the distribution of data for a parameter - which represents the 'lowest' performing links. The difference between the P1 and spec value can then serve as a reasonable indicator of margin for the parameter in question. Fig 4,5 depicts the distribution of minimum Rx optical power and minimum Rx SNR over a day. Similar margins can be deduced for other key parameters to identify areas for improvements.

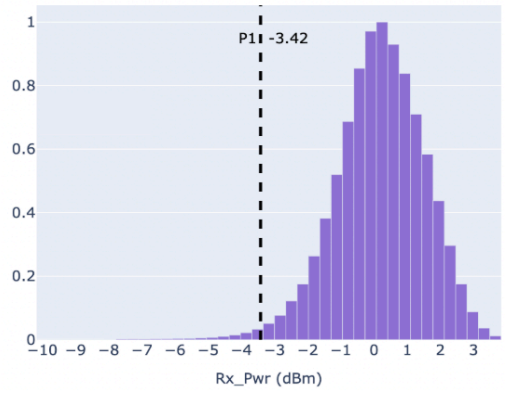

Fig 4: Min Rx Power over a day for a specific optics type

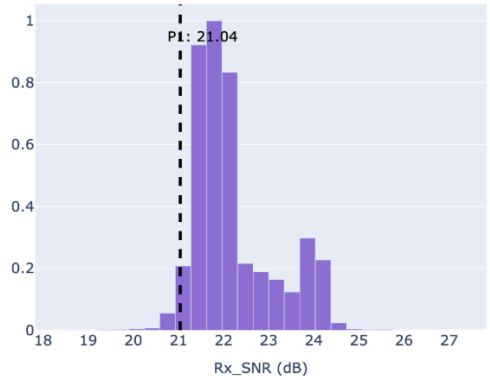

Fig 5: Min Rx SNR for Ch-0 over a day for a specific optics type

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. However, these metrics have limitations in accurately capturing the unique reliability requirements of AI infrastructure. To address this, we are focusing on Annual Interruption Rate (AIR) as a key metric, which more closely aligns with AI infrastructure's efficiency needs. A link event can be described as link down, link flap or error packets. Not all link events are caused by optical failures. For instance, switch software events or electrical events could cause link flaps. Hence having a comprehensive end-to-end view of the link including all failure nodes will be beneficial in understanding overall link quality.

V. AUTOMATED TRIAGE PROCESS

To ensure timely remediation of link failures, Meta has built an automated triage system that integrates with telemetry infrastructure to detect, diagnose, and resolve optics-related faults. The triage system continuously monitors telemetry data and identifies signals of abnormal behavior. It localizes faults to the correct component -

whether the optics, switch port, or upstream system - and triggers either automated or operator-guided remediation workflows.

Triage workflows can be triggered by threshold violations, persistent link instability, or schedule analytics jobs that detect slow performance degradation. Anomaly detection logic compares telemetry against static thresholds and dynamic fleet baselines to identify outliers. Once an anomaly is flagged, the system performs fault localization by correlating data from both ends of the link, nearby ports, historical event logs, and prior swap history. The root cause is inferred using Meta's internal rules engine, known as Next Generation Triage (NGT) [7-8], which generates suspected root causes.We always escalate to repair unless the issue is determined to be transient – health checks are run before sending to repair. The automated repair workflow could involve scheduling a module replacement, issuing work orders, or flagging the issue for operator follow-up. In ambiguous cases, further diagnostics, such as loopbacks or PRBS, may be scheduled. Fig 6 shows the automated triage process workflow.

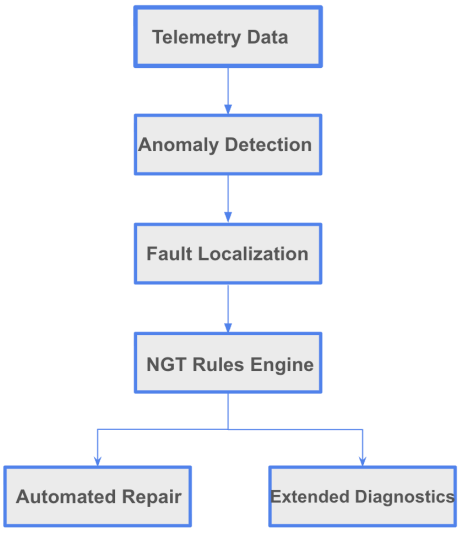

Fig 6: Automated Triage Process Workflow

The NGT database also keeps track of all triage outcomes. Its extensible run framework supports onboarding of new hardware types and telemetry sources as the network evolves. The deployment of this automated triage pipeline has reduced Mean Time to Detect (MTTD), shortened Mean Time to Repair (MTTR), and significantly lowered the rate of unnecessary part swaps caused by "No Trouble Found" (NTF) cases.

VI. CONCLUSION

Robust optics monitoring - with rich telemetry, key metrics and automated triage - is crucial for sustaining reliable, high-performance data center infrastructure at Meta. These systems detect issues early, reduce downtime, and ensure service continuity, directly supporting Meta to run resilient, large-scale AI/ML applications.

REFERENCES

[1] Engineering @ Meta: "Building Meta's GenAI Infrastructure," - K. Lee, A Gangidi & M Oldham March 12, 2024

[2] Engineering @ Meta: "OCP Summit 2022: Open hardware for AI infrastructure," - A. Bjorlin October 18, 2022

[3] Engineering @ Meta: "GenAI Training In Production," - J. Lee, KR Kishore & A Gangidi June 12, 2024

[4] D. Alduino, "AI Clusters at Scale: Opportunity & Impact of Optical Connectivity," ECOC 2024, Frankfurt, Germany, 2024

[5] V. Lowalekar and A. Chakravarty, "Laser Reliability Performance of High-Speed Optical Interconnects in Hyperscale Datacenters for PAM4 applications," in Frontiers in Optics + Laser Science 2023 (FiO, LS), Technical Digest Series (Optica Publishing Group, 2023), paper FM5D.6.

[6] V. Lowalekar and A. Chakravarty, "Long Term Reliability Methodology of Next Gen Pluggable Optical Modules for PAM4 Applications in Hyperscale Datacenters," in Frontiers in Optics + Laser Science 2024 (FiO, LS), Technical Digest Series (Optica Publishing Group, 2024), paper FM1C.2.

[7] E. Chou, A. Mohan, C. Berry, C. Powers, and M. Morales, "Novel In-Line Triage Methodology for High-Speed Optical Transceivers in Hyperscale Datacenters," in Optical Fiber Communication Conference (OFC) 2024, Technical Digest Series (Optica Publishing Group, 2024), paper M4E.2.

[8] C. Berry et al., “Automating Triaging of Network Circuit Flaps and Port Failures,” OCP 2022, San Jose, CA, 2022