Novel In-Line Triage Methodology for High-Speed Optical Transceivers in Hyperscale Datacenters

Elaine S. Chou, Arun Mohan, Chris Berry, Chet Powers, and Mario Morales
Meta, One Hacker Way. Menlo Park, CA 94025
eschou@meta.com

Abstract: A novel in-line triage methodology has been developed by leveraging data collected from optical transceivers and network switches. A success rate of ~68% was achieved by correlating diagnosis from triage to failure analysis from vendors. © 2024 The Author(s)

1. Introduction

Optical modules being returned (RMA) to vendors are typically subjected to a switch-level traffic test for several hours, and if no packet drops are observed, the module is determined to be *No Trouble Found* (NTF). Data centers (DCs) and vendors have long suffered from high rates of NTF. This adversely affects Meta in two ways. (1) Resources and cost are unnecessarily spent swapping potentially good optics and shipping them long distances for vendor failure analysis (FA). (2) The resolve and resources of vendors are strained in processing so many NTFs. Fig.1 illustrates this issue: the average NTF rate of all optics over the last four quarters was ~79%.

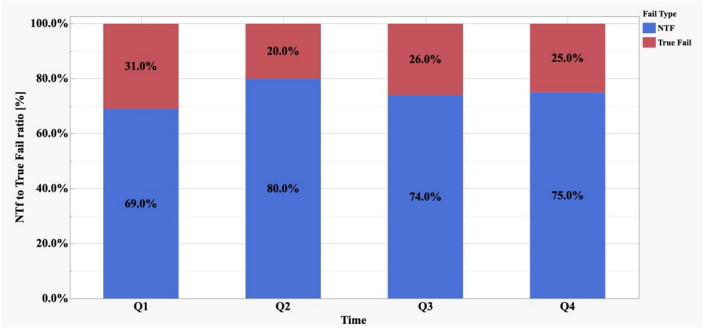

Fig.1. The NTF rate for all high-speed optical modules over four quarters.

Broadly speaking, the root cause of the high NTF rate is (a) Incorrect diagnosis at DC: link flaps may be incorrectly associated with optical module failure. Meta has tried to address this issue by developing tools for automated diagnosis of network issues [1]. (b) Incorrect diagnosis of NTF: NTF determination made at the vendors by a short traffic test may not always be accurate. The rest of the paper focuses on efforts to address these issues using a novel triaging methodology developed by Meta.

2. Automated link triage

With our constantly expanding data center network and emerging use-cases such as AI that require prompt resolution of link issues, automation is important to keeping services running smoothly without overwhelming manual maintenance efforts. To that end, we developed a pipeline to automate network monitoring and failure resolution, with the goal of quickly repairing broken links with minimal human intervention and ultimately enabling a more stable, reliable, and scalable network. In a data center network, many switches, optics, and fiber operate in a dynamic environment, so even if individual failure rates are low, many link down or link flap events happen each day. Identifying the source of every failure can be difficult and time consuming, motivating many optical network rootcause studies [2, 3]. Our pipeline incorporates knowledge of optics and signals with the experience gained from past issues into a comprehensive triaging flow to programmatically narrow down the potential source of issues.

The strength of our automated link triage flow lies in its comprehensive diagnostic logic that analyzes a wide range of data to triage issues from incompatible software configuration to optics laser instability. Snapshots of the hardware state are continuously collected from switches and stored in memory. When a failure such as a link flap is detected, a decision tree identifies the devices and circuits involved and leverages the snapshot data around the time of the link event to run a root cause analysis. Decisions are based on detailed data over the entire link from the switches, line cards, and optics over a few minutes spanning the link event. By focusing on a short interval, the data snapshots provide high time resolution views of the sequence of events leading up to and following a link failure.

The following case studies demonstrate two issues detectable by our triaging methodology, one optics module issue and one non-module issue.

2.1 Case study: module issue

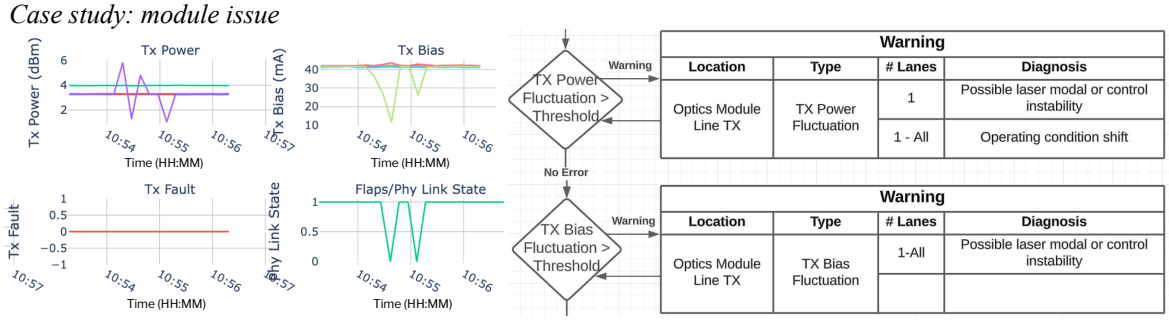

Fig. 2. A laser failure causes link flaps lasting tens of seconds each.

In Fig. 2, the transmit (Tx) laser bias and power on one of the four lanes fluctuates, causing link flaps. The Tx bias change happens infrequently and lasts only a few seconds, so even a long soak test may not reproduce the failure.

2.2 Case study: non-module issue

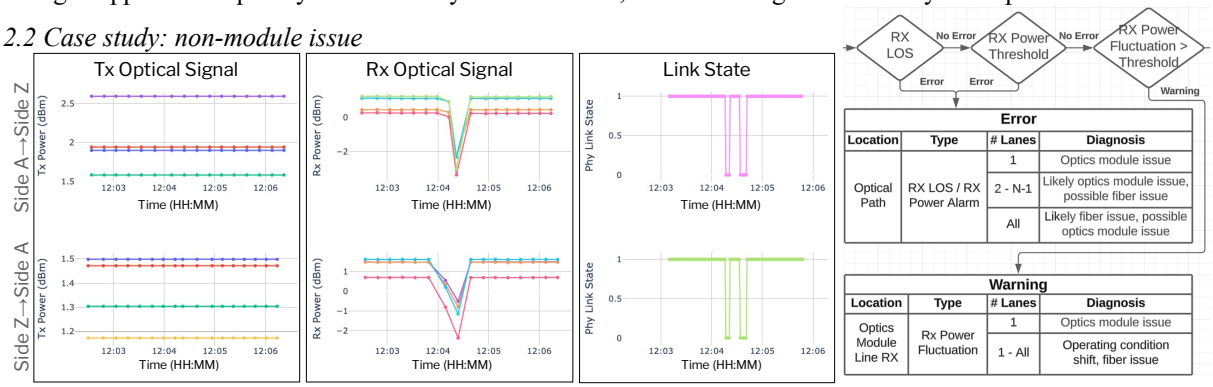

Fig. 3. A fiber disruption causes power on all lanes in both directions to drop momentarily.

In Fig. 3, a short fiber disruption causes the received (Rx) optical power to drop by 3 dB on all lanes in both directions, resulting in several link flaps. Optical fibers are sensitive to disruption from installation work or reflections [4]. For such ephemeral issues resolved without intervention, we can definitively rule out the optical transceiver as a cause of failure and avoid unnecessary swaps.

3. Automated triage vs. RMA analysis

3.1 Efficacy of automated link triage

The modules diagnosed by our automated link triage were sent to the vendors to perform their own testing. During this data collection phase, automated diagnosis was performed but results did not inform DC optics swap decisions, so a balance of optics failure and non-optics rootcauses were represented in the data set. As described in section 1, vendors perform a switch-level traffic test to determine if a module is NTF or failed. Failed modules are subjected to further FA to determine the failure mode. Table 1 below represents a comparison of our diagnosis vs. vendor diagnosis. As shown below, the diagnosis match of our automated triage and vendor FA is ~68%. This in turn implies if we accept the automated diagnosis as a source of truth in DC link repair, the NTF rate would be much lower than the current ~79%.

Table 1. Correlation between vendor and NGT diagnosis.

Percentage of matched diagnosis (%)	Total number of transceivers
68.4%	2927

3.2 Breaking down the efficacy of automated link triage

To better understand and improve the efficacy of the automated diagnosis logic, we looked further at the breakdown of rootcause as shown in Fig. 4a.

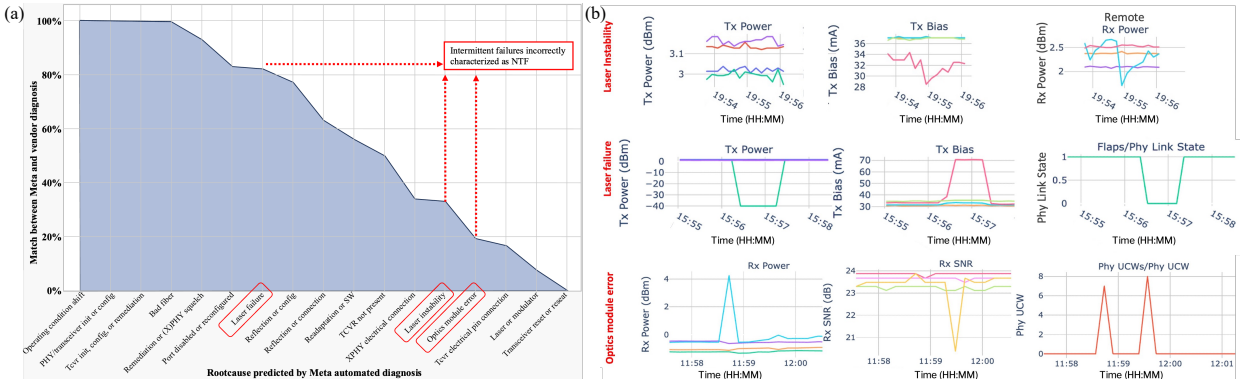

Fig. 4. (a) Efficacy of automated triage: percentage match of Meta's rootcause prediction vs. vendors failure analysis. (b) Intermittent issues caught by automated diagnosis that are often determined to be NTF by vendors.

Typically, vendor FA results on RMA optics have been accepted as correct, but a deeper dive into the snapshot data revealed this not to be the case, especially in cases where failures are intermittent. Three optics Tx failure modes included in our decision tree logic that cause infrequent link disruptions are shown in Fig. 4b; *Laser Instability*: Tx laser bias on one lane is unstable, and although the Tx power measurement looks stable, the Rx power at the opposite endpoint confirms optical power fluctuations with the laser bias. At any given time, all power levels are healthy, but the change over time can cause link flaps. *Laser failure*: The Tx laser bias on one of the four lanes suddenly jumps and Tx power drops, causing a link flap. The Tx power loss is infrequent and not persistent, so it is hard to catch without continuous polling. *Optics module error*: An Rx power alarm is temporarily raised on a module and the link goes down, but it is not a permanent failure, so the link state transitions between up and down unpredictably. All of these optical transceivers were misclassified as NTF by the vendor.

As many of these failures are intermittent, a simple traffic test screen often used by the vendors to determine true failures is insufficient to isolate these issues. Meta has developed a modified traffic test with temperature cycling to better address intermittent issues. We subjected 23 modules to a standard Traffic Test (TT) at nominal temperature and then further subjected them to a *Temperature Cycled Traffic Test* (TCTT), where the temperature of the modules was cycled multiple times from 15ºC to 65ºC.

Table 2. Summary of data collected with Temp Cycled Traffic Test (TCTT)

Result of Traffic Test (TT)	Result of Temp Cycle Traffic Test (TCTT)	Automated triage rootcause
Pass	Fail	IPHY_ELECTRICAL_PIN_CONNECTION
Pass	Fail	LASER_FAILURE
Pass	Fail	LASER_FAILURE
Pass	Fail	OPTICS_MODULE_ERROR
Pass	Fail	LASER_INSTABILITY
Pass	Pass	OPTICS_MODULE_ERROR
Pass	Pass	LASER_INSTABILITY

Table 2 summarizes the TCTT results, showing the seven samples that automatic triage identified as possible transceiver failures with the corresponding rootcause. Of these seven samples, five also failed TCTT. TCTT has ~70% efficacy in determining possible transceiver related fails. Effort is currently ongoing to improve confidence in this screening methodology.

4. Conclusion

In-line triage presented here is important for overall cost reduction and network up time. This work is even more important for AI workloads where any link loss may lead to have the entire training state to be reloaded. In future work we hope to leverage further statistical data to determine failures for which the assignment of blame to optical modules is not obvious. Leveraging advanced DSP monitoring features as described in [4] will further improve in-line triaging.

5. References

[1] C. Berry *et al.*, "Automating Triaging of Network Circuit Flaps and Port Failures," OCP 2022, San Jose, CA, 2022.
[2] C. Tremblay *et al.*, "Detection and root cause analysis of performance degradation in optical networks using machine learning," ECOC 2023, Glasgow, Scotland, 2023.
[3] K. Abdelli, *et al.*, "Fault monitoring in passive optical networks using machine learning techniques," ICTON 2023, Romania, 2023.
[4] A. Ulhassan *et al.*, "Statistical method for multi-path interference detection in IMDD optical links," *JLT* vol. 41, no. 14, July 2023.