Data Analytics Practice for Reliability Management of Optical Transceivers in Hyperscale Data Centers

Jianqiang Li[1,*], Zhicheng Wang[2], Chunxiao Wang[2,3], Qin Chen[2], Peng Wang[2], Rui Lu[2], Songnian Fu[3], Chongjin Xie[4]

[1]*Alibaba Group, Bellevue WA, USA,* [2]*Alibaba Group, Hangzhou, China*
[3]*Huazhong University of Science and Technology, Wuhan, China,* [4]*Alibaba Group, Sunnyvale CA, USA*
* *jason.li@alibaba-inc.com*

Abstract: There are limitations when directly interpreting reliability information of optical transceivers from manufacturers to end users. Data analytics in a large optical transceivers' population is studied for data center operators with a case study. © 2020 The Author(s)

1. Introduction

Over the past few years, data center operators have been continuously deploying and upgrading high-speed optical-transceivers in data centers to meet fast-growing bandwidth demands. Hyperscale data center operators such as Alibaba Cloud typically have millions of optical transceivers with different types and manufacturers in their systems, which poses great challenges on optical transceiver reliability management [1]. Data center operators also want to have a deeper understanding of how reliable these transceivers perform and what characteristics and factors are associated with their failures, as this knowledge can improve optical transceiver design, development and manufacturing, and eventually increase the reliability of these optical transceivers and thus their networks. In practice, most of the reliability and failure mode information is obtained from optical transceiver manufacturers based on accelerated life tests in small populations and historical return merchandise authorization (RMA)/failure analysis (FA) databases [2]. However, there are a number of limitations using this information in production networks with a huge number of optical transceivers for end users, as in the storage sector [3]. For example, the information often isolates the connections between optical transceivers and specific deployment characteristics or historical operation conditions. The data center operators would benefit from running their own data analytics to characterize the reliability and failure modes of optical transceivers matched with their own production networks.

Most hyperscale data center operators has established their own infrastructure to collect, store and process the essential data which is believed to be good indicators of the operating status for in-field optical transceivers, including Alibaba Cloud. In this paper, we present the online data analysis activities and experiences for optical transceiver reliability management in a large production environment with a specific case study. The statistics of operation data from a large number of transceivers are believed to be able to offer more facts and insights.

2. Reliability Management System

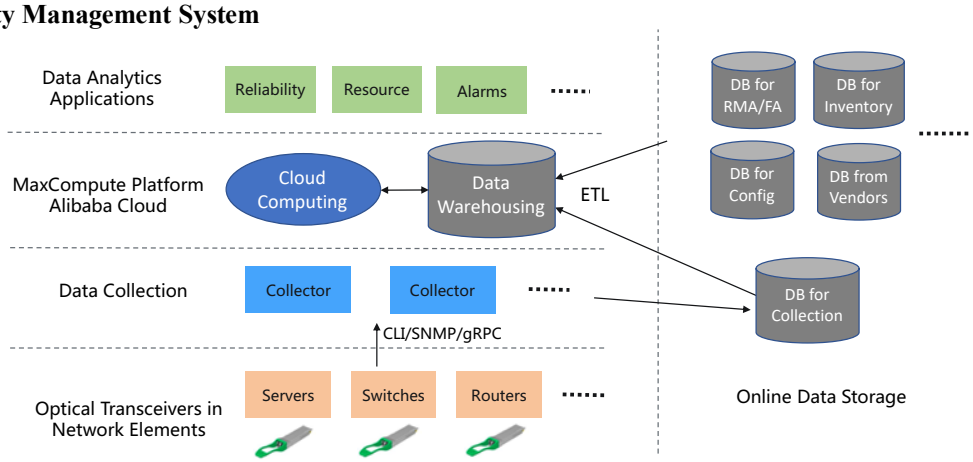

Fig. 1. The architecture of optical transceiver reliability management system

As shown in Fig.1, reliability management is one of the essential applications of the system built in Alibaba Cloud that collects, stores and analyzes online optical transceiver data from servers, switches, and routers across the entire production data center networks. In the data collection layer, multiple running collectors pull or subscribe different data of optical transceivers from all network elements (e.g. the servers, switches and routers) in all data centers at

frequencies of up to minute level. The most recent raw data is temporally stored in an online database, and is periodically transferred to the MaxCompute platform with the help of integrated extract, transform, load (ETL) tools for data persistence and subsequent processing. Note that MaxCompute is a general purpose, fully managed, multi-tenancy data processing platform for large-scale data warehousing developed by Alibaba Cloud. In addition, several other online databases are also integrated into the data warehouse for data fusion. Pre-processing of the raw data is necessary since data can be lost or corrupted during collections from a large population of units. With all the pre-processed data of in-field optical transceivers, data analyses are implemented to enable diverse applications including reliability evaluation and failure management.

3. Data Analytics for Optical Transceiver Reliability Management

3.1. Digital Diagnostic Monitoring (DDM) Statistics

As the majority of the collected data, DDM attribute-value pairs as a function of sampling time are the major available indicators of online performance for in-field optical transceivers. Therefore, the DDM data is periodically aggregated to get various statistics in terms of manufactures, part numbers, hardware/firmware versions, host device models, host device OS version, clusters, data centers, failure modes, accumulated power-on age, and so on. The trend is characterized by connecting time-ordered statistics at sequential timestamps within a given observation window. The DDM statistics are also associated with the concurrent traffic and packet error data in the same fiber links. Besides the typical DDM parameters, we are customizing more parameters that can be read from transceiver EEPROM in several new versions of optical transceivers. The above efforts are assisting us to identify multiple common problems, such as early infant mortality failure modes, incompatibility between optical transceivers and host devices, the unfavorable operation conditions in several data centers, clusters, and host device models. All of these will help initiate countermeasures to improve the effective reliability of the entire production networks.

3.2. Reliability Characterization in Production Networks

First, manufacturers and end-users may have different views on what is a failure for optical transceivers. An end user tends to consider an event where one optical transceiver deteriorates the traffic-level performance in a user-prescribed manner as a failure. It is no surprise to do so since the optical transceiver is no longer suitable to keep running at that time and replacement must be immediately initiated. In fact, there is a high probability for these replaced optical transceivers to be shown as “no trouble found (NTF)” after manufacturers’ FA process. There may be some hidden malfunctions to be dug out behind NTFs. Therefore, end users often identify a failure event when an in-field optical transceiver was replaced and put into the RMA process. Here, we refer to the manufacturers’ failure rate as “failure rate”, and refer to the end-users’ failure rate as “RMA rate”. Both rates need to be tracked by end users, but the RMA rate is in much more real time since RMA/FA process takes time. Second, in production networks, new optical transceivers are often deployed and powered on in a unit of cluster or point of presence (POP). Over a given observation window, different groups of optical transceivers have diverse start times for in-service. Therefore, one should be careful when calculating the RMA or failure rate. In order to better characterize the reliability, we selected annualized RMA/failure rate by considering the actual power-on hours of each individual optical transceiver [4,5]. Annualized RMA/failure rate gives the estimated probability that an optical transceiver will fail during a full year of use. In Alibaba, we are tracking the annualized RMA/failure rates over three observation windows (one year, half year, and one month) in terms of manufacturers, part numbers and purchase orders. More importantly, the rates are also broken down by the age an optical transceiver was when it was replaced, which forms the well-known bathtub curve in reliability engineering. This kind of manipulation helps monitor the reliability over the entire lifecycle and identify several abnormal behaviors, which will be exemplified by the case study in Section 4.

3.3. Proactive Identification of Misbehaviors and Failures

It is easy to understand that most data center operators want to implement advanced data analytics to identify or predict potential optical transceiver problems or failures prior to traffic interruption. Proactive misbehavior or failure identification is crucial to guarantee the service-level agreement especially for large scale production networks. Besides the constant observation of DDM statistics discussed in Section 3.1, we also pay attention to the normality of a DDM parameter over a given set of optical transceivers, such as the optical power. The technical logic behind this is the well-known central limit theorem which establishes that, when small independent random variables are added, their properly normalized sum tends toward a normal even if the original variables themselves are not normally distributed. In principle, the received optical power collected from

an optical transceiver is related with a large set of random factors, such as fabrication process of components, chip models, packaging, optical connectors, temperature, fibers, and monitoring accuracy etc. Temperature is often playing the most significant role in affecting the optical power. Our prior study indicated that, it is quite likely for the received optical powers of optical transceivers from one part-number, one manufacturer, and host devices in one model to be normally distributed, since this set of optical transceivers have similar temperature statistics, as seen from Fig. 2. We are implementing two parallel surveillance tasks to appreciate the above conclusion. First, normality tests are periodically carried out over pre-selected sets of optical transceivers. This helps proactively identify several potential common misbehaviors or even design/fabrication defects in group. Second, a Gaussian-fitted probability density function (PDF) library is established based on historical DDM data for each given set of optical transceivers. It has a high chance to predict the failures of individual optical transceivers prior to traffic loss with the library. Moreover, we are leveraging several dedicated diagnosis algorithms to filter the high-risk optical transceivers when the failure patterns are clearly identified.

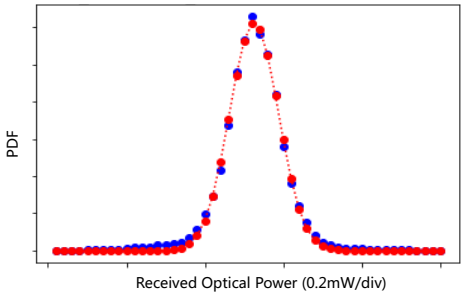

Fig. 2. The PDF of received optical power from an exemplary optical transceiver set (Blue: real, Red: Gaussian fitting)

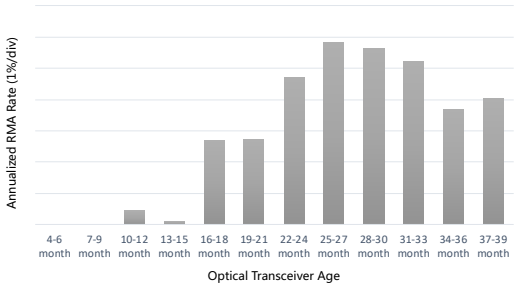

Fig. 3. Annualized RMA rate broken down by optical transceiver age

4. Case Study: Massive Failures in a Similar Mode

Massive failures in a similar failure mode is a disaster to data center operators with a large number of optical transceivers. These massive failures often occur in a set of optical transceivers from one manufacturer, one part number and one purchase order. There must be several common misbehaviors or design/fabrication defects. It is of a great value for data center operators to identify this kind of hidden event as early as possible. Then immediate countermeasures can be taken to prevent upcoming massive in-field failures in production networks. With the data analytics capability elaborated in Section 3, one event was successfully identified by annualized RMA rate tracking in groups. As shown in Fig.3, the annualized RMA rate curve broken down by optical transceiver age shows a distinct deviation from the classis bathtub curve. The observed group of optical transceivers tends to fail after working one and half year. The failure mode analysis was done with a clear root cause and recognized failure pattern where the optical power undergoes a slow climbing prior to a failure. As shown in Fig.4, a dedicated block-slope-based algorithm was developed to proactively select the optical transceivers that are on the way to a failure.

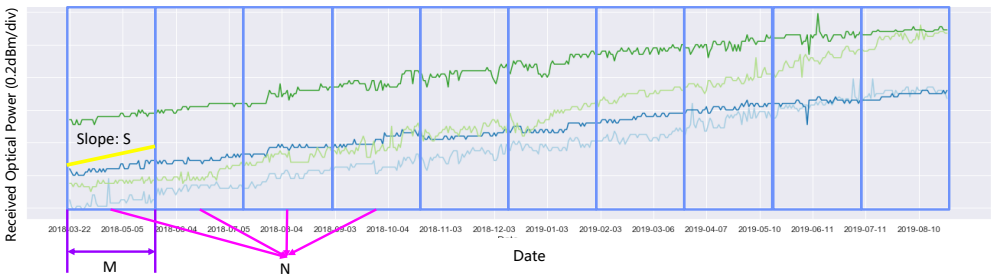

Fig. 4. Illustration of the dedicated diagnosis algorithm

5. References

[1] A. Chakravarty, *et al*, "Characterizing Large-Scale Production Reliability for 100G Optical Interconnect in Facebook Data Centers Data Centers," Frontiers in Optics, Washington, D.C. United States, ISBN: 978-1-943580-33-0, 2017.
[2] "Reliability Issues for Optical Transceivers," DfR Solutions, White Paper, www.dfrsolutions.com
[3] E. Pinheiro, "Failure Trends in a Large Disk Drive Population," Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07), Feb. 2007
[4] "Diving into "MTBF" and "AFR": Storage Reliability Specs Explained," Inside IT Storage, Seagate. Apr 2010.
[5] "Hard disk drive reliability and MTBF / AFR," Seagate, www.seagate.com/support/kb/hard-disk-drive-reliability-and-mtbf-afr-174791en/