Explaining SSD Failures using Anomaly Detection

Chandranil Chakraborttii
University of California Santa Cruz
cchakrab@ucsc.edu

Heiner Litz
University of California Santa Cruz
hlitz@ucsc.edu

ABSTRACT

NAND flash based solid-state drives (SSDs) represent an important storage tier in data centers holding most of today's warm and hot data. Even with the advanced fault tolerance techniques and low failure rates, large hyperscale data centers utilizing 100,000's of SSDs suffer from multiple device failures daily. Data center operators are interested in predicting SSD device failures for two main reasons. First, even with RAID [2] and replication [5] techniques in place, device failures induce transient recovery and repair overheads, affecting the cost and tail latency of storage systems. Second, predicting near-term failure trends helps to inform the device acquisition process, thus avoiding capacity bottlenecks. Hence, it is important to predict both the short-term individual device failures as well as near-term failure trends.

Prior studies on predicting storage device failures [1, 6, 7, 9] suffer from the following main challenges. First, as they utilize black-box machine learning (ML) techniques, they are unaware of the underlying failure reasons rendering it difficult to determine the failure types that these models can predict. Second, the models in prior work struggle with dynamic environments that suffer from previously unseen failures that have not been included in the training set. These two challenges are especially relevant for the SSD failure detection problem which suffers from high class-imbalance. In particular, the number of healthy drive observations is generally orders of magnitude larger than the number of failed drive observations, thus posing a problem for training most traditional supervised ML models.

To address these challenges, we propose to utilize *1-class ML models* that are trained only on the majority class. By ignoring the minority class for training, our 1-class models avoid overfitting to an incomplete set of failure types, thereby improving the overall prediction performance by up to 9.5% in terms of ROC AUC score. Furthermore, we introduce a new learning technique for SSD failure detection, *1-class autoencoder*, which enables interpretability of the trained models while providing high prediction accuracy. In particular, 1-class autoencoders provide insights into what features and their combinations are most relevant to flagging a particular type of device failure. This enables categorization of failed drives based on their failure type, thus informing about specific procedures (e.g., repair, swap, etc.) that need to be applied to resolve the failure.

For analysis and evaluation of our proposed techniques, we leverage a cloud-scale dataset from Google that has already been used in prior work [1, 8]. This dataset contains 40 million observations from over 30,000 drives over a period of six years. For each observation, the dataset contains 21 different SSD telemetry parameters including SMART (Self-Monitoring, Analysis, and Reporting Technology) parameters, the amount of read and written data, error codes, as well as the information about blocks that became non-operational over time. Around 30% of the drives that failed during the data collection process were replaced while the rest were removed, and hence no longer appeared in the dataset. As a result, we obtained approximately 300 observations for each healthy drive (40 million observations in total) and 4 to 140 observations for each failed drive (15000 total observations). We treated each data point as an independent observation and normalized all the non-categorical data values to be between 0 and 1.

One of our primary goals was to select the most distinguishing features that are highly correlated to the failures for training. We used three different feature selection methods, Filter, Embedded, and Wrapper [4] techniques, for selecting the most important features contributing to failures for our dataset. The resulting set of top features selected were correctable error count, cumulative bad block count, cumulative bad block count, cumulative p/e cycle, erase count, final read error, read count, factory bad block count, write count, and status read-only. The dataset containing only the top selected features is then used for training the different ML models.

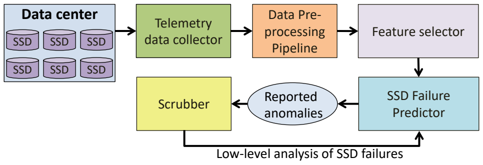

Figure 1: Block diagram of the Deployed System

In a datacenter, we envision our SSD failure prediction technique to be implemented as shown in Figure 1. The telemetry traces are collected periodically from all SSDs in the datacenter and sent to the preprocessing pipeline transforming all input data into numeric values while filtering out incomplete and noisy values. Following data preprocessing, feature selection is performed to extract the most important features from the data set. The preprocessed data is then either utilized for training or inference. For inference, device anomalies are reported and classified according to our 1-class autoencoder approach. SSDs can then be manually analyzed by a technician or replaced directly. As an alternative, a scrubber can be leveraged to validate the model predictions by performing a low-level analysis of the SSD.

To evaluate the five ML techniques, we first label all 40 million observations in the dataset to separate between healthy and failed drive observations. We then perform a 90% - 10% split of the dataset into a training set and an evaluation set respectively. For training the 1-class models we remove all failed drive observations from the training set, however, the evaluation set is kept identical for our proposed 1-class techniques and the three baselines. We use ROC AUC score as a metric for comparing the performance of our approaches with chosen baselines, which is inline with prior work [1] and use 10-fold cross-validation for evaluating all approaches. The 1-class autoencoder model utilizes 4 hidden layers comprising of 50, 25, 25, and 50 neurons, respectively. The neurons utilize a *tanh* activation function, *Adam* optimizer, and the model is trained for 100 epochs. We use early stopping with a patience value of 5 ensuring that the training of the model stops when the loss does not decrease after 5 consecutive epochs. Increasing the number of hidden layers beyond 4 increases the training time significantly without providing performance benefits. Figure 2 illustrates the comparative performance of different ML techniques for predicting SSD failures one day ahead. Among the baselines, Random Forest performs best, providing a ROC AUC score of 0.85. Both our 1-class models outperform the best baseline. In particular, 1-class isolation forest achieves a ROC AUC score of 0.91, representing a 7% improvement over the best baseline while 1-class AutoEncoder, outperforms Random Forest by 9.5%.

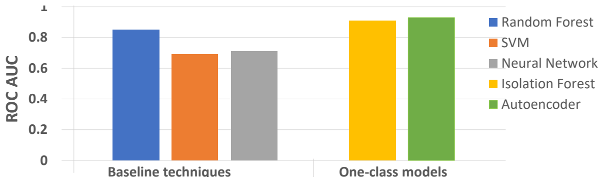

Figure 2: ROC AUC comparison of evaluated methods

This work introduces 1-class autoencoders for interpreting failures. In particular, our technique exposes the reasons determined by our model to flag a particular device failure. This is achieved by utilizing the reconstruction error generated by the model while reproducing the output using the trained representation of a healthy drive. The failed drives do not conform to the representation, hence, generate an output that differs significantly from the actual input producing a large reconstruction error. We study the reconstruction error per feature to generate the failure reasons. The features which contribute more than average error per feature to the reconstruction error, is defined as a *significant* reason. The results show that many failed drives show a higher than normal number of *correctable_errors* and *cumulative_bad_block*, however they were selected as a reason for failure only for 35%, and 30% of the cases respectively. Hence, our analysis shows that there exist particularly relevant features that indicate device failures in many cases, however, only the combination of several features enables accurate failure prediction.

To conclude, this paper provides a comprehensive analysis of machine learning techniques to predict SSD failures in the cloud. We observe that prior works on SSD failure prediction suffer from the inability to predict previously unseen failure types motivating us to explore 1-class machine learning models such as 1-class isolation forest and 1-class autoencoder. We show that our approaches outperform prior work by 9.5% ROC-AUC score by improving on the prediction accuracy for failed drives. Finally, we show that 1-class autoencoders enable interpretability of model predictions by exposing the reasons determined by the model for predicting a failure. A more comprehensive evaluation of our approach in discussed in [3], where we show the adaptability of 1-class models to dynamic environments with new types of failures emerging over time and the impact of predicting further ahead in time.

REFERENCES

[1] Jacob Alter, Ji Xue, Alma Dimnaku, and Evgenia Smirni. 2019. SSD failures in the field: symptoms, causes, and prediction models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 75.

[2] Mahesh Balakrishnan, Asim Kadav, Vijayan Prabhakaran, and Dahlia Malkhi. 2010. Differential raid: Rethinking raid for ssd reliability. *ACM Transactions on Storage (TOS)* 6, 2 (2010), 1–22.

[3] Chandranil Chakraborttii and Heiner Litz. 2020. Improving the accuracy, adaptability, and interpretability of SSD failure prediction models. In *Proceedings of the 11th ACM Symposium on Cloud Computing*. 120–133.

[4] G Chandrashekar and F Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.

[5] Evangelos S Eleftheriou, Robert Haas, Xiaoyu Hu, and Roman A Pletka. 2014. Reliability scheme using hybrid SSD/HDD replication with log structured management. US Patent 8,700,949.

[6] Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso. 2007. Failure trends in a large disk drive population. (2007).

[7] B Schroeder, R Lagisetty, and A Merchant. 2016. Flash reliability in production: The expected and the unexpected. In *14th {USENIX} Conference on File and Storage Technologies 16)*. 67–80.

[8] B Schroeder, R Lagisetty, and A Merchant. 2017. Reliability of NAND-based SSDs: What field studies tell us. *Proc. IEEE* 105, 9 (2017), 1751–1769.

[9] Yong Xu, Kaixin Sui, Randolph Yao, Hongyu Zhang, Qingwei Lin, Yingnong Dang, Peng Li, Keceng Jiang, Wenchi Zhang, Jian-Guang Lou, et al. 2018. Improving service availability of cloud systems by predicting disk error. In *2018 {USENIX} Annual Technical Conference ({USENIX}{ATC} 18)*. 481–494.