Published invited paper at International Electron Devices Meeting (IEDM) 2019: https://ieeexplore.ieee.org/document/8993481

Enabling Prognostics of Robust Design with Interpretable Machine Learning

Jay Sarkar and Cory Peterson
Advanced Technology Development, Western Digital Corporation, California, USA.
email: {jay.sarkar, cory.peterson}@wdc.com / jsarkar@utexas.edu

***Abstract*—** Design of robust systems needs to fully account for reliability physics, operational stresses and interactions thereof - while accommodating range of stresses from qualification to field. This research demonstrates the method of empirically analyzing system-internal parametric data of Solid-State Storage devices (SSD) with Machine Learning (ML). ML is shown to be a necessary, effective and novel means of proactively assessing and interpreting prognostics of the resilient system design. The methodologies and results also bear strong relevance to assessment of current and future designs for evolving usage models and new application areas.

I. Introduction

SSD technology is finding increasing adoption across enterprises, hyperscale data centers, high-performance computing architectures and consumer devices. Robust design of SSD memory subsystem can be modeled as an integrated resilient system comprising NAND/NVM (non-volatile memory) media with co-working controller-firmware enacted architecture to mitigate naturally expected defect growth over lifetime [1-2]. Shown in Fig. 1, the population robustness of SSD memory subsystem under data throughput stress has been empirically evaluated and modeled as an integrated system, from perspectives and methodologies of classical system reliability [3]. In this paper, we focus on the potential of SSD-internal parametric data to enable prognostics and enhanced functionalities at the level of individual SSDs. There are significant practical benefits that can result from such an approach, such as individual SSDs proactively communicating prognostic information to the host(s) and associated tailored usage models. Crucially, if such parametric data is leveraged in alignment with system design and reliability physics such as to naturally embed fundamental causality in the prognostics, meaningful prediction frameworks spanning architectures and tuned to specific designs can be achievable.

Robust design of SSD memory subsystem is predicated on a distributed architecture of internal data layout across silicon components (i.e. NAND Flash memory dies) and electrically isolable units (i.e. erase blocks) within such components, as shown in Fig. 2. Such a design of parallelized internal data layout achieves simultaneous benefits of enabling high throughput of data access for the host, as well as, a high system-level robustness. The latter is achieved in SSD design with distributed resilience schemes leveraging independent, uncorrelated probabilities of generated defects across physically separate NAND dies and electrically isolable erase blocks, given sufficiently high-quality silicon. Associated considerations include NAND flash organization and operational idiosyncrasies (such as lack of direct overwrite), performance and application usage models requirements [1-2].

II. Machine learning framework

A. Context of Design and Reliability Physics in Data

Given the distributed architectural design of SSD memory subsystem: parameters representing multiple internal units (i.e. blocks, planes, dies) along with proactive and reactive management of tolerated defects - naturally result in high dimensional data representative of an SSD as a system. This high-dimensional parameter set lends well to a supervised ML classification framework - where the state of an individual SSD can be described in terms of the parameters fashioned as ML "features". For prognostics, labeling each drive sample with their state of health (e.g. healthy, degraded, or failed) as the "target" ML variable can therefore complete a supervised classification framework based on appropriate training data. It is however apparent that the quality and context of training (and validation) data is key to the process of algorithmic learning, especially when such data describes a complex engineered system, such as SSD. In such a context, training on observational or noisy data can run the risk of fitting a model to data that may reflect an unknown, embedded design flaw or a missed validation of complex system design. This is because a trained ML model does not necessarily nor automatically reflect a causal relationship between the feature set and the target variable; yet can strongly leverage correlational relationships to deliver predictions. Hence, in our approach, we obtained clean, noiseless data from a design of experiment (DoE), backed by physical characterization and well understood failure physics. On the other hand, challenges and opportunities in ML-based prognostics on large-scale SSD populations in data-centers have been effectively explored in [4], wherein notable differences between SSD design classes and specific models were clearly evident.

For enabling prognostics fundamentally aligned with reliability physics and robust system design, data from our DoE resulted from accelerated throughput stresses along usage model perspectives, on $N = 120$ enterprise-class SSDs with mature design and qualification. Fig. 3 shows Throughput (i.e. Drive Writes per Day or, Program/Erase Cycles per day) Acceleration Coefficient (TAC) stress causing SSDs to fail within designed lifetime, from fundamentally well understood physics of word-line shorting precipitated by high cycling rate of NAND program operations [3, 5]. Applying throughput and JEDEC JESD219 enterprise [7] workload accelerated stress for sufficiently long duration of 1.25 years, significant and expansively distributed growth of program status failures, defective blocks and die failures eventually resulted in breaching the distributed resilience inherent in robust SSD design. Fig. 4 affirms that despite the fundamentally stochastic failure mechanism of shorted word-lines, component Failure in Time (FIT) rate correlated with the system-level stressor.

Thereafter, detailed failure analysis of samples as exemplified in Fig. 5, accompanied definitive labeling of the ML target variable as either healthy or failed. Thus, population system reliability was ascertained in coherence with the stress-strength interaction principle of reliability science, as shown in Fig. 6.

B. Machine Learning Model

The focus on aligning algorithmic training with SSD design and reliability physics makes a highly interpretable ML model a crucial necessity [5]. For good interpretability balanced with high performance; nonlinear, high-dimensional feature space; natural feature interaction; and limited availability of data samples - Random Forest (RF) was chosen as the ML model [6]. RF offers the balance of bias and variance with natural resistance to over-fitting and has found significant applications across other similar domains of statistical learning.

III. ML Prognostics Results and Interpretation

The classification results measured by various ML metrics are shown in Table I, showing excellent (> 90 %) performance. This result demonstrates the possibility of prognostics with embedded causality, by being aligned with SSD design and physics. The full feature set's categories spanned counts of proactive resilience function activations, managed defect measures and reactive resilience function activations [6]. Feature importance metrics were used as a methodologically sound and practically meaningful technique for model interpretation. While the originally available set comprised ~ 40 features, it was condensable to ~ 25 features based on RF's Gini Index based feature importance ranking, and also interpreted based on design knowledge, as shown in Fig. 7. A further significant feature set compression was possible with Permutation Feature Importance (PFI), by excluding features with PFI ≤ 0, and resulting in a succinct set of ~10 features, as also shown in Fig. 7. Such principled methods of feature and dimensionality reduction, and associated model compression, significantly improves model interpretability. These methods are superior for model interpretability in alignment with domain knowledge, relative to other methods of dimensionality reduction (such as Principle Components Analysis) - by retaining a subset of the original, contextually meaningful features. In this problem domain, the feature set is naturally multicollinear by virtue of resilient system design [8], enabling such compression while retaining model performance.

In this context of resilient design, evident dominance of features associated with proactive firmware resilience function activations is in accordance with robust design. As throughput and workload stresses generated more physical defects, proactive resilience functions more frequently activated prior to eventual failures. Conversely, proactive resilience activations are also most predictive of classifying failures from survivors. This fact is also clearly affirmed in the example distribution analysis shown in Fig. 8, where *more* of the failing samples show *greater* activation of the resilience function. This shows that the distributions are not simply different between the classes (thus enabling ML classification), but they are also meaningfully aligned to the context of resilient SSD design, applied DoE stress and interactions thereof. On the other hand, individual counts of managed physical defects of NAND contain less predictive information because of the distributed, resilient architecture of SSDs [1,2,6,8]. Thus, any expectation of grown defective block counts alone being predictive of SSD health is simplistic without consideration of the larger context of distributed design and architecture. The fact that the two ML classes lack fully non-overlapping distributions for any of the features, as in Fig. 8, further underlines the distributed architecture, as well as the necessity and efficacy of ML for prognostics [6]. Fig. 9 shows a visualization of two key informative features, illustrating their decision boundary suited for non-linear ML models, including deep neural networks [6].

Temporal prognostics of individual SSD health is shown in Fig. 10, by assessing classification metrics from historical parametric data gathered during the DoE. The observed gradual degradation in metrics is expected because of diminishing classifiable information content in the features at milder degradation states, when the eventually failing samples more closely resembled the less degraded samples. Based on such prognostics, effective communication between individual SSD to the host can be possible prior to the point of actual failure. Workload effects have also been interpretably analyzed in [8].

IV. Summary and Conclusions

We have discussed our interpretable approach to, and demonstration of, design and physics aligned ML based prognostics of SSD technology. We point out that such an approach embeds causality in the prognostics by ML, and additionally aids the validation of robust system design through the interpretation process. Thereby, the methodology lends well to a general framework for analysis, prognostics and tailored usage of complex system designs, such as for SSD technology.

Acknowledgment

The authors gratefully acknowledge the support of the Quality organization at Western Digital.

References

[1] Y. Cai, S. Ghose, E. F. Haratsch, Y. Lui, O. Mutlu, "Error Characterization, Mitigation, and Recovery in Flash-Memory-Based Solid-State Drives", Proc. of the IEEE, pp. 1666-1704, Sep. 2017.

[2] N. R. Mielke, R. E. Frickey, I. Kalastirsky, M. Quan, D. Ustinov, V. J. Vasudevan, "Reliability of Solid-State Drives Based on NAND Flash Memory", Proc. of the IEEE, pp. 1725-1750, Sep. 2017.

[3] J. Sarkar, C. Peterson, Y. Zhang, S. Lock, "Robust Error-management and Impact of Throughput in Solid State Storage – Characterization and First System Reliability Model", Proc. of the 2017 Annual Reliability and Maintainability Symposium (RAMS '17), pp. 1-6, 2017.

[4] I. Narayanan, D. Wang, M. Jeon, B. Sharma, L. Caulfield, A. Sivasubramaniam, B. Cutler, J. Liu, B. Khessib, K. Vaid, "SSD Failures in Datacenters: What? When? and Why?", Proc. of 9th ACM International Systems and Storage Conference (SYSTOR '16), Article No. 7, 2016.

[5] C. Molnar, *Interpretable Machine Learning*, 1st Edition, 2019.

[6] J. Sarkar, C. Peterson, A. Sanayei, "Machine-learned assessment and prediction of robust solid state storage system reliability physics", Proc. of the 2018 IEEE International Reliability Physics Symposium (IRPS '18), pp. 3C.6-1 - 3C.6-8, 2018.

[7] Solid State Drive (SSD) Endurance Workloads - JESD219A, JEDEC Solid-State Technology Association, July 2012.

[8] J. Sarkar, C. Peterson, "Operational Workload Impact on Robust Solid-State Storage Analyzed with Interpretable Machine Learning", Proc. of the 2019 IEEE International Reliability Physics Symposium (IRPS '19), pp. 1-8, 2019.

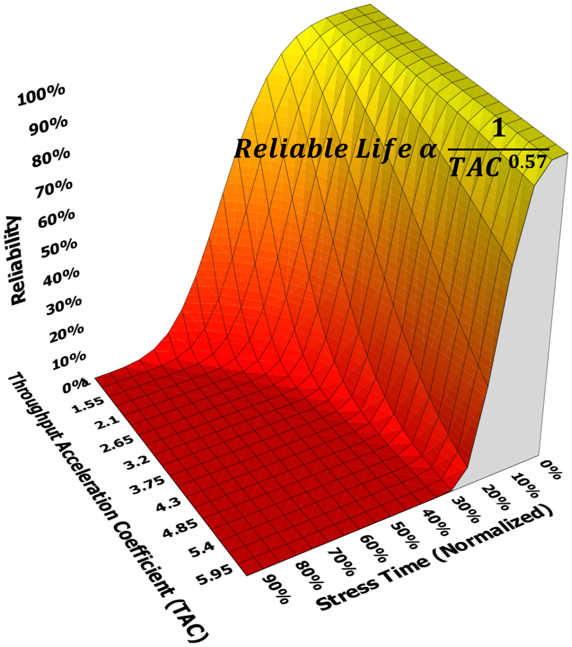

Fig. 1. Surface of population system reliability as a function of throughput acceleration coefficient (TAC) and stress duration, derived from DoE data and modeled as an inverse power law [3].

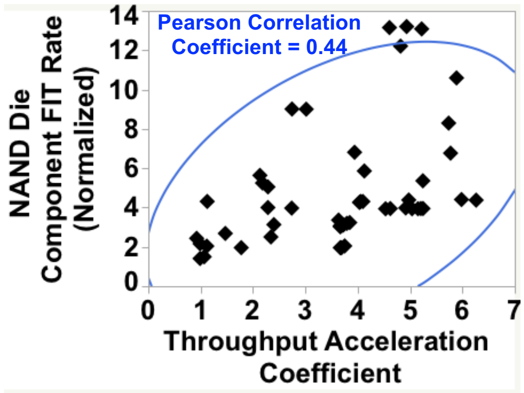

Fig. 4. Component NAND failure FIT rate correlated with host-level stressor metric of TAC (or, Program Erase Cycling rate), affirming the physics of system stress [3].

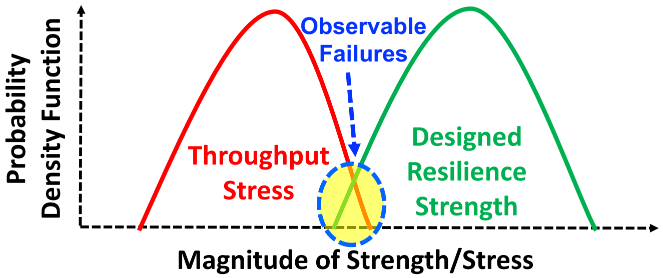

Fig. 6. Conceptual underpinning of stress-strength interaction in precipitating system-level failures of the robust system. The region of interaction between applied throughput and workload accelerated stresses, and the designed resilience strength of the system (SSD), led to the precipitated failures from the DoE.

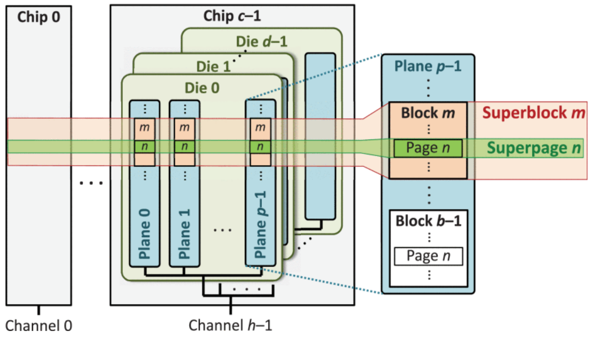

Fig. 2. Flash memory organization in a distributed architecture within solid-state storage, reproduced from [1]. © 2017 IEEE

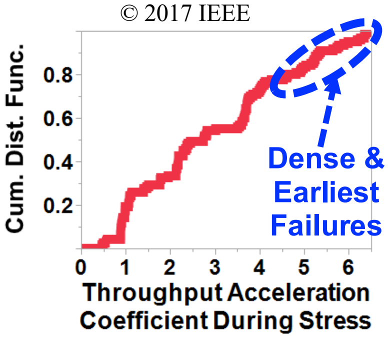

Fig. 3. Usage model aligned spectrum of applied stressor, with failures resulted from physics aligned understanding of robust system design and backed by causational failure analysis [3].

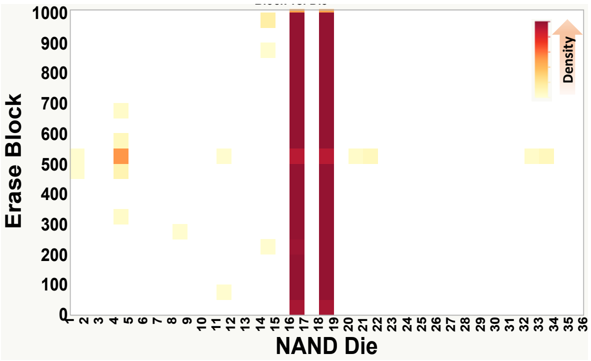

Fig. 5. Topography of grown defects across span of NAND dies and erase blocks of a representative SSD in final failed state. The accelerated stress driven defects were largely stochastically distributed, except for dies 16 and 18 on different channels of this sample SSD bearing highly dense and broad defect distributions.

ML Classification Metric	Performance
Accuracy	94 %
True Positive (Recall)	95 %
False Positive Rate	6 %
True Negative Rate	94 %
False Negative Rate	5 %
Precision	92 %
F1 Score	93 %

Table I. ML classification performance for failures (“positive” ⇔ failing samples) from throughput accelerated stress of DoE.

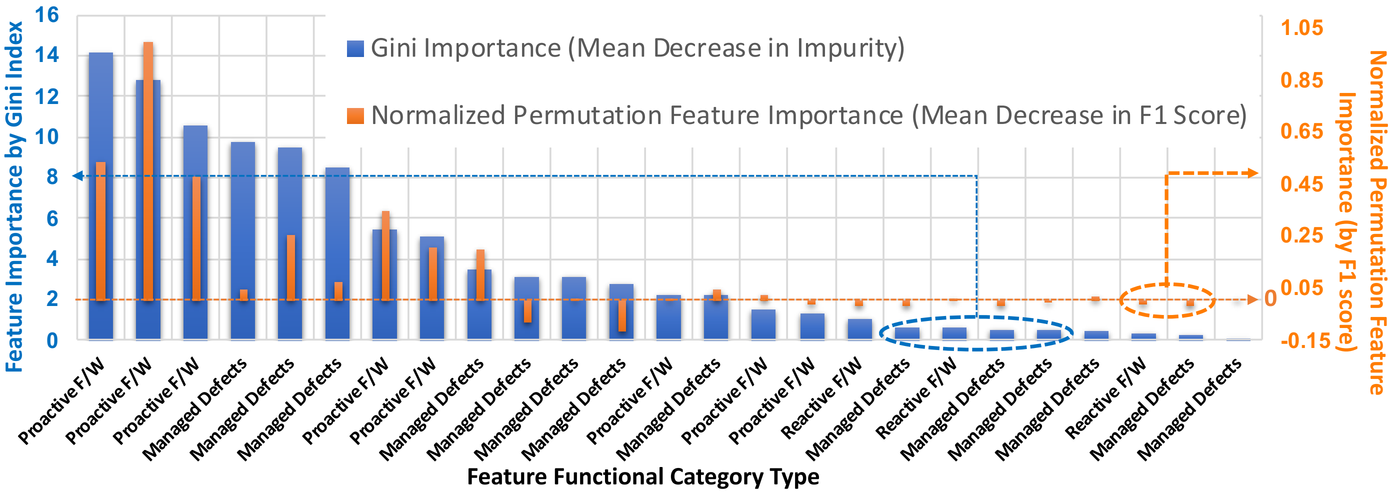

Fig. 7. Feature importance assessments by Gini Index measure of RF (left axis) versus Permutation Feature Importance (right axis), illustrating an effective approach to model interpretation and feature-set compression. Removing features with PFI ≤ 0 enables significant feature reduction, while retaining a subset of original, contextually meaningful features and model performance (F/W ⇔ Firmware implemented resilience).

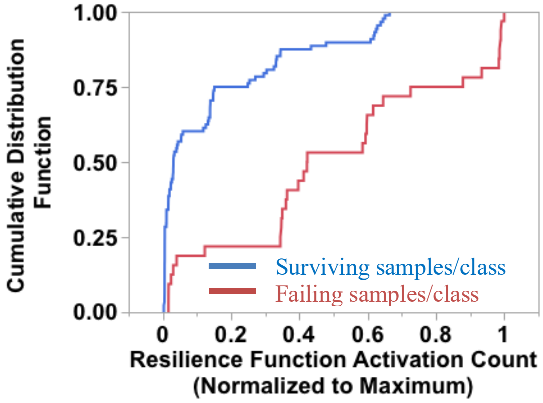

Fig. 8. Normalized ML feature denoting an architected resilience function activation during DoE stress illustrates partial distribution-level distinction between failing and surviving samples/classes. While all ML features bear such significantly overlapping distributions, more of the failing samples have larger values than the surviving samples.

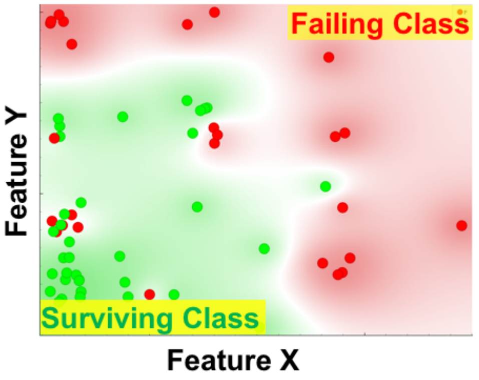

Fig. 9. Non-linear decision boundary in 2-D feature space illustrated for two of the features, implying optimality of non-linear ML models for classification including deep neural networks, aside from Random Forest.

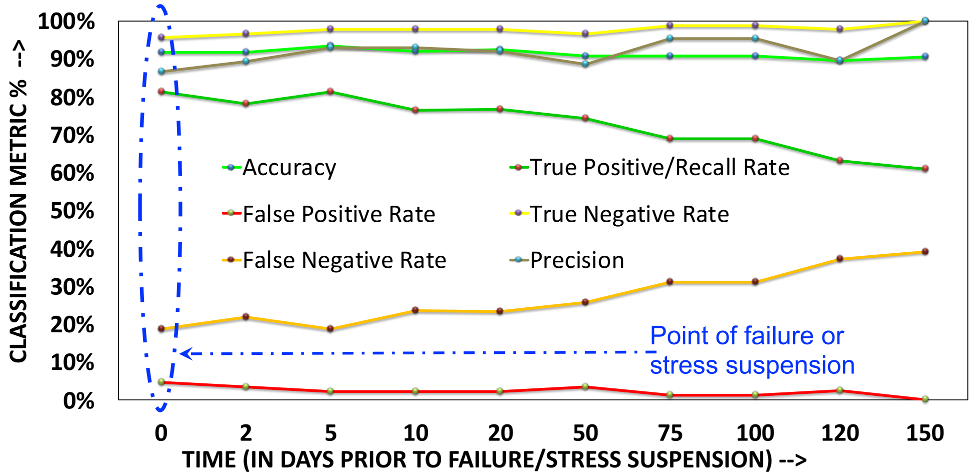

Fig. 10. Temporal prognostics in terms of ML performance metrics evaluated on historical parametric data prior to actual failure or stress suspension, demonstrating predictive health assessment of individual SSDs. This conservative assessment was made based on mixed write workload stresses from the industry standard JESD219 enterprise and a pseudo-sequential workloads (the latter with highly imbalanced classes due to few failures, degrading the model performance) [6]. Such workload impacts have been separately analyzed with interpretable ML in [8].