A Machine Learning based Disk Health Monitoring for Supporting Always-On Extreme Scale Storage Systems

Song Fu*, Hsing-bung (HB) Chen** and Zhi (George) Qiao*

*Department of Computer Science and Engineering, University of North Texas
**HPC-DES Group, Los Alamos National Laboratory

Introduction

Computations and simulations help advance knowledge in science, energy, and national security. Over the years, they have become more accurate to generate more realistic outcome, and as a result, the demand for computational power and much larger storage system also increased. A typical HPC simulations on LANL's Trinity requires hundreds of PBs of data to be written out in order to capture the entirety of simulation data.

At such scale, disk failures and associated data loss become the norm, and data recovery process is worsen due to the increased disk capacity. For a helium-filled hard drives, an extensive rebuild time could be approaching 5 days. In large RAID arrays, a second disk drive can fail before the first is rebuilt, which result in data loss and significant performance degradation.

Meanwhile, existing storage systems are mostly passive. But with increasing performance and decreasing cost of processors, storage manufactures are adding more system intelligence at I/O peripherals. The processing capability is provided at the storage enclosure level, which indicate the possibility of offloading certain services to storage systems.

Desired Features

We envision the extreme scale storage system would include following features.

1. **Always-on:** the service is always on so data availability is guaranteed all the time. Failure will mitigate by system and cause little performance degradation to services and applications.
2. **Active and intelligent:** processing capability are enabled at storage enclosure level so that a drive can participate in service and management.
3. **Automatic repair:** as failure become norm and recover time extend, automatic repair will become standard mode of operation for fixing corruptions and failures.
4. **Aforehand replacement:** Leverage the machine learning methods to replace disk before its failure point, and rescue data aforehand, so it never lose or damage.

Methodology

We propose a Machine Learning based Disk Health Status Assessment, Failure Prediction, and Pre-Failure Data Recovery Approach for supporting Always-On Extreme Scale Storage Systems. We prototype a new fault-resilience solution on the ZFS file system, Key-Value storage, or Object Storage Systems (OSS). Our solution includes following components.

Active Storage

We characterize and model the performance, reliability and power consumption of active storage systems built from HGST Open Ethernet Drives (OED). Each OED consist of a ARM CPU, RAM, block storage drive, with a standard 3.5" HDD form factor that have Ethernet connectivity and running Linux OS. The Active Storage can offload certain data management task from compute node, such as encoding / decoding erasure code segment at each drive. OEDs perform as small Linux servers yet consume same amount of power as hard drives. Aggregated computation power and low energy cost make OED ideal for future active storage systems.

PASSI

A Parallel, Reliable and Scalable Software Infrastructure (PASSI) is designed for active storage systems [1]. It harness the processing capacity of each disk drive to perform parallel erasure coding, which provides cost-effective data integrity. In PASSI, each object is encoding to $K+M$ segments where K are the original data segments and M are the redundant code segments. It then evenly distributed among $K+M$ OEDs to achieve balanced data placement. PASSI erasure coding scheme can tolerate a loss of up to M data/code segments.

ZFS+

Leverage the ZFS filesystem on extreme scale storage to ensure high performance, high scalability, and storage resilience to disk failures, memory errors, and silent data corruptions. We work on the ZFS's Storage Pool Allocator (SPA) to include MLFP module for drive pre-failure replacement and recovery.

MLFP

The Machine Learning based disk Failure Prediction (MLFP) project [2] includes a tool for monitoring disk *SMART* data, and a runtime for assess the health status and predict failure timing. Based on the result, it can proactively/aforehand replace a failing disk and clone the data to new disk before the failure occurs.

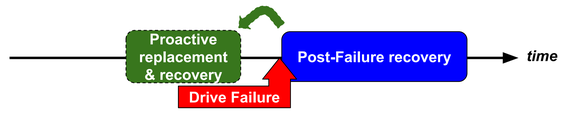


Figure 1: Proactive replace & recover eliminate the heavy calculation during disk rebuild, just clone & delta-resilvering

MLFP is enhanced to gain intelligence towards failure-preventative fault management, and proactive pre-failure data rescue. It eliminates the time-consuming, expensive disk rebuilds and disk repair activities, therefore minimize the performance degradation caused by post-failure data recovery and ensures data availability.

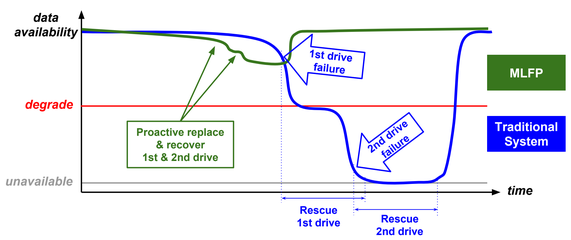
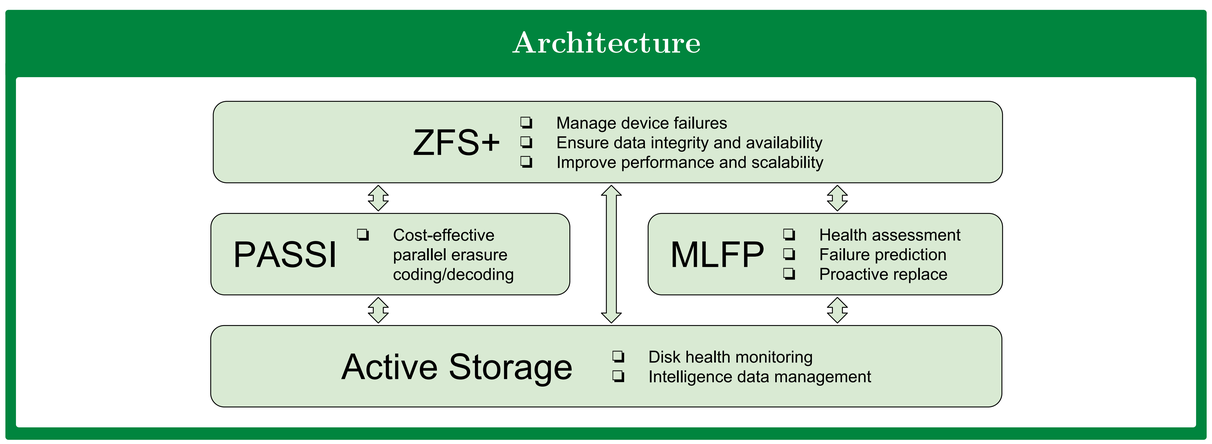


Figure 2: Proactive replacement algorithm ensures data availability

Architecture



Conclusion

We aim to improve the reliability and scalability of storage systems that extreme scale science requires by supporting design and prototyping of the next-generation of active storage environments. To this end, we integrate storage system analysis, machine learning algorithmic design, and system prototyping implementation. Our design help build high-performance, scalable, and energy-efficient extreme scale storage systems and support for intelligent extreme scale scientific computing and knowledge discovery. It can significantly impact scientific computing at an extreme scale by ensuring the storage systems that can be counted on for availability and data integrity. This allows large scientific applications to run a correct solution efficiently.

References

[1] Hsing-bung Chen and Song Fu.
Passi: A parallel, reliable and scalable storage software infrastructure for active storage system and i/o environments.
In *the 34th IEEE International Performance Computing and Communications Conference (IPCCC)*, pages 1–8, 2015.

[2] Hsing bung Chen.
Mlfp - a machine learning based disk failure prediction on hpc storage systems.
The PathFinder project, Los Alamos National Lab, 2017.

[3] Robert Ross and Scott Klasky et al.
Storage systems and input/output to support extreme scale science.
DOE Workshops on Storage Systems and Input/Output, 2014.

[4] Song Huang, Song Fu, Quan Zhang, and Weisong Shi.
Characterizing disk failures with quantified disk degradation signatures: An early experience.
In *IEEE International Symposium on Workload Characterization (IISWC)*, pages 150–159. IEEE, 2015.

[5] Hsing-Bung Chen and Song Fu.
Improving coding performance and energy efficiency of erasure coding process for storage systems-a parallel and scalable approach.
In *the 9th IEEE International Conference on Cloud Computing (CLOUD)*. IEEE, 2016.

Acknowledgements

This publication has been assigned an LANL identifier LA-UR-17-23134.