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.

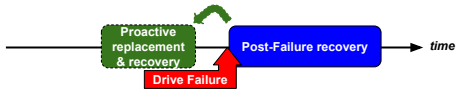


Figure 1: Proactive replace & recover eliminate the heavy calculation during disk rebuild, just clone & delta-resilvering

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.

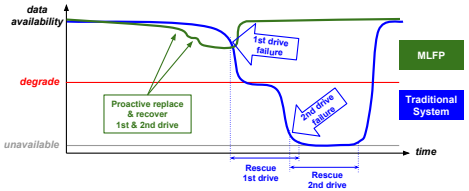
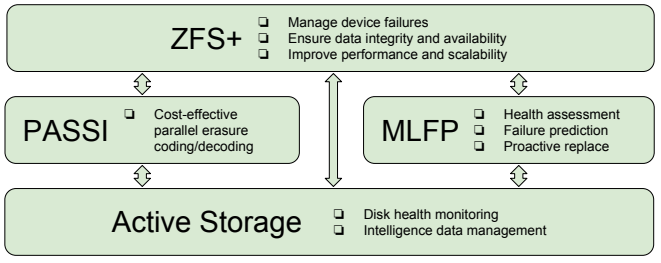


Figure 2: Proactive replacement algorithm ensures data availability

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.

A Machine Learning based Disk Health Status Assessment, Failure Prediction, and Pre-Failure Data Recovery Approach for Supporting Always-On Extreme Scale Storage Systems

Extended Abstract

Song Fu[1], Chen, Hsing-bung (HB) Chen[2] and Zhi (George) Qiao[1]
[1] Department of Computer Science and Engineering, University of North Texas,
[2] HPC-DES Group, Los Alamos National Laboratory
song.fu@unt.edu, hbchen@lanl.gov and zhiqiao@my.unt.edu

ABSTRACT

High performance I/O is critical as storing and retrieving an immense amount of data can greatly affect the overall performance of the system and applications. As exabytes of data need to be stored on hundreds of thousands of disk drives, disk failures and associated data loss become the norm. Time to rebuild a failed drive is extended due to the increased disk capacity, resulting in hours of data processing disruption. In this paper, 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 (IDFP). The goal is first to leverage the machine learning methods on extreme scale storage's fault management problem and is second to meet the scaling and resilience need of extreme scale science by ensuring that storage systems are pervasively intelligent, always available, never lose or damage data.

1. INTRODUCTION

Computations and simulations help advance knowledge in science, energy, and national security. As these simulations have become more accurate and therefore more realistic, their demand for computational power has also increased. This results in growth of simulations to, which in turn increases the demand for much larger storage systems. For instance, launching a suite of 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. Therefore, high performance I/O becomes critical since storing and retrieving such an immense amount of data can greatly affect the overall performance of the system and applications.

Existing storage systems are mostly passive, which means that the disk drive's stage and drain memory/burst buffer checkpoints with little intelligence. With the increasing performance and decreasing cost of processors, storage manufacturers are adding more system intelligence at I/O peripherals. The processing capability is provided at the storage enclosure level, while the end disk drives are still passive. We envision that future HPC storage systems will be clever and can provide system intelligence at each disk drive so that a drive can participate in storage services and management [1].

Additionally, since exabytes of data need to be stored on hundreds of thousands of disk drives, storage scalability presents a large challenge. At such a scale, disk failures and data loss become the norm in exascale storage environments. As data generated by HPC applications, such as checkpoint/restart data sets and disk drives become larger the time to rebuild a failed drive is extended, resulting in hours of data processing disruption. For helium-filled hard drives, the larger capacity and slower increase in performance result in an extensive rebuild time—calculated to be approaching five days. In large arrays, a second disk drive can fail before the first is rebuilt, which results in data loss and significant performance degradation. We envision that dealing with data corruptions and disk failures will become the standard mode of operation for storage systems; additionally, these tasks will be performed by machines with little human intervention due to the overwhelming storage scale. Therefore, it should cause little data access or performance degradation to systems and applications.

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 (IDFP). The goal is first to leverage the machine learning methods on extreme scale storage's fault management problem and is second to meet the scaling and resilience need of extreme scale science by ensuring that storage systems are pervasively

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

intelligent, always available, never lose or damage data and energy-efficient.

2. METHODOLOGY

The proposed IDFP system a) explores the machine learning models/methods on Disk Health Status Assessment and Disk Degradation/Failure Prediction [4][5][7], 2) provides a proactive solution for Pre-Failure Data Recovery to support always-On Extreme Scale Storage Systems [8][9], and 3) provides a new prototype of fault-resilience solution on the ZFS [2] file system (commonly used on Lustre and object storage systems), Key-value-store [10][11], or object storage systems [12]. We conduct the following research tasks.

(1) Characterize and model the performance, reliability and power consumption of active storage systems built from disk drives. The derived models are then utilized to develop IDFP storage software I/O infrastructure.

(2) Develop a self-aware storage monitoring software tool to ensure storage resilience to disk failures, memory errors and silent data corruptions.

(3) Develop intelligent disk replacement algorithms, failure-preventive fault management methods, and proactive pre-failure data rescue methods to reduce time-consuming, expensive disk rebuilds and disk repair activities.

(4) Explore the processing capacity of each disk drive to design a parallel erasure coding mechanism for cost-effective data integrity. The generated erasure data segments are distributed evenly among the drives.

(5) Utilize machine learning based disk degradation result and prototype a new fault-resilience enhancement software on the ZFS file system, KV-Store, or object storage systems.

3. CONCLUSIONS AND IMPACT

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. We design enabling technologies that help build high-performance, scalable, and energy-efficient extreme scale storage systems and support for intelligent extreme scale scientific computing and knowledge discovery. IDFP 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, Proc. of IEEE IPCCC Conference, 2016.

[2] Don Brady & Justin Gibbs, ZFS fault management, 2016 openZFS Developer Summit.

[3] DOE 2014 report, Storage Systems and Input/Output to Support Extreme Scale Science, Report of the DOE Workshops on Storage Systems and Input and Output, 2014.

[4] Hsing-bung Chen, MLFP – A Machine Learning based Disk Failure Prediction On HPC Storage Systems, The PathFinder project, Los Alamos National Lab, 2017.

[5] Song Huang, Song Fu, Quan Zhang and Weisong Shi, Characterizing Disk Failures with Quantified Disk Degradation Signatures: An Early Experience, Proc. of IEEE IISWC Conference, 2015.

[6] The Opportunities and Challenges of Exascale Computing, Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee, DOE, 2010.

[7] Mirela Botezatu, Ioana Giurgiu, Jasmina Bogojeska, Dorothea Wiesmann, Predicting Disk Replacement towards Reliable Data Centers, Proc. of ACM KDD Conference, 2016.

[8] Hsing-bung Chen and Song Fu, Parallel Erasure Coding: Exploring Task Parallelism in Erasure Coding for Enhanced Bandwidth and Energy Efficiency, Proc. of IEEE NAS Conference, 2016.

[9] Qin Xin, Understanding and Coping with Failures in Large-Scale Storage Systems, UC-Santa Cruz, Technical Report UCSC-SSRC-07-06, 2007.

[10] Peng Wang Guangyu Sun, Song Jiang, Jian Ouyang, and Shiding Lin, An Efficient Design and Implementation of LSM-Tree based Key-Value Store on Open-Channel SSD, Proc. of ACM EuroSys Conference, 2014.

[11] Leonardo Marmol, Swaminathan Sundararaman and Nisha Talagala, Raju Rangaswami, NVMKV: A Scalable, Lightweight, FTL-aware Key-Value Store, Proc. of USENIX FAST Conference, 2015.

[12] Object Storage A Fresh Approach to Long-Term File Storage, Dell Technical White Paper, Dell Inc., 2010.