Sunil Template

Contents

Chapter 1

The Convergence of Enterprise, Internet Scale, and High Performance Computing Storage Infrastructures

Jay Lofstead

Sandia National Laboratories

Eric Barton

Intel

Matthew Curry

Sandia National Laboratories

Carlos Maltzahn

University of California, Santa Cruz

Robert Ross

Argonne National Laboratory

Craig Ulmer

Sandia National Laboratories

Abstract

Large scale storage infrastructures have been significantly impacted by the growth in data analytics applications. High Performance Computing storage infrastructures, once the extreme end of the storage scale spectrum, must now adapt to technologies optimized for large scale data analytics applications. Hardware changes, such as storage class memory, are also affecting how the exascale storage stack will be constructed. We examine use cases, trends, convergent technologies, and new opportunities generated by this technology blending.

1.1 Introduction

HPC infrastructures have grown around the requirement to handle large, decomposed data structures for parallel computation. Single data objects may be as large as 100s TB spread across the entire machine. Paralell storage systems have adequately grown and addressed performance and storage requirements while maintaining backwards compatibility with the standard POSIX interface. Big data application, on the other hand, focus on searching through immense volumes of tiny items looking for patterns or correlations that may lead to insights. Some science applications, such as genomics, have a workload pattern similar to these big data applications.

Traditionally, the HPC market has focused on supporting coherent and consistent output methods from parallel sources to parallel targets. This requirement is driven from validating that the output of a single item is complete and correct. Largely, the workload is write-intensive during the expensive, at scale computation process with a read-intensive phase lasting months on cheaper machines or at small scale with low priority. File systems like the dominant Lustre [**?**] and GPFS [**?**] systems have been carefully optimized to address these workloads.

The big data market has opposite priorities. The big computation phase requires reading in large data quantities for processing at scale. The output from this process can be handled at much smaller scale later and is orders of magnitude smaller. Given the small item focus, the overhead inherent in coherent and consistent storage for write intensive workloads is both unnecessary and a heavy cost. Instead, distributed object-based storage technology has been embraced with independent, uncoordinated dat access. The profit potential for this market has caused an explosion in specialized products aimed at accelerating this style processing. The optimizations targeting this market,

such as Kinect [**?**], offer a native object interface for the devices connected directly to a network.

Adding complexity to this storage environment is the relentless performance improvements and cost reductions for solid state storage, like NAND-based flash memory. These devices have already rendered 15,000 RPM disk drives obsolete. The 10,000 RPM disk drives will not survive for more than a few more years. New disk technology like shingled drives [**?**] offer a path for disks to survive longer. The enormous capacities for read intensive, write infrequently workloads is very attractive for many communities. For example, storing images created sequentially for later read-intensive processing can yield a better cost/performance balance.

We will investigate how the HPC environment can and must adapt to this new storage envionment. We will also consider the planned reintegration of large scale computing from the split of big data applications from simulation-based computing with both the necessary and forced integration of these large, expensive platforms for multi-use.

1.2 Object-Based Stores

placeholder cite [1].

Here we want to talk about how object-based key-value stores are used for big data applications summarizing the specific features that identify this market segment.

1.2.1 Big-Data/General Computing Optimized Object/Key-Value Stores

Wisconsin [Chou, et. al] is the granddaddy 1985. Talk about others and how they differ.

memchached [2003] from LiveJournal.

1.2.2 HPC Oriented Object Stores

Parallel file systems inherently have an object-like layer beneath the surface. The requirement to spread a single file across multiple devices for capacity reasons alone prompts this approach. The actual implementation may vary, such as using individual files within a local file system, each representing part of parallel file. Popular examples include Lustre [**?**], ...,

1.3 Next Generation HPC Storage Systems

Here we want to talk about, at a proposal sort of level, what we think needs to be done.

Talk about the disconnect between metadata and storage and the complications it introduces and some ideas on what we plan to do about it.

Talk about the major efforts

1.3.1 Lustre/DAOS

The FFSIO phase II sort of info to start. Keep it acceptable. This seems to me to be just a pure object store.

1.3.2 Kelpie/Data Warehousing

Native multi-level key-value store

1.3.3 Hybrid Models

SSIO project from ORNL/Sandia

1.4 Conclusions

This is our overall view on things

Acknowledgements

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

1.5 Glossary

Adaptable: An adaptable process is designed to maintain effectiveness and efficiency as requirements change. The process is deemed adaptable when there is agreement among suppliers, owners, and customers that the process will meet requirements throughout the strategic period.

Bibliography

[1] M. Ilyas, I. Mahgoub, and L. Kelly. *Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems*. CRC Press, Inc. Boca Raton, FL, USA, 2004.