MarFS - A Scalable Near-Posix Metadata File System with Cloud Based Object Backend

Gary Grider, Dave Montoya, Hsing-bung Chen, Brett Kettering, Jeff Inman, Chris DeJager, Alfred Torrez, Kyle Lamb, Chris Hoffman, David Bonnie, Ronald Croonenberg, Matthew Broomfield, Sean Leffler, Parks Fields, Jeff Kuehn, John Bent[1]

Los Alamos National Lab,
Los Alamos , New Mexico 87545, USA
Contact: {dmont,hbchen,kyle_lamb}@lanl.gov

EMC Corporation[1]
176 South Street, MA 01748

Abstract – Work In Progress

Many computing sites, LANL being one of them, have a requirement for long-term retention of mostly cold data. Although the main function of this storage tier is capacity, it does also have a bandwidth requirement. For many years, tape was the best economic solution for this requirement. However, over time, data sets have grown larger more quickly than tape bandwidth has improved. We have now entered a regime in which disk is the more economically efficient medium for this storage tier. Also more and more, data dominates the computing world. LANL has been aware of this HPC tiered storage transition and has initiated efforts to extend capabilities to take advantage of architectural changes and also allow current infrastructures and approaches to be preserved. There is a "sea" of data out there in many different formats that needs to be efficiently managed and effectively used. "Mar" means "sea" in Spanish. We introduce a new hybrid storage system named MarFS. MarFS is a Near-POSIX Global Name Space Management Infrastructure using scale-out commercial/cloud for data and many POSIX file systems for metadata services. MarFS is an approach to support a data lake for HPC that sits on industry based commodity storage hardware and is a software layer that provides a global namespace and near POSIX semantics. MarFS provides the capability to serve as an umbrella over a variety of underling storage layers.

In MarFS, currently we use GPFS for namespace metadata and object based store for data storage. Additionally, we improve upon GPFS scalability by introducing a layer of indirection (i.e. MarFS) that allows us to build a single namespace from a collection of GPFS namespaces. We further improve upon GPFS scalability by separating the directory metadata from the file metadata. The directory namespace is sharded by MarFS across a set of GPFS name-space. The file namespace is shared by MarFS across a separate set of GPFS namespace. MarFS provides scalability of name space by sewing together multiple POSIX file systems both as parts of the tree and as parts of a single directory allowing scaling across the tree and within a single directory.

The proposed MarFS Near-POSIX file system has two major contributions:

a) MarFS provides the capabilities for future HPC tiered storage objects systems to provide massive scaling with convenient and efficient erasure coding techniques. People need folders and many of the features that POSIX metadata gives them. Also most software written more than 10 years ago utilizes a POSIX interface and there is insufficient effort to upgrade this software to an Object Storage solution. Leveraging the best of POSIX namespace management in a scalable way over the best of Object Systems has huge economic appeal.

(b) MarFS meets the challenges of future HPC tiered storage requirements. There are many challenges to provide this capability including the mismatch of POSIX and Object metadata, security, update in place semantics, and efficient object sizes. Further, the HPC need includes billions of files in a single directory and single files that are even zeta bytes in size, so scale is a huge factor.

Our early experience of using MarFS has shown promising parallel I/O feature and highly scaling capability. Currently we are deploying the MarFS system on LANL's Trinity machine and we plan to build the campaign storage with 30PB to 100PB capacity. We plan to conduct more field studies of MarFS on the LANL's Trinity machine and will enhance and revise MarFS phase two features mentioned in the MarFS requirement document.