MarFS and libNE Utility Development

Author: *Daniel Perry,* Cornell University

Mentors: *David Bonnie* and *Garrett Ransom*

As the capabilities of high performance computing (HPC) supercomputers continue to grow, so do the sizes of the datasets being computed on these machines. Object storage systems have proven themselves capable of scaling to adequately accommodate storing large amounts of data while also enabling high-speed accesses. However, many users and applications still expect a POSIX filesystem interface, as opposed to the Representational State Transfer (REST) semantics utilized by most object stores. MarFS is an open-source software developed at Los Alamos National Laboratory (LANL) which implements a near-POSIX interface over a scalable multi-component object store that offers few compromises. The laboratory's current production system provides 60PB of storage with access speeds approaching 25 GB/sec. LANL is currently in the middle of a complete rewrite of the MarFS code base to provide stability, functionality, and performance improvements to the filesystem.

An essential component of this new implementation is libNE, a library that handles the multi-component functionality of MarFS through parallel erasure coding to provide both high performance transfers and failure tolerance. Part of libNE is its data abstraction layer (DAL), which allows the use of "hot-swappable" underlying storage systems. In addition to other development work within libNE, we implemented several DALs within the library, such as an AWS S3 and recursive DAL, along with others for testing/benchmarking purposes to complement the default DAL which interfaces with POSIX-based filesystems.

MarFS is intended to be interfaced with in one of two ways: either users access it interactively through a filesystem in userspace (FUSE) mount, or data is transferred through batch jobs handled through pftool, a parallel file transfer tool also developed by LANL. We also created interfaces which allow both of these utilities to interact with the filesystem. These interfaces are designed to ease future development and integration efforts for long-term support and maintenance.

LA-UR-21-27780