U.S. Department of Energy Best Practices Workshop on
File Systems & Archives
San Francisco, CA
September 26-27, 2011
Position Paper

Ruth Klundt
Sandia National Labs
rklundt@sandia.gov

Stephen Monk
Sandia National Labs
smonk@sandia.gov

John Noe
Sandia National Labs
jpnoe@sandia.gov

James Schutt
Sandia National Labs
jaschut@sandia.gov

ABSTRACT

We here present an overview of our current file system strategies, and a brief mention of planning for the future. The focus of the discussion is the link between usability issues and implementation decisions.

INTRODUCTION

Our primary drivers in the design of file system solutions are reliability and performance. In addition we attempt to provide solutions covering a spectrum of user needs, which also include convenience of use, backup capability, and high availability.

Overview of Current File Systems

User requirements in the storage arena are often difficult or impossible to satisfy simultaneously with a single global solution. Simulation codes generate a large quantity of restart data that must be stored quickly, as a defense against system outages. Most of this data is transitory, so does not need to be backed up. Other types of user data such as application codes and input data must be stored reliably. During periods of maintenance, it is important to users for the continuity of their work that some portions of the infrastructure remain available.

At present we maintain three basic categories of storage.

- Site-Wide Parallel File System

  Our parallel file system is implemented using Lustre [1] running on commodity servers, backed by DDN 9900/9550 raid cabinets. This file system serves ~2PB of fast scratch space to 4 different clusters, via LNET routers. Testing is under way on an upgrade to DDN SFA10K hardware providing ~3PB space for the new TLCC2 installation. Software support for Lustre is provided by Whamcloud [2].

- Intermediate NFS File System

  On all clusters, a large storage space is delivered by means of Sun Unified Storage (7410) using ZFS. This is not purged, and not backed up.

- Traditional NAS

  Less than 100TB, provided by NetApp hardware backed up to corporate archives. Stable, safe, slow location serving user /home and /projects space commonly across the clusters.

Usability Impact

The parallel file system satisfies the need for fast storage of large data sets. Although no backups can be done at this size, all possible efforts are made to avoid data loss, by means of hardware RAID configurations and continuity by means of Lustre failover and Multi-path IO. The local Red Sky Lustre implementation, which requires use of software RAID on the Sun equipment, has encountered some difficulties due to increased operational demands and is slated to be shutdown in favor of site file systems.

The intermediate NFS file system provides an alternate location for users to continue work during maintenance periods on the parallel file system. The longevity of the Sun 7410 platform is not clear given the lack of a clear hardware roadmap from Oracle. Although it has proven to be a solid product within this role, we are moving to a solution that is less of a "black box" from the view of the hardware (see below).

The NetApp filers serving /home and /projects have a fairly long history of providing robust reliable service here, although of limited size. New or different solutions have a high bar to meet in order to be considered as replacements for this functionality.

Future Plans

- GPFS NAS

  Some DDN 9550 cabinets are currently being re-purposed for use with IBM's GPFS file system [3] as an alternate highly available storage space, implemented at minimum cost. Production deployment is imminent.

- Ceph

  An effort is in progress to test the robustness, usability, and performance of the Ceph file system [4]. Early results show promise for this open source solution as a potential alternate in the NAS file system space in the near future. In addition, a variety of use cases other than HPC are being actively explored elsewhere, such as the ability to export as NFS, integration with PNFS [5], and access via user space clients. Interest in Ceph from disparate data storage venues can only improve the robustness of the implementation, and a broad user base provides some confidence that the file system has a productive future ahead.

  Some key design elements that make Ceph a high performance file system of interest:

  - Workload scalability (lots of servers/clients)
  - On-line expansion (easy to add capacity and performance)
  - Data replication (fault tolerance without RAID controllers)
  - Adaptive meta-data server (scalable)
  - Ability to reliably use commodity storage platforms

In conjunction with the Ceph testing effort, a heterogeneous test bed is being expanded and shared as a release test platform for production machines.

CONCLUSIONS

Challenges in maintaining multiple types of storage might be mitigated in the future, with improvements in current parallel file systems with respect to reliability and availability. Ideally a single global file system solution with pools of storage configured for different use cases would streamline the delivery of the disparate services needed. A single solution capable of providing sufficient bandwidth to parallel platforms, differential backup capabilities, and 24/7 availability to users does not yet exist.

REFERENCES

1. Lustre http://www.lustre.org
2. Whamcloud http://whamcloud.com
3. GPFS http://www-03.ibm.com/systems/software/gpfs/
4. Ceph File System http://ceph.newdream.net
5. PNFS http://www.pnfs.com