U.S. Department of Energy Best Practices Workshop on

File Systems & Archives

San Francisco, CA

September 26-27, 2011

Position Paper: Reliability and Availability

John Gebhardt
Lockheed-Martin/U. S. Air Force Research Laboratory
DoD High performance Computing Modernization Program
john.gebhardt@wpafb.af.mil

Thomas Kendall
U. S. Army Research Laboratory
DoD High performance Computing Modernization Program
thomas.m.kendall4.civ@mail.mil

Cray J. Henry
DoD High performance Computing Modernization Program
cray@hpcmo.hpc.mil

ABSTRACT / SUMMARY

The DoD High Performance Computing Modernization Program (HPCMP) has implemented a multilayered storage approach to cost effectively meet the storage needs of a diverse customer base. Users' can wait in the batch queue indefinitely (but typically start within seventy-two hours) and then can run for up to fourteen days (or longer with special arrangements). To maximize systems availability, several layers of storage and storage use policy are implemented.

INTRODUCTION

The HPCMP has layered several storage and file systems within the environment. Each system has specific reliability and availability characteristics and use policy driven by system availability requirements.

This paper discusses the reliability and availability of the following types of storage constructs within the HPCMP:

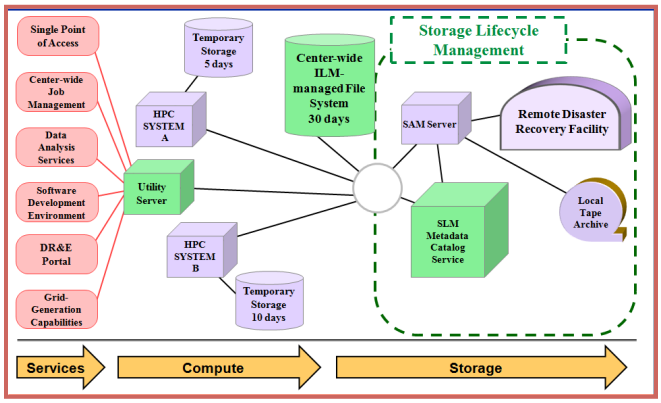
- HPC scratch space file system
- HPC Home and Applications file system
- Root services file system
- Center Wide File System (CWFS)
- Lifecycle Management System
  - Archive system
  - Tape storage

Data Center Facilities

The Department of Defense Supercomputing Resource Centers (DSRC) operates twenty-four hours a day, seven days a week. Each of the DSRCs utilize different combinations of UPS, redundant commercial power feeds and diesel backup power generation capabilities to allow operations to continue through minor power fluctuations and allow for graceful equipment shut for prolonged power outages. In the event of a prolonged power or cooling failure, procedures are activated to shutdown the systems, which in turn quiesces the storage. This approach nearly eliminates unplanned, abrupt outages.

. **File System Overview**

Each DSRC hosts a CWFS which supports the lifecycle management system, the utility server and each HPC system. Each HPC system typically includes a combination of RAID storage devices that are logically decomposed into three file systems -- Scratch space, Home and Applications and Root services. The utility server is similarly configured and the lifecycle management system includes multiple data stores, archival servers and agents described later.



Center Wide File System

The HPCMP has recently deployed Center Wide File Systems (CWFS) at each DSRC. The purpose of the CWFS is to provide users with a fast central storage capability that can be easily accessed by all major HPC systems and servers within the center. It is intended to serve as the "near-HPC" intermediate storage between scratch file system on each HPC system and the long-term archive. It has been sized to providing a minimum of thirty days of quick intermediate storage. Through CWFS users can move their entire data sets among the scratch file systems and the archive file system. They can perform pre and post processing on their data conveniently avoiding the slower access times associated with archived data. This approach affords users the time necessary to make more thoughtful decisions on what data, for example after a large run, really needs to be archived and what can be deleted.

The CWFS is not backed up. It does contain all the redundancy features of HPC scratch with the addition of check sums on read from storage to host.

HPC Scratch Space File Systems

The HPC systems are in high demand. The data sets used to set up runs and the data resulting from runs is very large and very transitory. In order to assure there is sufficient scratch space to stage the next job, HPCMP policy allows for user data to exist on HPC scratch storage for 10 days. Within ten days after data creation, users must move their data to the center wide file system or archive. After ten days the data it is subject to removal to make space available for the future jobs.

Like the CWFS, the HPC scratch space is typically not backed up due primarily to the transient nature of the data and the amount of data.

HPC scratch storage systems are normally procured with the HPC system. The HPC vendors propose the file systems and storage architecture as components within an overall HPC system. The HPCMP request for quotations (RFQ) for HPC systems states that the storage system must be architected to be resilient and robust; highly reliable components are to be utilized.

The HPCMP's RFQ defines the minimum aggregate data transfer rates between the compute nodes and the disk subsystem are based on specified ratio values for total system memory bandwidth (GB/s) to 1000 times the disk subsystem I/O bandwidth (GB/s). These ratios are 2.07, 1.68, and 1.34 for read, write, and full-duplex respectively. For example, a 200 node system with 50 GB/s of memory bandwidth per node would have total system memory bandwidth of 10000 GB/s. To meet the minimum full-duplex requirement, the system would require an I/O bandwidth of 7.46 GB/s (i.e. (200*50 GB/s)/(1000*1.34)). The minimum formatted usable disk storage size must be at least 40GB per processor core.

HPC vendors must also commit to monthly interrupt counts and overall systems availability (> 97%). This encourages the vendors to offer

reliable storage systems due to the penalties imposed for not meeting the system availability commitments.

The files systems end up on RAID protected storage that is either RAID 5 or RAID 6. RAID 6 is becoming more common place which is due to the increasing scratch space sizes which are architected with increasingly larger and lower costs SATA disk drives. These large drives take much longer to rebuild leaving the RAID set vulnerable to another drive failure. Vendors architect the storage with redundant paths from the hosts and multiple controllers. Metadata redundancy and availability is expected.

In order to maximize performance HPC scratch storage does not have end-to-end protection mechanisms or check summing.

Downtime for preventative maintenance may be executed at the recommendation of the HPC vendor. Typically, these downtimes are to update the HPC operating environment and not necessarily for the storage systems.

Support for the systems and storage is twenty four hours a day, seven days a week, four hour onsite support for hardware problems.

HPC Home and Applications

Home and application file system on the HPCMP HPC systems are considered more permanent then the HPC scratch file system. These file systems typically are hosted on the same storage as the HPC scratch space and utilize the same file system software.

With only minor exception, home and application file systems protected in the same manner as the HPC file system. Home and application file system are backed up on a daily basis.

Root File Systems

Root drives on major infrastructure servers and key elements in an overall HPC system (login nodes, admin nodes, etc) predominately are architected with multiple disk drives that are protected with RAID 1 or RAID 10. The costs for additional disk drives vs. performance make this a very worthwhile architecture decision.

Compute nodes within an HPC system that are architected with disk drives do not include redundancy. Since the disk drive images on compute nodes are easily reproducible, redundant drives in a RAID configuration are not employed.

Archive File Systems

Arguably, the tape archive systems are one of the HPCMP's most important systems which require the highest level of availability. Prior to integrating the CWFS, and the short time to live for data on HPC scratch the archive had to always be available to the user.

In addition, a user can have their HPC batch jobs in the work load management system queues in some cases up to fourteen days prior to the job running on the HPC system. Jobs that require input data from the archive system would not necessarily want to move their data immediately with the short time to live of the data in HPC scratch.

The DSRCs have architected redundant servers, redundant server component, redundant SAN switches, tape drives, very high speed RAID 5 or RAID 6 disk caches and metadata devices for the archive systems.

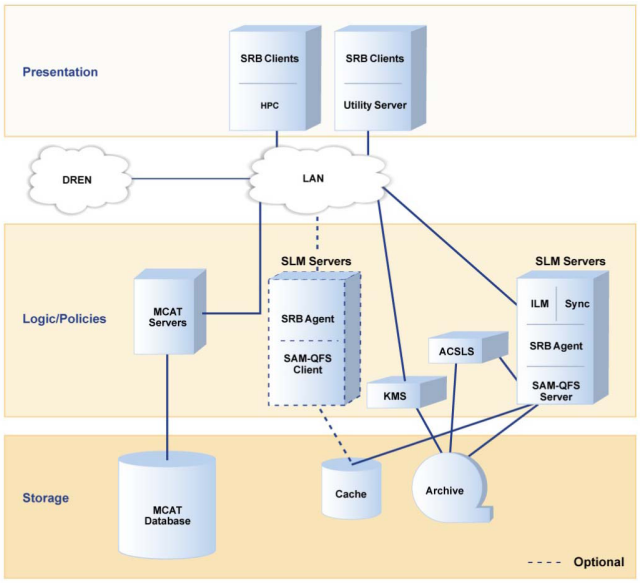
To maintain this high availability, the archive systems have been maintained at twenty four hours a day, seven days a week, with four hour response time.

Implementation of an active-active redundant archive server solution remains a priority.

Tape Archives

The HPCMP currently provides the user by default, two copies of their data on tape. One copy is at the DSRC and the other copy is sent via network to an archive system in another facility and then copied to tape. Archive file system metadata is backed-up daily and stored locally as well as remotely.

Storage Lifecycle Management



The HPCMP is currently implementing storage life cycle management (SLM). SLM tightly couples the current archive systems with an Integrated Lifecycle Management (ILM) management tool. The ILM is an Oracle Real Application Cluster (RAC) environment that will contain the file metadata from the archive as well as user applied metadata such as whether or not to make a disaster recovery copy, when the data can be deleted, what project(s) that data belongs to, etc.

The ILM is architected with multiple Oracle servers via Oracle RAC for redundancy and scalability. Additional reliability features incorporated in the design include redundant server components, a performance disk subsystem for the Oracle databases, utilizing multiple fiber channel paths per server, RAID 5 and RAID 10 volumes, redundant controllers, and redundant network interfaces connected to redundant switches.

CONCLUSIONS

In order to maximize HPC cycles for researchers, the HPCMP will continue to employ redundancy and other availability measures where practical to maintain availability of the systems, file systems and storage. The HPCMP would like to see the vendor community continue to develop the capabilities for end-to-end data protection (e.g T10DIFF) to ensure bit error rates are extremely low and that bit errors are identified and corrected. As storage space on disk drives continues to grow, new RAID schemes to decrease rebuild times and maintain file system performance are desired and would be implemented.