U.S. Department of Energy Best Practices Workshop on

File Systems & Archives

San Francisco, CA

September 26-27, 2011

HPC Enhanced User Environment (HEUE) Position Paper

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

John Gebhardt
Lockheed-Martin/U. S. Air Force Research Laboratory
DoD High performance Computing Modernization Program
John.gebhardt@lmco.com

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

ABSTRACT / SUMMARY

The DoD High Performance Computing Modernization Program (HPCMP) is now implementing a major change to at all its DoD Supercomputing Resource Centers (DSRC) through the introduction of a center-wide file system (CWFS) and an integrated life-cycle management hierarchical storage manager (ILM HSM).

Following discussions with its top consumers of archival capacity, the HPCMP architected a strategy to enable its customers to reduce archival requirements. The key elements of the HPCMP's strategy are:

- Provide tools to enable customers to associate project specific metadata with files in the archive; enable automated scheduled actions keyed against specific metadata; enable users to control second copy behavior; and enable user specified logical data constructs suitable for building case management features.
- Provide an intermediate level of storage between the HPCMP's traditional two tier scratch and archive architecture. This intermediate storage (i.e. CWFS) will enable customers sufficient time to analyze results, and archive analysis results rather than 3-dimensional restart files. The center-wide file system is sized to allow 30 days of analysis before transfer to archive.
- The introduction of the center-wide file system also creates the opportunity to enhance and upgrade interactive customer support with high performance graphics and large memory to support the analysis efforts.

INTRODUCTION

The HPCMP built forecasts of future archival storage capacity needs based upon past history and concluded that the current growth rate was unsustainable. Left unchecked, storage would consume the majority of the HPCMP's budget by the end of the decade. A key finding of the subsequent analysis was that the archive costs remained in an affordable range if the archive

growth rate was constrained to 1.4 times the growth of the previous year's growth. This finding is tied to an assumption that industry doubles tape capacity every 24 months. If the growth rate of the archive exceeds the rate of tape capacity increase, the HPCMP has to fund tape libraries, slots, and potentially licensing for the additional capacity.

After gathering input from the principal investigators of the projects that consumed the vast majority of the program's archival capacity, several recurring themes emerged:

- Existing storage tools were insufficient to manage large datasets and the use of filenames to capture relevant metadata was no longer practical.
- Raw computational outputs were being archived due to insufficient analysis time for data stored in scratch space.
- Performing analysis using batch resources was adding to the problem of insufficient time for analysis.

These observations were further vetted and ultimately formed into requirements for the HPCMP's next generation storage solution. A working group, with representatives from the HPCMO, the DSRCs, and user advisory groups, was formed. The group was chartered to further develop and refine the requirements and to develop the architecture for data flow within the HPCMP.

The architecture that the storage working group arrived at included a combined information lifecycle management and hierarchical storage management layer.

A subsequent effort surveyed the information lifecycle management and high performance storage markets, leading to the creation of an acquisition strategy. A key element of this strategy was the separation of hardware and software requirements and provisioning.

A market survey determined, to no great surprise, that a mature information lifecycle management solution integrated with hierarchical storage management did not exist. The strategy that emerged was to seek a partnership with industry aimed at fostering the integration of a leading information life cycle management solution with a leading hierarchical storage manager. Our software requirement allowed for an initial capability that could evolve into a fully integrated solution over 10 years.

The combined ILM HSM requirement was called "HPCMP Storage Lifecycle Management." A Request for Proposals was released in March of 2009 and a contract awarded in August of 2009.

With the ILM+HSM addressing the software requirements for improved tools to manage data, the remaining primarily hardware requirements were for the Center Wide File System and the Utility Server. This second component of the acquisition strategy was focused on the required hardware to deliver the new services.

Much of the requirements for the Center Wide File System (CWSF) and utility server derived from the winning ILM+HSM solution. Subsequently, two Requests for Proposals were issued. The RFPs required responses for the six DSRC locations and for a range of file system capacities and performance levels. They also required the inclusion of a 10 gigabit network fabric for connection of the Center Wide File System components, the utility server nodes, and the HPC system login nodes at each DSRC. The storage capacities requirements ranged from 250 TB to 2 PB. The I/O performance requirements ranged from 8.0 to 40.2 GB/s and from 70,000 to 320,000 file open/creates per second.

Awards for the CWFS and utility server were made by Lockheed-Martin in September 2010 and deliveries were completed in December, followed by acceptance and integration. The systems were transitioned into production sequentially center by center between June and August 2011.

In 2011, the HPCMP also took steps to refresh its tape archive hardware. Based on the earlier analysis showing that a doubling of tape capacity every 24 months was a key component of cost containment, the program compared commodity LTO drives with the proprietary Oracle T10000

line. Although the LTO family drives have a lower initial purchase price than the T10000 family drives, the lifecycle costs for LTO were found to be significantly higher. Drivers were the need to replace the media with each generation of LTO drive in order to realize the increase in capacity and the slightly slower capacity growth rate (15x over ten years for LTO and 10x over five years for T10000).

CONCLUSIONS

It is too early to gage the impact on the growth of the HPCMP's archive as a result of the acquisition and deployment of the HPCMP Enhanced User Environment. The Program's advisory bodies have responded positively to the goals and progress. The initial feedback for the additive analysis capability provided by the utility server and Center Wide File System has been positive. Once the Storage Lifecycle Management solution is fully deployed for production use in October 2011, the full effects of HEUE will be measured and reported.

In terms of the adoption of commodity hardware, the tape industry would need to seek ways to reduce the frequency of complete media replacements while meeting or exceeding the 1.4 compound annual capacity increase target.