LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-CONF-497278

Best Practices Workshop Position Paper - Reliability

M. R. Gary

September 1, 2011

The 5th DOE Workshop on HPC Best Practices: File Systems and Archives
San Francisco, CA, United States
September 26, 2011 through September 27, 2011

Disclaimer

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

Mark Gary
Lawrence Livermore National Laboratory
mgary@llnl.gov

ABSTRACT / SUMMARY

This position paper addresses the reliability and availability of storage systems. Specifically it introduces LLNL storage best practices in the areas of resilient architectures and daily operations which contribute to enhanced availability, reliability and computational integrity in LLNL's 24x7 HPC centers.

INTRODUCTION

Livermore Computing operates multiple 24x7 "lights on" HPC environments and has done so for over 40 years. Availability, reliability and computational integrity are of paramount concern in the Center due to the tremendous investment in, and the importance of, our HPC machines and the data they generate. This position paper outlines, at a very high level, some of the storage system best practices followed by LLNL. The two focus areas covered in this paper are:

- **Resilient Storage Architectures**: Best practices surrounding the storage hardware architectures employed in the LC and their impact on availability and data integrity.
- **Daily Operations**: Storage system best practices surrounding daily operations (from outages and maintenance to training and communications).

Within these areas I very briefly identify best practices as fodder for Workshop discussion.

Resilient Storage Architectures

Allowing storage operations to continue in the face of failure or outage is critical. Among the hardware architecture best practices followed in the LC are:

- ***Scalable Unit Architecture***

Following the lead of our computing platforms we deploy storage hardware using the concept of a Scalable Unit (SU). An SU is the smallest unit of hardware (storage and associated servers) by which you can grow a storage subsystem. Well identified *identical* SUs allow for ease of repair, maintenance, administration, expansion, and sparing. The purchasing power of buying identical hardware in volume is an added benefit.

- ***Leveraging Compute Platform Hardware***

In our file system and archive environments we, whenever possible, leverage and duplicate the server hardware technologies used on our compute platforms. As in the SU area, this helps ease repair, maintenance, administration, expansion and sparing and takes full advantage of the purchasing power and technology investigation efforts made during platform procurements.

- ***Failover Partners***

The LC has nine very large Lustre file systems. The Object Storage Server (OSS) nodes controlling subsets of disk are architected into failover pairs allowing a failed OSS to have its disk taken over by its healthy partner. This

architecture is leveraged constantly and aids not only the case of node failure, but also increases availability during software and firmware deployments and electrical work.

- ***Targeted Use of Uninterruptible Power Supplies (UPS)***

The amount of electrical power required by LC compute and infrastructure hardware makes full coverage with backup power/UPS power economically infeasible. Instead the LC uses its UPS budget in a targeted manner with a concentration on support of metadata services, network infrastructure, and home directory file systems. This strategy focuses on the protection of the most critical data and aides in a rapid return to service of Center operations following a power outage.

- ***Dual Power Sources***

The majority of LC storage infrastructure racks are wired in a redundant manner to be able to survive the planned or unplanned loss of any one electrical subpanel. This is particularly critical in our very dynamic machine room environment during this era of enhanced electrical safety rules.

- ***Archive Dual Copy***

While budgets do not allow us to keep multiple copies of the PetaBytes of simulation data stored in the LC, our HPSS systems do allow a user to direct that dual copies be made of targeted files. The two copies are stored in geographically separated locations cross-Laboratory.

- ***Degraded Mode Archive Operation***

Because of the distributed, multi-level hierarchy architecture of our HPSS archive implementation we are commonly able to provide users with various levels of degraded mode service during outages. In degraded mode not every file is accessible, but typically the most recently written files and those with dual copies can be accessed.

Daily Operations

On a daily basis the LC implements a number of operational best practices surrounding our storage systems including:

- ***"Lights On" Operation***

It is our philosophy, in part driven by the tremendous dollar investment in our computing environment, that the LC provide 24x7x365 customer service and compute availability. Our cross-trained operations staff is always on site monitoring our systems and answering off-hours user questions - around the clock. In the event of an environmental emergency (e.g., loss of cooling) they can immediately react following prescribed/ordered power down procedures. Operations staff are trained in basic file system and archive administration and do hardware repair as well. They have full access to on-call storage system and archive administrators at any hour.

- ***Self-maintenance***

Much of our hardware maintenance is performed by LC employees. This allows us to perform maintenance immediately when a problem occurs, eliminates security escort requirements, and allows us to closely track and learn from system failures. The fact that storage hardware leverages platform hardware procurements means that our personnel need be trained on only a limited number of equipment types.

- ***Hardware/Spare Burn-in***

We have a full hardware spare/RMA center supporting our local maintenance operations. Rather than pulling spare parts off of the shelf, we maintain a burn in environment where we have spare storage hardware under continuous test, exercise, and burn in. When equipment fails, hardware is pulled directly from the burn in environment. Before tape drives are allowed to be placed into service they first undergo a suite of performance tests and integrity tests.

- ***Testbeds***

Livermore Computing has a variety of testbeds in which we test pre-production hardware and

software. These testbeds range from single racks, to the well-known multi-vendor Hyperion test environment where software can be tested at scale with thousands of clients.

- ***Data Integrity Checking***

Continuously in the background, the LC runs a tool called DIVT - the Data Integrity Verification Tool. DIVT checks the data integrity of our archive and our parallel file systems by writing known data patterns to files from different platforms, forcing the data to flow through file systems and down to archival tape, and then checking data integrity upon fetch back to the platform. Over the years DIVT has caught data corruption ranging from on-platform component problems to corrupting drive firmware.

- ***Planned Downtimes***

The LC has a philosophy that planned downtimes happen during the work week from Tuesday through Thursday unless particular circumstances dictate otherwise. While this has an impact on interactive users, Center resources are fully subscribed 24x7 including weekends. Our philosophy allows us to have experts from all disciplines on hand in case of problem in order to improve availability. Fridays are avoided to limit the introduction of problems impacting the weekend. Mondays are avoided to allow users to process the results of their weekend runs.

Software rollouts are planned in such a manner as to minimize impact on the programs supported by the Center. We rollout software to our unclassified systems first which allows us to bring outside experts to bear on problems encountered. Recently we leveraged a Six Sigma quality project to improve our software rollout process and reduce the length of planned downtimes.

- ***Impacts and File System Meetings***

Every Monday representatives from every facet of the LC (including Facilities) have a formal meeting to manage any outage or operation that has impact on Center customers or has cross-cutting impact among Center discipline areas. This meeting has tremendous value and allows us to combine outages and plan forward in order to maximize the availability of all center resources including storage. A separate meeting which pulls together Operations staff, storage system administrators, and hardware repair personnel occurs weekly. This meeting improves file system specific communication across all involved disciplines and shifts.

CONCLUSIONS

Large HPC environments are extremely complex. They require that particular attention be paid to operational and architectural storage system best practices in order to ensure availability, reliability and computational data integrity.

* This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-CONF-497278