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

Wayne Hurlbert
Lawrence Berkeley National Laboratory
wehurlbert@lbl.gov

ABSTRACT / SUMMARY

This position paper aims to provide information about techniques used by the Mass Storage Group at the National Energy Research Scientific Computing Center (NERSC) to accomplish technology refresh, system configuration changes, and system maintenance while minimizing impact on users and maximizing system availability and reliability. In particular, it addresses the Center's position that shorter, scheduled outages for archival storage system changes, occurring at familiar times, minimizes the likelihood of unscheduled or extended outages, and so minimizes impact on users.

INTRODUCTION

For the purposes of this discussion, it is taken as given that much of the activity involved in technology refresh, along with configuration changes and other system maintenance, requires systems to be off-line. NERSC's approach to these activities in the archival storage systems, is largely driven by the need to minimize the impact on our users. The notion of minimum impact encompasses the scheduled activity itself, potential fallout from the activity, and the concept of preventive maintenance. This has resulted in a conservative attitude toward system maintenance that favors incremental rather than radical change. In the following I will discuss the motivation and benefits of this approach, and mention some of the real world steps taken at NERSC to implement the approach.

Little by Little

While it can be tempting to "just do it", an incremental approach to technology refresh and other system maintenance activities is usually a viable alternative to the more significant outages often required to accomplish the changes in a single sitting.

Types of projects for which this approach might be helpful include:

- Server upgrades or replacement.
- Significant application, OS, or layered software upgrades.
- Replacement or reconfiguration of disk and tape resources.
- Replacement or reconfiguration of large infrastructure components such as SAN switches.

These projects will often take many hours, sometimes days, to accomplish and run a relatively high risk of unanticipated problems or complications.

An incremental approach indicates that these larger projects be broken up into smaller pieces which can be accomplished in an independent and sequential manner. Naturally, there are projects where this is not possible, for various reasons; our finding is that the reasons are typically not technical in nature.

The Benefits

There are several benefits provided by this approach:

- Less complexity of the tasks executed during an outage, which means a reduction in the likelihood of human mistakes in planning or execution of the tasks.
- Lower risk of aborted or extended outages due to unexpected or unanticipated complications. For example, because fewer tasks are being undertaken, there is a smaller window for hardware failure if devices or servers are being power cycled. Naturally, a device can fail during either an incremental activity or a major project, but the impact on workflow is likely to be smaller, and the impact on the user is likely to be less significant in terms of total time for the outage.
- Easier back out in the case of the need to abort the maintenance activity due to unexpected or unanticipated events.
- Lower likelihood of human error due to the fatigue and stress which usually occur during significant projects.
- When compared with forklift upgrades, lower risk of subsequent fallout due to as yet undiscovered bugs or defects. This is particularly true, obviously, for newer products.
- Where desirable, allows for completing system-down activities during business hours, because of the shorter outages. Business hours may be required in order to insure access to outside expertise.

User Expectations

The incremental approach to performing system maintenance subscribes to the notion that shorter, more frequently scheduled outages will ensure a more stable system, which will better serve users. Outages should be scheduled for a standard day and time, even if not at standard intervals e.g. weekly, with the intent that users will come to expect that time period and plan around it. For instance, on one end of the spectrum, users can simply plan to not run during the normal hours, on the normal day for outages. However, NERSC does provide a programmatic, network based mechanism for automated jobs to check system availability.

Further, NERSC has developed an effective protocol for suspending user storage transfers during short outages. Referred to as “sleepers”, user interface tools on the compute machines look for lock files which cause these clients to loop on the system sleep call until the lock file disappears. The result is that many user jobs simply pause until the outage is completed.

In annual user satisfaction surveys at NERSC, the archival storage resources typically receive high scores with regard to system availability and reliability. [1] [2]

Preventive Maintenance

Preventive maintenance, in the sense of avoiding unscheduled outages and the associated user interruptions, can be seen as primarily concerned with restarting, rebooting, and/or power cycling equipment. These activities usually take relatively little time, and fit nicely with shorter, more frequently scheduled outages. Examples include:

- Reboot to validate configuration changes made while the system is live, even if a reboot/power cycle is not strictly required.
- Reboot to flush out pending hardware failures, or to reset hardware that is in a confused state.
- Rebooting or power cycling also helps maintain familiarity with the way systems and devices behave during power-down and power-up.
- Restarting applications, and less importantly these days restarting operating systems, can

help avoid outages due to software defects such as memory leaks.

- Build rather than copy: when locally built software must be installed on multiple servers, building it on each server validates the installation and configuration of layered software (in addition to allowing debug activities on the various servers).

Example

Project: application upgrade on the current production server hardware, which requires OS and/or layered software upgrades.

The NERSC storage group will typically build a new system disk, from the ground up, on a second disk in the production server.

This will usually involve an outage to install the new OS followed by one to several 2-3 hour outages to install, build, configure, and test (as appropriate) layered software and application code. Each of these outages will involve a reboot to the second system disk for the work to be done, followed by a reboot back to the production disk.

This activity is usually spread out over a number of weeks, and is typically interleaved with other activities that may, or may not, involve preparation for the upgrade.

The upgrade is finalized by rebooting to the new system disk and performing any remaining activities required before going live.

CONCLUSIONS

A conservative approach to system outages for technology refresh, system reconfigurations, and other maintenance can be accomplished through a policy which uses multiple short outages to perform the work incrementally. This promotes greater system stability and minimizes the number of unscheduled outages, resulting in better service to users.

REFERENCES

1. NERSC 2010 High Performance Computing Facility Operational Assessment.
2. NERSC User Surveys. http://www.nersc.gov/news-publications/publications-reports/user-surveys.