# Challenges in Managing Multi-System Multi-Platform Parallel File Systems

## U.S. Department of Energy Best Practices Workshop on File Systems & Archives

## San Francisco, CA

## September 26-27, 2011

**Ryan Braby**
National Institute for Computational Sciences
rbraby@utk.edu

**Rick Mohr**
National Institute for Computational Sciences
rmohr@utk.edu

## ABSTRACT / SUMMARY

The National Institute for Computational Sciences (NICS) is looking to deploy one or more center wide parallel file systems. Doing so should reduce time to solution for many NSF researchers. Researchers who run on multiple systems at NICS will no longer need to move data between parallel file systems and hopefully this will reduce the amount of file system space used for extraneous data replication. However, there are a number of challenges in setting up a multi-system, multi-platform parallel file system. This paper discusses many of the identified challenges for deploying such a file system at NICS and supporting at least the following architectures; Cray XT5, SGI UV, and commodity Linux clusters.

## INTRODUCTION

The National Institute for Computational Sciences (NICS), a partnership between the University of Tennessee and Oak Ridge National Laboratory, was granted a $65M award from the NSF in September 2007. A series of Cray HPC systems, named Kraken, were purchased and deployed. Currently, Kraken is a Cray XT5 system with a peak performance of 1.17 PFlops. Lustre is the primary file system for Kraken, and it is built on top of DDN storage directly attached to special I/O service blades in the Cray. These blades act as the MDS and OSS servers for the rest of the system.

In the last year, NICS has deployed a new file system to be shared between Nautilus (a large SGI UV) and Keeneland (a cluster used for GPGPU development). An evaluation of file system technologies was done, and Lustre was selected for use here. Ideally, the scratch file systems will be shared across all NICS HPC resources. To this end, we have been planning and preparing to upgrade our Infiniband SAN, attach Kraken to this SAN, and migrate Kraken's current Lustre file system to be SAN attached.

While a number of sites have deployed multi-cluster Lustre file systems, unique site requirements prevent the creation of a one-size-fits-all solution. NICS supports a wide variety of platforms (Cray XT, SGI UV, and Linux clusters). Individually, these platforms can present challenges for a site-wide Lustre file system. Combining them further complicates matters.

### CRAY XT

Cray ships Lustre as part of CLE (Cray Linux Environment), but they are currently shipping an older version (1.6.5) with custom patches. While it is nice to have a vendor supported version, this version is older and lacks features that have been introduced in newer versions. As we move to a center wide file system, there are also concerns about version compatibility between the servers and all the clients.

While it should be possible to put a newer version of Lustre into CLE boot images, there are a number of possible complications with doing so. At this time NICS does not have a file system developer and it is not in our short term plans to

hire one. We could build and install Lustre, but we have minimal resources to test it on. Lacking a file system developer our abilities to fix issues with Lustre in CLE would be limited. The Cray XT systems use a proprietary SeaStar network, which requires it's own Lustre Network Driver (LND) and could complicate LNET routing. Further, Cray support might be hesitant to help on production issues when we are running our own version.

### SGI UV

The SGI UV, is a large NUMA architecture with a single system image. Running a single Linux kernel, this architecture tends to get poor IO bandwidth when compared to clustered systems of similar core count. NICS has spent time testing multiple file systems on our 1024 core UV system, and determined that in present day performance Lustre (1.8.6) was the winner (just barely).

Comparing the known road maps for the major parallel file systems, Lustre was the only one that has plans for improving SMP scalability and NUMA performance. In particular, it looks like some improvements in this area have already been added in Lustre 2.1.

Another challenge for parallel file systems on the SGI UV is effectively utilizing multiple network interfaces to our SAN. As a large single system image system, it is important for performance that a file system can drive multiple network interfaces at near line rate. We have had some success scaling Lustre read performance with multi rail infiniband on our UV. This is an area that we hope to see improvements to Lustre for in the future.

## LINUX CLUSTERS

Linux clusters with Infiniband interconnects are probably the most common platform for Lustre file systems. As such, including Linux clusters in a multi-cluster Lustre configuration adds some to the complexity. It is another platform to consider and keep track of, but it is also one that you can rely on the community for testing and development.

## MULTI-SYSTEM CHALLENGES

Deploying a Lustre file system that spans multiple systems and architectures introduces new challenges apart from the previously mentioned system-specific ones. For example, it may be desirable to run Lustre 2.1 on the SGI UV in order to address some of the SMP scalability issues. However, this would require running Lustre 2.1 on the MDS and OSS servers, which is not compatible with the Lustre 1.6 client on the Cray.

Maintaining compatibility between all of the clients and servers is the first major challenge to a multi-system Lustre deployment. Different platforms may require different patches, and in some cases require different client versions. Knowing which versions are compatible and testing the compatibility is critical to ensuring file system usability.

Some system vendors include a supported version of the Lustre client and publish supported client / server combinations. Merging these requirements from multiple vendors could lead to a situation where the supported versions are not compatible with each other. To reconcile this may require running a version not supported by one or more vendors. One approach to deal with this would be to purchase third party Lustre support.

Managing a multi-system parallel file system makes the file system more of an infrastructure service. Since multiple rely on the availability of the file system, the effects of any disruptions (like maintenance) must be carefully considered. Further, you have to plan upgrades carefully; ensuring that at all points in your upgrade plan you are on compatible versions and not unintentionally running an unsupported combination of server, router, and client versions.

Also, like any infrastructure service, there are possible contention issues. Performance on one system can and will be impacted by access from another system.

## CONCLUSIONS

NICS is planning to move to center wide Lustre file systems. There are a number of issues

involved in doing this. While we have identified many of the issues and have ideas of how to deal with them, we do not have the experience and history of implementing these ideas to determine if they are indeed best practices.

## REFERENCES


1. T. Baer, V. Hazlewood, J. Heo, R. Mohr, J. Walsh, *Large Lustre File System Experiences at NICS.* CUG 2009, http://www.cug.org/5-publications/proceedings_attendee_lists/CUG09CD/S09_Proceedings/pages/authors/11-15Wednesday/12B-Walsh/walsh-paper.pdf
2. G. Shipman, D. Dillow, S. Oral, F. Wang, *The Spider Center Wide File System; From Concept to Reality.* CUG 2010, http://www.nccs.gov/wp-content/uploads/2010/01/shipman_paper.pdf
3. Whamcloud JIRA for Lustre SMP salability: http://jira.whamcloud.com/browse/LU-56