A Center-Wide File System using Lustre

Shane Canon
H. Sarp Oral
Oak Ridge National Laboratory

May 8, 2006

Abstract

The National Leadership Computing Facility is currently deploying a Lustre based center-wide file system. This file system will span multiple architectures and must meet demanding user requirements. The objectives for this file system will be presented along with an overview of the architecture. A discussion of issues that have been encountered during the deployment, as well as current performance numbers, will also be provided. The paper will conclude with future plans and goals. KEYWORDS: Lustre, file system

1 Introduction

The National Leadership Computing Facility (NLCF) is a computer user facility funded by the Office of Science (SC) within the U.S. Department of Energy (DOE). Its mission is to provide leadership-class computing resources to a select set of projects in order to accomplish breakthrough science. To deliver this, the NLCF operates an 18 TF Cray X1E and a 25 TF Cray XT3. Furthermore, the DOE-SC has announced aggressive goals for delivering new capabilities at the NLCF in the coming years. These targets include 250 TF of capability in 2007 and 1 PF of peak computing capability in 2008. To deliver on these targets, the NLCF has partnered with Cray and has developed a strategy based on the continuation of the XT3 road map.

The NLCF systems are allocated by the SC for use by roughly 20 projects. These projects span a variety of scientific disciplines including climate, astrophysics, chemistry, materials science, and fusion. Most of these projects receive millions of node hours on the X1E or the XT3. Unique from other DOE-SC computing facilities, the NLCF can schedule **dedicated access** to a system to insure coordination with other constraints such as the availability of other resources, researchers, or deadlines. In addition to the large computational platforms, NLCF also operates resources geared towards data analysis and visualization. These include a 256 processor SGI Altix and a 64 node AMD Opteron based visualization cluster. NLCF is also exploring a new mode of data discovery, termed end-to-end, which couples large scale computations directly with data reduction, distribution, and simulation monitoring. An overview of the resources at NLCF is presented in Table 1.

The NLCF is in the process of deploying a center-wide file system called Spider. The file system will be based on Lustre [2] and will span the major production resources. The remainder of this paper will describe the motivation for this file system, the initial strategy for deploying the file system, the experience to date, and future plans.

Resource Name	Architecture	Use
Phoenix	Cray X1E	Computing
Jaguar	Cray XT3	Computing
Ram	SGI Altix	Data Analysis
Hawk	Cluster (Quadrics)	Visualization
Ewok	Cluster (IB)	End-to-End
Spider	Cluster	Center-Wide File System

Table 1: NLCF Resources

2 Motivation and Challenges for a Center-Wide File System

2.1 Data Management

One of the issues most frequently raised by NLCF users in doing large scale simulations relates to data management. The users often need to move massive amounts (e.g. tens of TBs) of data from remote locations. Following large simulation runs, the users need to do significant post-processing and data analysis to reduce the simulation data, create visual representations, or perform data discovery in order to better understand the results. Unfortunately, these tasks are typically carried out on different resources than the initial computation and require tedious effort to move and manage the data. Furthermore, this results in unnecessary duplication of data, as well as the need to insure data sets are consistent between storage systems. Having a centralized file system addresses many of these challenges.

2.2 Leveraging I/O Capability

In addition to the inefficiencies that separate storage resources impose on the user, multiple storage systems also result in inefficiencies from the hardware and administrative side. I/O bandwidth is one of the more costly components in high-performance computing. Providing high bandwidth storage systems on the large computing systems and then duplicating this capability on data analysis and post-processing systems results in added cost. For performance targets of a few gigabytes per second, this duplication may be affordable. However, as NLCF looks towards the I/O needs on a 1 PF system, it is evident that this system will require hundreds of gigabytes per second of I/O bandwidth, if not terabytes per second, to provide a balanced system. Delivering this capability represents a significant investment, and replicating this capability on other systems is impractical. So, as we look towards the future systems, it is imperative that the I/O system extend beyond the computational platform.

2.3 Challenges

While there are clear motivations for deploying a center-wide file system, this approach also entails numerous challenges. These challenges are both technical and policy related. Technical issues include the need for a multi-platform file system that supports various interconnects and allows routing between the networks. Furthermore, the system must be highly scalable and cost-effective to satisfy the current and future bandwidth requirements of the center.

From a policy vantage point, we must insure that the quality of service provided to the various systems is matched with the requirements and priorities of the system. For example, the large computing system represents the most significant investment and we must insure that it receives the bandwidth it requires to sustain application performance. Also, a centralized file system impacts the security stance of the center. Careful consideration must be given to the protection methods provided by the file system compared with the security requirements of the center.

3 Architecture for the Center-Wide File System

In 2005, the NLCF developed a strategy for delivering a center-wide file system built around Lustre. Much of this strategy has already been executed. For this plan, a 10 Gb Ethernet network was deployed to provide the needed bandwidth for the storage system. The backbone of this network is provided by a Force10 E1200 switch. Key systems will possess multiple 10 Gb links to achieve the required bandwidth. The meta-data server (MDS) and Object Storage Servers (OSSs) are be connected via 10 Gb as well. For the initial deployment, a target of 10 GB/s of aggregate bandwidth was set.

Since the initial plan was developed, we have made minor adjustments. Testing has started with a smaller configuration which can provide roughly 2 GB/s of aggregate bandwidth. Once we have functionally demonstrated the feasibility of the central file system on the key systems, the file system will be phased into production and scaled up to meet the 10 GB/s target.

4 Experience to Date

4.1 Deployment on the Spider Cluster

Currently, the Spider cluster consists of twenty OSSs and one MDS. The OSS systems are configured with dual dual-core AMD Opteron processors, 8 GB of RAM, a dual port 2Gb Fibre Channel card (QLogic QLA2342-CK) and a 10Gb Ethernet card (S2IO). The MDS is configured with dual dual-core AMD Opteron processors, 8 GB of RAM, and a 10 Gb Ethernet Card (S2IO). For the storage hardware, two DDN 8500 couplets with Fibre Channel disk are used. The DDN storage system provides roughly 2.4 GB/s of aggregate bandwidth and approximately 17 TB of formated Lustre space. The initial deployment of Lustre on this system has been smooth, and no serious issues have been encountered. The tests have used evolving versions of Lustre 1.4. Early single node tests demonstrated a single client connected via a single 10 Gb link could achieve nearly 300 MB/s.

4.2 Deployment on Cluster Systems

For the further testing of the Spider system, NLCF mounted Spider on the Hawk visualization cluster. Since Hawk is a fairly standard Linux cluster, this configuration presented the easiest target for testing. As expected, few problems were encountered. This configuration was used to do early performance and scaling test. While the Hawk cluster consists of 64 nodes, twenty of the nodes are dedicated to running a PowerWall system and other services. Consequently, only 44 nodes are available for testing. However, since the scale of this system and its network are well matched to the capability of the initial storage hardware, this platform is well suited to exploring the potential performance that could be obtained from the Spider system. Many of the studies that are described below used Hawk as the client system.

4.3 Deployment on SGI Altix

The NLCF SGI Altix system, Ram, is designated for data analysis. While it possesses a local XFS file system capable of delivering over 1 GB/s of file system bandwidth, it is desirable to have Ram coupled with the Spider file system. However, the architecture of the Altix, a very large SMP, presents several challenges.

One challenge is providing sufficient network bandwidth to the system. While cluster systems typically achieve high aggregate network bandwidth by leveraging individual links, on a large SMP system like the Altix, this isn't possible. Instead, many interfaces must be bonded together, or the file system must be capable of stripping I/O across several interfaces. To date, we have only tested the bandwidth of a single 10 Gb interface in a small (32 processor) Altix development system. Since the Altix has PCI-X interfaces, the most that can be obtained is roughly 6.5 Gb/s.

Another challenge on the Altix architecture is obtaining high performance with the Lustre client. Much of the Lustre development and testing to date has been focused on clusters. Consequently, Lustre has not been tuned to perform well on large SMPs, especially for the client. To date, we have only obtained around 300 MB/s of aggregate sustained performance.

It is still unclear how much effort will be required to achieve the target performance of multiple gigabytes per second. Also, the NLCF is still reviewing future plans for providing resources for data analysis. We intend to continue to test Lustre on the Altix and find ways to improve performance, but will defer a decision on any significant investment.

4.4 Deployment on Cray XT3

One of the most critical aspects of deploying the center-wide file system was providing access to the system on the Cray XT3. Since the XT3 is currently the most powerful resource at the center and the architecture represents the future direction for the center, it was imperative that this system have high bandwidth access to the file system. However, this presented some challenges. Typically, systems rely on the natural routing capabilities of TCP/IP to move data from compute nodes to and from nodes outside of the system. However, the light-weight kernel used on the XT3, Catamount, only natively supports the Cray Portals protocol. Furthermore, the native Cray Portals implementation is slightly different compared to that used in Lustre. To address this issue, NLCF contracted CFS, Inc., to develop a routing capability in Lustre. This ultimately resulted in a new networking layer in Lustre called LNET which is now the standard networking layer used by Lustre (starting with 1.4.6)[1]. LNET uses a Lustre Networking Device (LND) layer to implement different network types. An LND exists for Portals, which is the native network protocol for the Cray SeaStar network. LNDs also exist for TCP/IP, various InfiniBand stacks, and other interconnects, such as Quadrics and Myrinet. Most importantly, LNET supports routing between these networks in the Lustre Networking layer.

At the time of this writing, NLCF is in the process of testing this routing code on the XT3. Version 1.4 of Cray UNICOS/LC will include Lustre 1.4.6 which has the routing code. A development XT3 system is upgraded to this release and will allow NLCF to complete testing of the routing code. This will provide the last critical functional component of the Spider project.

4.5 Study of Efficiency of Multiple Object Storage Targets per OSS

As NLCF began testing of Spider, one open question related to the optimum number of Object Storage Targets (OSTs) per OSS. To address the question, we did a study looking at the performance for configurations with one, two and four OSTs per OSS. In each case we kept the total number of OSTs fixed at sixteen. Some of the results from this study can be seen in Fig. 1. Currently, the results are mixed. In general, lower OSTs per OSS perfrom better on reads. However, the 8 OSS configuration (2 OSTs per OSS) provided the best write performance.

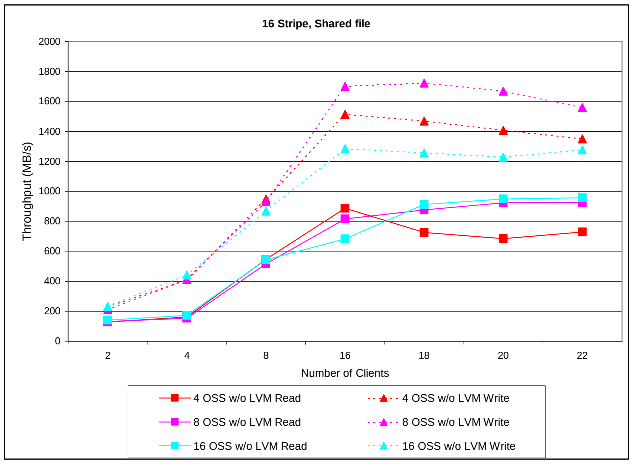

Figure 1: Comparisons of Performance with various OSS/OST configurations.

4.6 Impact of Linux LVM on Lustre Performance

Initially, we configured the OSTs to sit on top of a logical volume using the Linux Logical Volume Manager (LVM). LVM was configured so that a future study of backup approaches could be explored. The logical volumes were configured with a single physical volume that corresponded to a single LUN from the DDN. As we carried out some of the studies described in the previous section, we encountered some problems with device naming. To avoid these issues, we removed the LVM and immediately noticed an impact on the performance. This motivated us to study the performance impact of LVM in more detail. The results can be seen in Fig. 2. The impact of using LVM was observed to be inconsistent. In discussion with people familiar with LVM, we were informed that LVM should have little impact on performance unless snap shots are being used (which were not in our tests). We have since learned that a mismatch in the block layout and the DDN cache system may be blame. We plan to do further tests to resolve this question.

5 Future Plans

5.1 Lustre Road Map

CFS, Inc., maintains a published road map for the Lustre file system [3]. Many of these plans dovetail nicely with our plans for the centralized file system and the road map for the center. The road map includes scaling targets, clustered meta data servers, auto-balancing and improved management that are well aligned with NLCF objectives. Furthermore, an upper layer RAID mechanism, Lustre RAID, will provide added reliability by extending RAID across the OSTs. This feature should further improve reliability and availability.

5.2 InfiniBand Investigation

The NLCF is currently examining several interesting possibilities related to InfiniBand (IB). IB is interesting both as a storage fabric and as a next generation enterprise network. To explore these aspects, NLCF is evaluating a DDN 9500 with InfiniBand interfaces and has deployed a small IB network.

Clients of the DDN 9500 can use Storage Resource Protocol (SRP) to access storage targets over the IB network. Using IB and SRP could allow us to build a cost effective SAN that could include dozens of servers. These host interfaces are much more affordable than 4 Gb Fibre Channel and provide more bandwidth. However, these implementations are still relatively new and may require some time to mature.

In order to achieve the aggressive I/O bandwidth targets for the center, it will be critical that we have an affordable and scalable fabric to span the various systems. While this is technically feasible with 10Gb Ethernet, the cost to provide 100s of GB/s of bandwidth between systems is expensive. Building a similar infrastructure with InfiniBand is much more affordable. However, there are still open questions to the feasibility and reliability of this approach. The IB testbed is intended to help address some of these questions. For this testbed, PCI-E based Host Channel Adapters have been installed in the Spider nodes. These nodes are connected to a small 24 port IB switch. This switch is connected via fiber converters to the IB Switch for the 80 node Ewok cluster (see Table **??**). We will investigate Lustre running in various styles over the IB network, as well as scaling

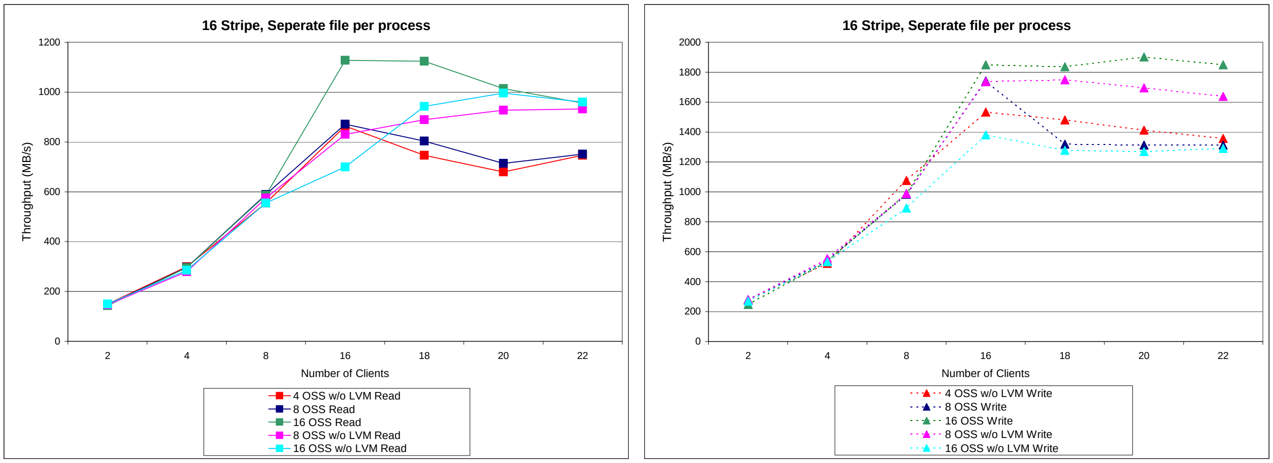

Figure 2: Comparisons of Performance with and without LVM for various OSS/OST configurations.

tests with Ewok. Also, partitioning and IB routing will be explored. These tests should help us to understand the readiness of IB and its place in the near future of the center.

5.3 Commodity Storage

Another area of exploration is the use of commodity storage with a Lustre file system. This would primarily be of interest as a tertiary storage area for large data sets and would augment the archival storage system. To investigate this, we are currently using storage nodes associated with Ewok, the end-to-end cluster, that are based on commodity hard drives and IDE based RAID controllers. From a cost comparison, a gigabyte per second for high-end storage typically runs around $100k, while this could potentially cost around $30k. However, it is difficult to achieve the same levels of reliability and availability with this storage versus premium storage. So, in addition to studying the performance characteristics, we will also be looking at reliability.

Another project initiated at NLCF is to develop a mechanism for the XT3 to offload I/O functions from Catamount compute nodes to the Linux SIO nodes. This approach could be useful for overcoming potential scaling issues with Lustre or supporting alternate file systems. The mechanism uses a library that is compiled into the application in a manner similar to liblustre, as well as a daemon that runs on service nodes. The library intercepts I/O functions and encapsulates the function into a portals message that is sent to the daemons. This is similar to how “yod” currently offloads I/O operations for tasks [4]. However, whereas “yod” has one daemon process per parallel job, this system will scale to multiple daemon processes for a parallel job. This should allow it to sustain higher I/O bandwidths.

6 Conclusion

Data management consistently ranks as one of the most tedious and time consuming aspects of large scale simulations. One approach to alleviating some of these issues is a high performance common file system that spans key resources in the center. NLCF has embarked on a project to deploy such a system based on the Lustre file system. To date this system has been tested on several resources.

Acknowledgments

The authors would like to thank the staff and colleagues who have contributed material to this paper. Also, we wish to acknowledge the work of Cluster File Systems, Inc., for the routing code. Research sponsored by the Mathematical, Information, and Computational Sciences Division, Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.

About the Authors

Shane Canon is the Group Leader for Technology Integration in the National Center for Computation Sciences at Oak Ridge National Laboratory. He can be reached by E-Mail: canonrs@ornl.gov. Sarp Oral is a member of the Technology Integration Group and is chiefly responsible for investigating options for the centralized file system. He can be reached by E-mail: oralhs@ornl.gov.

References

[1] Cluster File Systems, Inc. Lustre change log. Web page. http://www.clusterfs.com/changelog.html.

[2] Cluster File Systems, Inc. Lustre manual. Web page. http://www.lustre.org/manual.html.

[3] Cluster File Systems, Inc. Lustre road map. Web page. http://www.clusterfs.com/roadmap.html.

[4] S. M. Kelly and R. Brightwell. Software architecture of the light weight kernel, catamount. In *CUG Proceedings*, 2005.