StorNext SAN on an X1

Jim Glidewell, Boeing Shared Services Group

ABSTRACT: *With the acquisition of a Cray X1 in January 2004, Boeing needed a replacement HSM (Hierarchical Storage Management) system which could support a variety of clients, including the X1. ADIC's StorNext product was selected as the management software for a number of HPC (High Performance Computing) systems. This paper describes our early experiences with the StorNext Management Suite serving a Cray X1.*

Introduction

Boeing has been using Cray systems to assist our engineering staff in designing airplanes and related systems since 1980. Support for these Cray systems is provided by the Boeing Shared Services High Performance Computing (HPC) group. Our current HPC equipment includes two Cray T-90s, a 384-CPU Origin 3800, and a Linux cluster consisting of 384 dual-CPU nodes. All of these systems are connected to a group of three STK Powderhorn tape silos.

Our site has been using DMF (Data Migration Facility) on Cray systems since 1980, and have used it on our Origin 3800 as well. We have a total of 37 terabytes of migrated data from the Cray T-90's, and another 38 terabytes of data associated with the Origin 3800.

X1 Configuration

The Cray X1 is a fully-configured single liquid-cooled chassis, with 64 MSPs and 512GB of memory. Disk storage consists of a total of 26 terabytes of LSI RAID disk arrays, of which roughly 7 terabytes are currently managed by ADIC's StorNext Management Suite (SNMS). SNMS communicates with ACSLS 7.0 on the STK silos to coordinate tape loading and unloading of drives, as well as management of tape residency in the silo.

SNMS is hosted on a Dell 2650 server, running Linux. This system is connected to six STK 9840C tape drives mounted in the silos. These drives were chosen based on their compatibility with other services in the datacenter and their relatively fast load time. There are other sites within Boeing which make use of ADIC tape libraries.

The StorNext Management Suite

The StorNext Management Suite (SNMS) consists of a combination of software components which provide for a complete SAN-based HSM (Hierarchical Storage Management) system. A key component of SNMS is the StorNext File System (SNFS), a journaled file system offering simultaneous access from a variety of platforms. The division between the SNFS and the rest of the management suite is not absolute, however, and other SNMS components are essential to the functioning of the file system.

Much of the terminology used by SNMS is similar to DMF – terms such as “migrated”, “online”, and “offline” are common to both systems. But the management commands are entirely different, as are the methods for specifying migration policies.

Installation and Configuration

Our primary goal in configuring our SNMS managed filesystems was to provide maximal performance for a heterogeneous mix of file sizes, data access methods, and access patterns. The initial setup of the SAN was significantly more difficult than we expected. Neither the Boeing or Cray analysts who were developing our configuration were aware of all the constraints involving in configuring a StorNext file server. Our initial plans for SAN filesystem layout called for as many as 26 individual filesystems, but we were learned quite late in the process that SNMS could only support about eight, and that four was a better limit if we wanted to use failover between the primary and backup file system servers (FSS’s). Even though Cray had a SAN analyst on site for much of the time, it still took several months to get the SNMS server into a production state.

Performance

Much effort during the install phase was spent trying to improve performance, which failed to meet our expectations. Of particular concern were write speed and certain metadata functions – for example, removing 4000 files took almost 30 minutes (this was fixed with a updated release of the X1 SAN client). In an attempt to close the performance gap between direct attached disks and identical disks managed by SNMS, we eventually deviated from Cray’s standard configuration by enabling CTQ (Command Tag Queuing) and write caching on the LSI RAID controllers. Further performance enhancements were achieved by modifying the X1 mount options, though we suspect that we may be hiding an underlying performance problem with lots of caching.

Our experience with SAN performance is summarized by the table below:

	Reads	Reads	Writes	Writes
	Direct	SAN	Direct	SAN
cp	130 MB/s	250 MB/s	120 MB/s	30 MB/s
cvcp	367 MB/s	175 MB/s	300 MB/s	97 MB/s

As can be seen by this table, while read speeds are generally adequate for both local and SAN disk, SAN writes are slower by a factor of 3 to 4 relative to local disks, even after extensive tuning. The “cvcp” command is an optimized version of “cp” provided by ADIC to maximize data rates during data movement in and out of the SAN. Both ADIC

and Cray have told us that the write speeds that we are seeing are significantly less than what should be expected, and ongoing efforts are being made to improve performance.

Operational Issues

Our initial configuration and testing of the StorNext SAN exposed a number of operational issues, some minor while others are more serious. Overall, we found that SNMS was not as resilient as we had expected. In multiple cases, we were forced to restart various SNMS daemons, and in some cases reboot the file system server, in order to recover various error conditions. Such activities can cause outstanding I/O requests to abort, and in the case of a file system server reboot may require remounting of the SNFS file systems or even rebooting of the client X1.

The interface between SNMS and the STK silo management software, ACSLS 7.0, has a number of deficiencies as well. SNMS views itself as the owner of the entire tape library, which on our site is shared among several systems. The movement of tapes in and out of the silo is a delicate process, and one that we are still trying to understand. The "fsmedcopy" utility (which merges data from multiple tapes) is slow and awkward – we need to test this facility further.

It should be noted that we tested a large number of failure modes, including a simulated disaster during our pre-production period, which exposed some problems which would never be seen in normal production. Nevertheless, we have generated a very large number of SPRs (Software Problem Reports) related to the SNMS SAN.

User Experiences

In contrast to the issues facing the system administrators, the user experience of the SNMS SAN has been quite positive. Besides a very slight sluggishness in metadata access, the SAN has been fairly painless for our users. For the most part, our users were unaware that we had made the transition to the SAN, which we view as a very positive sign.

The lack of a client-side "dmget" analogue may become a serious issue when a significant percentage of our files are offline. ADIC has taken our request for such a facility, which will be of great benefit to our users when we have a large number of migrated files.

Production Experience

Our StorNext SAN went into full production, with the transfer of all user home directories and permanent data, on April 25, 2004. We have seen no service interruptions since the SAN went into production. Our users appear to be satisfied with the performance and general user experience.

Despite a number of operational issues, the StorNext SAN appears to be a solid product, which offers a number of features which are extremely valuable to our datacenter. The ability to manage data in excess of our physical disk capacity has been an essential part of of our HPC services. The ability of StorNext to offer high performance shared access from heterogeneous platforms provides us a number of options to address our user's needs for sharing of data between our HPC systems. The StorNext "trickle backup" has relieved us of the need for traditional weekly base plus daily incremental backups. And its support for multiple copies of backup media provides us the ability to meet our requirements for offsite disaster backup media as a routine part of the backup process.

The willingness of Cray and ADIC to address our concerns, and their quick response when a critical problem arose, factored heavily in our decision to move forward and put the StorNext SAN into full production.

Linux Server Platform

While the choice of Linux for the SNMS file system server is well-aligned with the general strategic direction of Cray and HPC in general, there are reasons to question its suitability for the role of an SNMS server at our site. SNMS on Linux is the newest platform that ADIC supports, and it appears that it may not be as mature, robust, and resilient on Linux as we had hoped. Furthermore, solid support of high-volume I/O is not as well developed on Linux as it needs to be to meet our requirements for performance. While Linux may be an ideal long-term strategic direction, Boeing needs a robust, well supported overall solution now.

Conclusion

Our experience with the installation of the ADIC SNMS management suite was significantly more difficult than expected, for both Cray and us. Installation and production readiness took longer than we had expected - months, rather than weeks. Cray's lack of a permanent SNMS test bed (which they now have) had significant impact on their ability to respond to problems.
In retrospect, it seems that getting ADIC directly involved with the problems we were seeing might have been beneficial. We do believe that other sites will benefit from the lessons learned at ours.

Boeing would benefit from a more "turn-key" solution to managed storage. We suspect that many of the operational issues that we experienced are unique to SNMS on linux (except for issues related to ACSLS). But we do not have experience with SNMS hosted on other servers.

StorNext appears to be a solid product, but for us it needs more polish in a number of areas, including ease of setup, HPC features (such as "dmget" functionality and ACL's), and handling of exceptions.

Performance still lags for our requirements, but was deemed sufficient for production use. We put our StorNext managed SAN into production on April 25, 2004. Despite concerns about operational issues, we are sufficiently confident in SNMS to entrust our users' data to that system. We look forward to continued improvements and thank Cray for their significant efforts to make the installation of SNMS at our site a success.

About the Author

Jim Glidewell has been a member of Boeing's HPC group for over twenty years, working on a variety of systems from Cray, SGI, CDC, and others. He is currently serving as the CUG Operating Systems SIG Chair. He can be reached at The Boeing Company, P.O. Box 3707 MC 7J-04, Seattle WA 98124-2207; E-mail: james.glidewell@boeing.com