Storage Management at Cray Research, Inc.

David J. Metcalfe and *Dave Thompson*

ABSTRACT: *Storage Management continues to play a key role in providing a complete solution to todays most challenging computational problems. This paper describes the current market trends and customer requirements driving Cray's continued commitment to providing an integrated storage management solution. Product directions for the areas of network backup, archive and hierarchical storage management (HSM) will be described.*

1 Introduction

Reliably managing large amounts of data and providing access to those who need it, when they need it and where they need it, is a problem in many of todays computing environments. The problem is getting more complex as the data volumes increase and access requirements expand to wider area, higher speed networks. Data and information are key corporate assets and maintaining the integrity and security of them while allowing timely access to those that need it is a critical component of any computer installation.

Cray Research has a long history in successfully managing high volumes of data. As Cray systems typically generate large amounts of data through complex simulations, we have had a vested interest in managing the data. By providing capabilities such as high-bandwidth I/O, high performance disks, tape libraries and industry leading hierarchical storage management software, we have provided solutions for the largest data management problems today. Many Cray sites manage multi-terabytes of data, with many gigabytes of new data being added daily.

In order to maintain our position of providing a storage management solution for our customer's growing requirements, we have undertaken detailed evaluations of various storage management products. Some of these, such as the previously announced OSM port, while attractive within their own market space, did not meet the high-end needs of our existing or prospective customer base.

This paper presents the customer driven requirements, the results of our product evaluations and the direction that Cray Research is taking to remain a leading provider of high-end storage management solutions. Three areas are focussed on: Network Backup, Archive and Hierarchical Storage Management. These are defined below. Other key factors in providing a complete storage management solution, such as local and distributed file systems, high speed network protocols and peripheral support are not discussed here.

2 Network Backup, Archive and Hierarchical Storage Management (HSM)

2.1 Network Backup

The primary objective of doing file backups is to be able to recover lost or corrupted information. Whether it is a user accidentally deleting a file, a disk crash, or a data center disaster - the goal is to find and restore damaged or missing information quickly. Backups must be performed on a regular basis. As a backup ages, it soon outlives its usefulness as a copy of the data. How often backups are performed depends upon the criticality of the data. The backup cycle is determined by the interim of new data and changes that you can afford to lose. The backup process is cyclical. Snapshots of the data are taken in time, stored in a safe place for a limited amount of time, after which the media is re-used for new backups.

The backup problem has become particularly acute with the growing use of networked computing. With many users on a network, it is impractical to require each one to be responsible for the backup of his/her data. Many organizations using workstation LAN's are moving to a network wide backup solution in which the backup is done automatically by the backup server according to a schedule set up by the network administrator.

2.2 Archive

The purpose of an archive is to keep important information that is no longer in everyday use but that may be required again at sometime in the future or that must be kept in order to comply with government, industry, or company regulations. While a backup is a "snapshot in time" of data in use, an archive is a copy of information no longer in use but that you want to keep for a relatively long time - in some cases, the goal is forever.

The emphasis of a file archive system is low cost of the storage media and high reliability. The media must be stable over time, economical and depending on the required retrieval rates, reasonably fast.

2.3 Hierarchical Storage Management (HSM)

The primary objective of an HSM is to preserve the cost effectiveness of how data is stored. The whole premise of HSM is based on the fact that the typical environments, as much as 80% of all disk requests are for 20% of the data. In other words, if it is possible to identify 80% of the so called "dormant" data that is not being used and move it to less expensive media, total storage management costs could be reduced. HSM manages the migration of unused files to secondary storage and also the retrieval of the files back to disk when the user requests them. Beyond an increased access time for migrated files, this process is carried out transparently to the end user or application.

HSM's are becoming increasingly common; however, workstation based solutions to date have not proved to be as successful as the number of products in this area would suggest. HSM's have proven to be very difficult to produce effectively. This, coupled with ever decreasing disk costs per megabyte and the inherent complexities involved with reliably and efficiently moving large volumes of data have produced a delaying factor in wide scale use of this technology. This is not true within Cray environments where the Data Migration Facility (DMF) has proved to be a highly effective HSM solution.

3 Strategic Vision

Cray's commitment is to provide a data storage architecture balanced to match computational speed requirements. The storage solution allows active data to be available at rates that do not impede effective use of the systems processing capability. Balanced with this need to support high performance processing and not delay user access to data, is a capacity requirement to store massive volumes of data while maintaining its security and integrity. The Cray solution scales to meet the increasing data storage requirements of our user community.

The same arguments hold whether the data is being generated on a Cray system or simply managed there and generated or accessed elsewhere. Data must be an open commodity. Data managed on a Cray system needs to be accessible both to the Cray itself, or other Cray systems, but also to the wider computing community in which the system lives. Support for industry standards such as NFS, DFS and FTP provide remote access to Cray managed data. In addition, we are exploring ways to provide a tighter integration between Cray storage and remote systems in order to provide a seamless storage environment across an enterprise.

Another factor in providing an open storage solution is the recognition that many sites already have centralized storage solutions that are served by non-Cray platforms. In such environments, it is important that the Cray systems are able to integrate into the existing scheme when it is appropriate to do so. A good example of this is that of centralized network backup systems. Cray needs to provide the "client" side of a number of network backup products to allow our systems to be integrated into the larger picture.

4 Industry Trends and Requirements

Cray Research systems are not the only machines generating large quantities of data. The information explosion is real and conservative estimates put the volume of data stored on-line as increasing by 15 to 40 percent annually. Within the scientific community, there is evidence to suggest that this number is currently nearer 100 percent annually. The data growth rate is fueled by growing computational power, an increased use of large scale computer simulation, new data-intensive applications (such as image processing and data mining), a reduction in disk storage prices, a reluctance to delete "old" data and the continued increase in the computer user-base. "Large" systems are measured in terabytes and those supporting petabytes will become a reality within the next 3-4 years. Even with decreasing disk storage costs, data continues to overflow the physical limits all too quickly.

Not only is the volume of data growing but its importance is too. Information is a key corporate asset. Timely access to it from all parts of an organization is becoming a vital component of a business. Whereas computing power is becoming increasingly decentralized through more powerful desktop systems, "information" is becoming highly centralized.

This requirement to store increasingly large volumes of data transcends all of our existing customer bases with clear examples within Government organizations, environmental, automotive, petroleum, entertainment, etc. industries. In order to assess where these sites are, and where they expect to go with data storage and their requirements, we have undertaken various market studies. A survey conducted at the Alaska Cray User Group meeting in the Fall of 1995, produced results which have proved representative of other research. Some of the results are shown below.

4.1 Data Storage requirements:

- 45% of the sites responding manage 1-10TB of data
- 19% of the sites responding manage 10-100TB of data
- 1 site manages over 100TB's of data.

These results closely match the wider Cray installation base for those using the Cray Data Migration Facility (DMF). In addition to the identified sites already managing +100TB of data, there are plans to move some sites to the petabyte level within the foreseeable future.

4.2 Data Growth Rates

Over the next 3 years all respondents expected their storage requirements to increase.

- 30% of the surveyed sites will grow by 50%.
- 53% of the surveyed sites will grow by 100%

4.3 Key Requirements for an HSM

The following were identified as key requirements for a successful HSM solution.

- Reliability and data integrity
- Storage capacity
- Scalability
- Security of the data
- Performance
- Network Access
- Openness (i.e. heterogeneous access)
- Easy transition from existing environments.

The first requirement stands alone as an "absolutely must have" condition. Without a certain, high, level of assurance that allowing an HSM to manage critical site data will not result in its loss or corruption, the HSM will not be considered.

Many of the other requirements clearly go hand in hand. Providing a fast, reliable solution to manage 200GB of data may be achievable. Using the same product to scale up to a terabyte may result in storing or retrieving data not being possible within the required time window.

Performance was also noted to include making effective and efficient use of the expensive tape and robotics equipment. For example, providing fast access to a tape-robot may not be acceptable if the speed is attained by dedicating the robot to the one storage application.

Cray Research used this market data and the customer requirements to review where we are and where we need to head with our storage management solution.

5 Product Evaluations

Research into storage management products has concentrated on three areas: Network Backup, Archive, and HSM solutions.

5.1 Network Backup and Archive

The more urgent customer-directed requirement here was to provide a network backup capability. The specific requirements were identified as follows:

- Reliability and capacity
- Open/network access, i.e., a network wide solution
- "Reasonable" performance

The need for an open solution, conforming to existing site practices indicated that a Cray provided solution would not be of significant interest. Customers require third-party products to be available under UNICOS to integrate the Cray systems into their existing backup strategy. Specifically, the client component of the network backup software such that Cray data may be transferred to a backup server running on a different system.

As such, we have worked with a number of 3rd-party providers of network backup software, providing porting assistance when necessary, to provide this capability. Four products are now available, and supported by the respective companies:

- IBM's ADSM V2R1, available 4/1/96. This is an IBM Service Offering.
- OpenVision's AXXiON-Netbackup (formerly OpenV*Netbackup). Available today.
- MultiStream system's HYPERtape. Available today.
- Software Moguls SM-*arch*. Available today.

Please contact the respective companies for additional details.

It is recognized that there are many alternative network backup products available today. As customer requirements and business needs dictate we will pursue additional 3rd-party relationships

We expect the archive capability to be provided along similar lines. That is, through integration of 3rd-party products. As yet, no relationships have been formed to provide that capability.

In addition to these 3rd-party products, many Cray sites utilize existing Cray software to provide backup and archive capabilities. For sites with data volumes too large to effectively move off the Cray system via a network, the optimized dump/restore utilities still provide a very efficient means for local backup activity. Cray systems are also used as central storage servers for networks of non-Cray systems. For example, some sites run 3rd-party network backup clients on a variety of platforms which automatically transfer data into a Cray managed file system. This file system is under the control of Cray's HSM - the Data Migration Facility - which moves the data off-line. This combination matches the flexibility and openness of the 3rd-party storage products with the respected capacity, performance and integrity of Cray's DMF facility.

5.2 Hierarchical Storage Management (HSM)

The key requirements for an HSM solution were noted in section 4. Using these evaluation metrics, coupled with data on market presence and end-user feedback, we evaluated a number of HSM solutions. Each have different characteristics, strengths and weaknesses but each have the underlying goal of providing an "infinite storage" capability. That is, to allow over-subscription of the physical disk resources in a user transparent manner.

5.2.1 Computer Associates Open Storage Manager (OSM)

OSM (which was owned by Legent when our evaluation took place) is a family of storage management technologies that, when combined, provide automated storage management across a network of clients and servers.

In 1995, Cray Research announced plans to provide the OSM product set under UNICOS. This was premature. A detailed technical evaluation proved that OSM would not meet the requirements laid down by our customers. OSM does have many attractive qualities and does meet a number of the requirements, especially in the area of openness that we were looking for. However, we found that the product is optimized for a lower level of data storage requirements than is typical within a supercomputer environment. This is true now and certainly with the growth rates that we see happening over the next few years.

As such, it was decided not to pursue the OSM technology and, instead, continue to investigate alternative strategies to meeting the storage management requirements.

5.2.2 EMASS' AMASS

EMASS is a subsidiary company of E-Systems dedicated to storage management solutions. EMASS markets several storage management products including AMASS, the AMASS data manager, AMASS Admin and Fileserv (see section below). The AMASS product provides an infinite store file system capability. That is, files may be written to and read from an AMASS file system which is automatically mapped to off-line storage.

While the AMASS product has a definite place within the high-end storage management industry, there was not seen to be a high level of "product fit" between what our customer base was requesting and what the AMASS product provides. It was also felt that our existing Data Migration Facility offers sufficient product overlap to AMASS that it would not be productive to try and support two such technologies, nor of significant benefit to force customers to switch from DMF to AMASS. As such, there is no activity between EMASS and Cray Research to port the AMASS product to UNICOS.

5.2.3 EMASS' Fileserv

Fileserv is developed and sold by EMASS. It is (in words from EMASS' world wide web pages):

"This software solution balances mass storage tape media with a client computer's on-line disk media in such a way that the tape data appears to be on-line to the user. The FileServ software is very flexible and can be configured on-line to the users' needs. FileServ software features DataClass groupings to segregate data and control the data migration between disk and tape media. This mechanism conserves valuable disk space by migrating inactive files to tape media. Later, when the user needs the file again, all that he needs to do is access the file using standard UNIX commands - FileServ software will then retrieve the file to disk automatically."

FileServ is very similar in capability to a Cray's earlier versions of DMF. Following a similar argument for not embracing the AMASS product, we saw no customer benefit, nor significant market opportunity in providing this product. Discussions with EMASS concluded with EMASS deciding not to complete the port of this product to UNICOS.

5.2.4 UniTree

From the NSL UniTree home page:

"NSL UniTree is a hierarchical storage management system designed for environments where storage requirements range from a few hundred Gigabytes to Petabyte sized archives and input/output rates are in the tens of megabytes per second. Based on the IEEE Mass Storage System Reference Model, NSL UniTree provides an open system storage solution for High Performance Client/Server (HPC/SI) computing environments. NSL UniTree provides users with transparent access to virtually unlimited storage space. Acting as a virtual disk, NSL UniTree manages users files automatically by migrating files to less expensive storage according to a programmable migration strategy."

UniTree has had a large following for some years. It has proved to be a successful piece of software in terms of units sold. As such, it was an obvious candidate for Cray Research to consider. An initial port was done by Titan in 1994. However, it was found that for similar reasons to AMASS and Fileserv, there was no real market opportunity for UniTree within Cray's typical storage management scenarios. It did not offer the scaling, nor data integrity necessary to meet the highest levels of end-user requirements.

In addition, it is recognized that any significant UniTree enhancements in the future are being funneled more towards the High Performance Storage System. See below. As such, Cray Research does not and has no future plans to offer the UniTree product under UNICOS.

5.2.5 High Performance Storage System (HPSS)

The next generation storage system under development at the NSL is called the High-Performance Storage System (HPSS). HPSS is a collaborative development effort with numerous industry, government and University participants.

The objective of the HPSS project is to develop a standards based, distributable, hierarchical storage system focused on scalability, in size and performance, and modularity of software components. Features of HPSS include: support for a large number of storage devices, multiple storage hierarchies, and a variety of API's including a parallel I/O storage system API, Parallel FTP (and sequential), NFS, with AFS, DFS and DMIG to be supported in the future.

While Cray Research sees significant promise in this activity, the timing of expected product availability (it is currently at "special project" status) is such that it can not play a part in solving our existing customer requirements. We will continue to monitor activity in this area and take customer input and at such time that it is likely to become a storage management mainstream product, we will re-evaluate. Interoperability with HPSS systems on non-Cray platforms will be addressed as required.

5.2.6 IBM's ADSM

IBM's ADSTAR Distributed Storage Manager (ADSM) is a backup/recovery solution for heterogeneous networks. It backs up and archives data to an ADSM server. It provides a graphical user interface, security features, different storage media, and several types of client machines.

IBM claim that "ADSM is IBM's premier enterprise storage management solution that provides storage management services in a multi-vendor, multi-platform computer environment."

For the market segment that IBM is targeting for the ADSM product, it provides a very effective solution. We recognize the large following that this product is attracting and supported the actively that allowed the network backup client component to run under UNICOS. However, the product is aimed more at the workstation environment and does not readily scale to the super-

computing environments which typically deal with an order of magnitude more data.

At this time we do not see any added benefit in having more than the client component of ADSM run under UNICOS. The nature of the ADSM product would not allow effective use of the Cray architecture, nor scale to the levels demanded of our storage solution.

5.2.7 Cray's Data Migration Facility (DMF)

The Cray Data Migration Facility (DMF) is a highly-used storage management service running under UNICOS. There is, arguably, more data currently under DMF's management than under any other Unix storage management system. DMF is in use on approximately 200 Cray systems around the world.

DMF is a comprehensive space management tool. It maintains free space on disk by moving old or large files (with details under administrator control) to more efficient, off-line media. File migration can be manual or automatic. Automatic migration is transparent to the user. DMF is implemented as an extension of the native UNICOS file system (NC1FS) and maintains full POSIX compliance. It provides seamless interaction to UNICOS system software (such as quotas and multi-level security) together with allowing network data access via NFS, DFS and FTP. DMF provides Cray users with a transparent "virtual disk" capability and does so with a minimal increase in system overhead. Typically, there is around a 5% system overhead increase observed when running DMF.

Most DMF installations manage in excess of 1 terabyte of data, with a significant number managing multiple terabytes and growing. Larger systems span into the 100's of terabytes, some serving greater than 4 million files. Other sites are more concerned with a so called "data velocity" requirement which is a measure of how much data is moved through the DMF in a given time period (which is also related to the ratio of data to physical disk, or the over-subscription ratio). There are examples of sites moving two terabytes of data through DMF on a daily basis. Oversubscription rates of 10 or more are not uncommon.

5.3 HSM Conclusions

In order to satisfy existing and future customer requirements for an HSM solution, our analysis shows that DMF offers the best path forward. However, the analysis also shows areas in which DMF needs to develop for it to remain a primary HSM solution. Some of these are more internal in nature, others affect the overall structure of the product. The DMF product plan details both short term activity, improvements and longer term directions.

Some of the planned changes recognize the fact that as DMF licenses are approaching 200, reliance upon field support personnel to fix certain conditions is no longer practical. For example, DMF 2.4, released in February 1996 does not use the facility called a "pre-migration" directory to stage files before they are migrated. Although rare, it was possible for an unscheduled system interrupt (such as a disk spindle failure) to leave DMF in an inconsistent state. This could require expert intervention to repair. By changing how the pre-migration directory is used, in DMF 2.4, this particular problem will not occur.

Another example is in the database technology used by DMF. The current database software will be replaced by a more robust, 2-phase commit database package in DMF 2.5. Existing DMF administrators will see minimal conversion impact (which will be supported by a database transition utility) but will see reduced down-time due to unscheduled interrupts. Once the DMF databases have been converted, end users will see no usability differences. DMF 2.5 will be available in 3Q96.

In terms of the overall DMF structure, the direction is to provide a true client server architecture. In earlier releases of DMF, the code structure was monolithic in nature. There were few clear distinctions between the code to interact with the file system (e.g. to select files for migration) and that to interact with the tape subsystem. Starting in DMF 2.2 and developing with DMF 2.3 (called "Distributed DMF"), the structure of DMF now consists of client and server components. At this time, although the client and server components may run on different Cray systems, there is a reliance on Cray's Shared File System (SFS) technology.

DMF 3.0 will remove the dependency on the SFS, allowing DMF to operate across a network of Cray Research systems in a true client/server, or open, fashion. This will allow DMF to support systems which do not have direct tape support. That is, data from Cray system running a DMF client will be transferred to another Cray system running the DMF server and then to secondary storage. This process of migrating and, equally, restoring files will be under end-user control, or remain transparent to them as required.

DMF 3.0 will also remove a number of internal dependencies on Cray's NC1 file system. DMIG (Data Migration Interface Group) activity is being tracked and we would expect to support any standards in this area as they become available. This will allow the DMF product to be quickly adapted to new environments as requirements and business reasons dictate.

More details of the existing DMF product and its development path are described within the paper "Data Migration Development Update", Barcelona CUG, 1996 by Neil Bannister and Thomas W. Lanzatella.

6 Summary and Conclusions

Cray Research has an established reputation for efficiently managing the highest levels of data storage. This reputation is built partially on hardware characteristics such as high-bandwidth I/O and support for high performance disk, tape and robotic devices. However, a significant portion of the success is a result of Cray's Data Migration Facility. After an extensive evaluation process of the available storage management technology, we have identified no products that meet the customer requirements for an HSM solution as effectively as DMF. As such, Cray Research is re-affirming its commitment towards the DMF capability and proposing a development path which

expands its capabilities while solidifying the core requirements of managing today's largest storage management environments.

In addition to enhancing this key component of our storage management strategy, Cray Research also recognizes the need to support a key set of 3rd-party storage management products. In order to integrate into existing environments, we have worked with customer recommended vendors to provide various network backup client products. As requirements dictate, we will evaluate expanding the number of 3rd-party products and provide similar support for the archive capability.