File Migration on the Cray Y-MP at the National Center for Atmospheric Research*

Ethan L. Miller
Computer Science Division
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720
elm@cs.berkeley.edu

June 25, 1991

Abstract

The supercomputer center at the National Center for Atmospheric Research (NCAR) migrates large numbers of files to and from its mass storage system (MSS) because there is insufficient space to store them on the Cray supercomputer's local disks. This paper presents file migration data collected over a 10 month period and some analysis of the data. The analysis shows that requests to the MSS are periodic, with one day and one week periods. Read requests to the MSS account for the majority of the periodicity; write requests are relatively constant over the course of a week. The intervals between requests to the MSS appear to be bunched, since well over half of the intervals are shorter than half the mean interval length. The latencies to access the first byte from disk, automated tape library, and manually-mounted tape are examined. In addition, information about file size distribution is given. We found that the NCAR system was different from systems studied earlier in several ways, including the ratio of local disk to tertiary storage. These differences affect the way that file migration can be done, so previous migration algorithms may not be effective.

1 Introduction

Over the last decade, computers have made incredible gains in speed. This speedup has encouraged the processing of larger and larger amounts of data; however, storing this data on magnetic disk is not feasible. Instead, most data centers with large data sets use tertiary storage devices such as tapes and optical disks to store much of their data. These devices provide a lower cost per megabyte

*This research was supported by contract S9128 with the University Corporation for Atmospheric Research (UCAR) which is sponsored by the National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of UCAR nor the National Science Foundation.

of storage, but they have longer access times than magnetic disk. By studying the tradeoffs between cheaper and slower tertiary storage and more expensive and faster disk storage, response time can be improved without increasing storage costs.

This paper will first cover the history and current solutions to the problem of mass storage. In the second section of the paper, we will present file migration data gathered on the Cray Y-MP used by the National Center for Atmospheric Research (NCAR). We will present the data and do some elementary analysis of it, particularly stressing its similarities to and differences from previous file migration studies. Finally, we will summarize current file migration policies and how they would perform on the system at NCAR.

2 Background

2.1 History

File migration systems are used by most large computer installations to store more data than that which would fit on magnetic disk. Tertiary storage, which usually consists of tape and optical disk, lies at the bottom of the "storage pyramid," which is shown in Figure 1. Cost and speed increase going up the pyramid, while the size of the memory level increases towards the bottom of the period. CPU cache, or perhaps even CPU registers, are at the top of the pyramid; they have the highest cost per byte and are the smallest and fastest of the levels. At the bottom of the pyramid are tape and optical disk, which have slow access speeds, on the order of seconds or minutes, and very low cost, under \$10/GB.

Early mass storage systems used manual tape mounting, since it was cheaper to hire system operators than it was to have a robot manage tape mounts. However, by 1978, several companies had introduced automated tape systems [Boy78], and automated tape storage became part of the mass storage systems in most major systems. Several systems were studied in the early 1980s; these included Brookhaven National Laboratory [EP82], the University of Illinois [LRB82], and the Stanford Linear Accelerator Center [Smi81b,Smi81a]. These will be discussed in a later section.

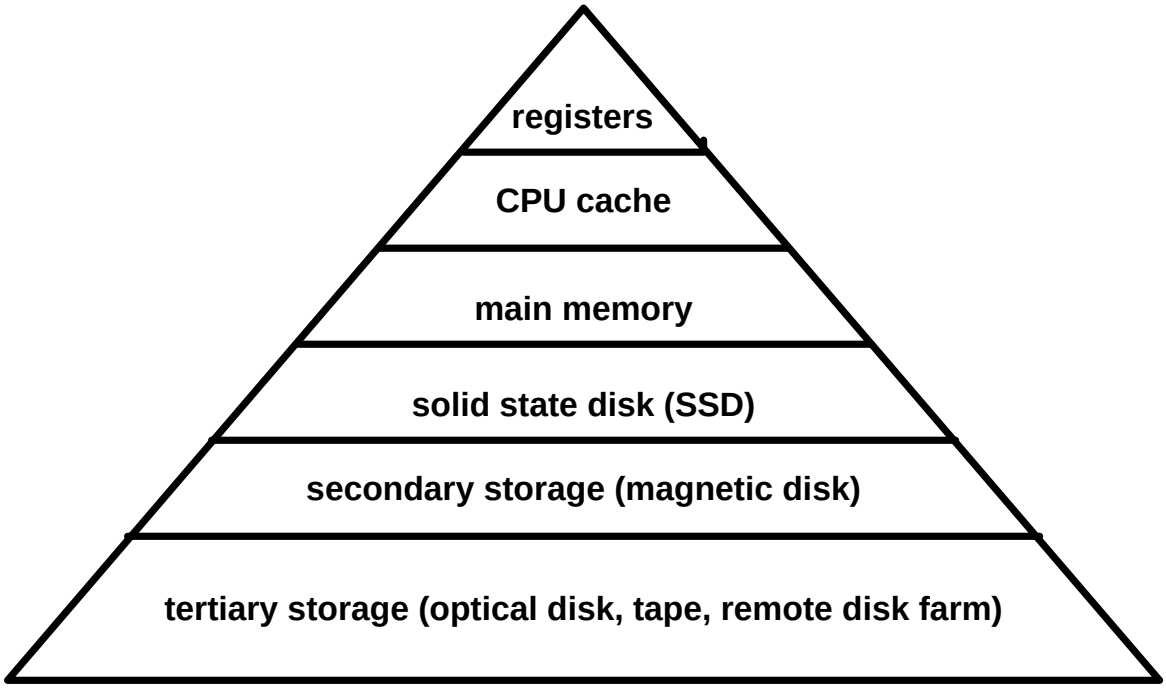

Figure 1: Memory hierarchy in large systems.

Since these studies, many complex mass storage systems have been implemented [McC87, NKMH87,HP89]. However, no studies on these systems have been published. Instead, the data management staff at these sites collect huge amounts of data to justify new equipment purchases and tune their systems. While this guarantees good performance for each system, it does not provide any guidelines for building future systems.

2.2 Mass Storage Devices

Currently, there are two major types of tertiary storage devices in common use—tape and optical disk. Both of these are high-density removable media. The tradeoffs between the two media are presented in Table 1. Two types of magnetic tape, helical scan and longitudinal scan, are presented. The numbers for the tapes come from [Woo88], while the optical disk statistics come from [Spe88]. The only change from the published numbers are for the IBM 3480, which is a longitudinal scan tape. This tape has now come out with a double density version, so the numbers for capacity per tape and cost per megabyte will be adjusted accordingly.

Category	Optical Disk Jukebox	Linear Tape	Helical-Scan tape
Media capacity (GB)	1.2	0.4	2.0
Random access speed	7 sec	13 sec	60 sec
Transfer rate (MB/sec)	0.25	3.0	0.25
Media cost/GB	$250	$35	$7

Table 1: A brief comparison of optical disk and tape.

The major tradeoffs among the three media are access latency and transfer bandwidth. Optical disks have a much lower access latency than either type of magnetic tape, but their bandwidth is also considerably lower. Thus, a system which performs many small I/Os to tertiary storage, such as a database system, would be best served by optical disk, since the dominating factor in calculating time per byte is access time to the first byte. For supercomputing installations, however, magnetic tape is better. While the time to get the first byte of data is longer for tape than for optical disk, the time to get *all* of the data is often lower for tape. Files on supercomputing installations tend to be large [Wal91], so the difference in transfer time between optical disk and tape is substantial. The two types of tape differ mainly in cost per MB and in transfer rate. Helical scan is a newer technology, and transfer rates are expected to rise. Another new technology, optical tape [Spe88], also looks promising because of its high density storage and high transfer rate.

Another primary consideration is price per gigabyte. As can be seen in Table 1, magnetic tape has a lower cost per gigabyte stored than optical disk. For systems with terabytes of data stored on tertiary storage, such as NCAR, this cost difference alone is enough to favor using tape exclusively as the tertiary store. The lower cost and higher transfer rate make magnetic tape the obvious choice for supercomputer centers which deal with sequentially-read large files.

Currently, the IBM 3480 tape format is standard at most supercomputer installations, though some are beginning to move to higher-density tapes such as Exabyte. The IBM 3480 uses linear recording, which provides high speed at the expense of recording density. The Exabyte drive, on the other hand, uses helical scan techniques (similar to conventional VCR recording) to greatly increase recording density; however, transfer rate is currently quite low on such tape drives.

Most installations today have one or more cartridge tape robots to automatically mount some of their tape libraries. An example of a tape robot, or automated cartridge system (ACS) is the StorageTek 4400 [LYSK87]. This system can provide access to 1.2 TB of data (6000 IBM 3480 cartridges, holding 200 MB each) without human intervention. Loading a cartridge takes approximately 6 seconds; from there, tape characteristics are identical to the IBM 3480.

2.3 Previous Work

There have been several studies of actual file migration systems, but they are quite old and deal with different computing environments. We will summarize them here, and in a later section will compare the results of studying the current NCAR environment with the results of the earlier studies.

In [Smi81b] and [Smi81a], Smith studied the file system at the Stanford Linear Accelerator Center. His data dealt with Wylbur text editor data sets, and tracked the references to those data sets. He found that the best algorithms had access to the entire reference string for a file. Since this is often not feasible, the migration criterion he suggested was to migrate off disk the files with the highest value of $last\ reference\ time^{1.4} \times file\ size$. This algorithm, called Space-Time Product (STP**1.4), was the best of the algorithms examined which did not make use of any file history other than the last reference time. The analysis in the paper also did not consider the possible effects of transfer time and access latency in minimizing average file reference time; instead, the analysis attempted to minimize file miss rate.

Smith also made several observations about file system activity. He noted that usage followed a weekly pattern, with activity highest on weekdays and lower on weekends and holidays. He also has extensive data on file sizes and interreference intervals; because of the size of the data set in the NCAR study (over 600,000 files), it would be very difficult to perform the same computations over the entire file set. The data set in the paper has a granularity of one day and does not distinguish between reads and writes.

For the data set in the paper, none of the acceptable migration algorithms would have had

much effect on average file access time. As noted in the paper, a miss ratio of 1% would mean a loss of 6.26 man/minutes per day, given the file usage rates and the number of users on the system. For STP, this miss ratio would require a disk system that held 1.5% of the total tertiary storage, and would require 300 tracks, or about 1 MB, of data to be transferred each day.

Lawrie, et. al., in [LRB82], considered the file migration patterns on the University of Illinois Cyber 175. Again, the system examined is quite different from the NCAR system studied in this paper. Interestingly, Lawrie reported that, though his system was quite different from SLAC, his results matched Smith's closely. This paper also examined several migration algorithms, and compared them against Smith's STP algorithm on their data. They found that STP was better than the algorithms they tried, which included pure LRU, pure length (migrate large files first), and SAAC, which migrated files that became less active. In all cases, STP outperformed these algorithms, though only by a slim margin.

Other papers have simply presented data gathered from existing mass storage systems without analyzing the data and suggesting possible algorithm changes. Systems analyzed include Brookhaven [EP82], NCAR [TH88,AN88], and NASA [HP89]. In addition, many large sites internally publish a summary of statistics gathered from their machines. They use these statistics for two purposes: to better tune their systems, and to justify new equipment purchases.

3 NCAR System Configuration

In this section, we describe the system on which the file migration traces were gathered. Rather than describe the entire NCAR network, we will focus on the parts which are relevant to the study. However, the rest of the network will be briefly described, since the mass storage system is shared by all of the systems at NCAR, so their presence might have an effect on mass storage systems performance.

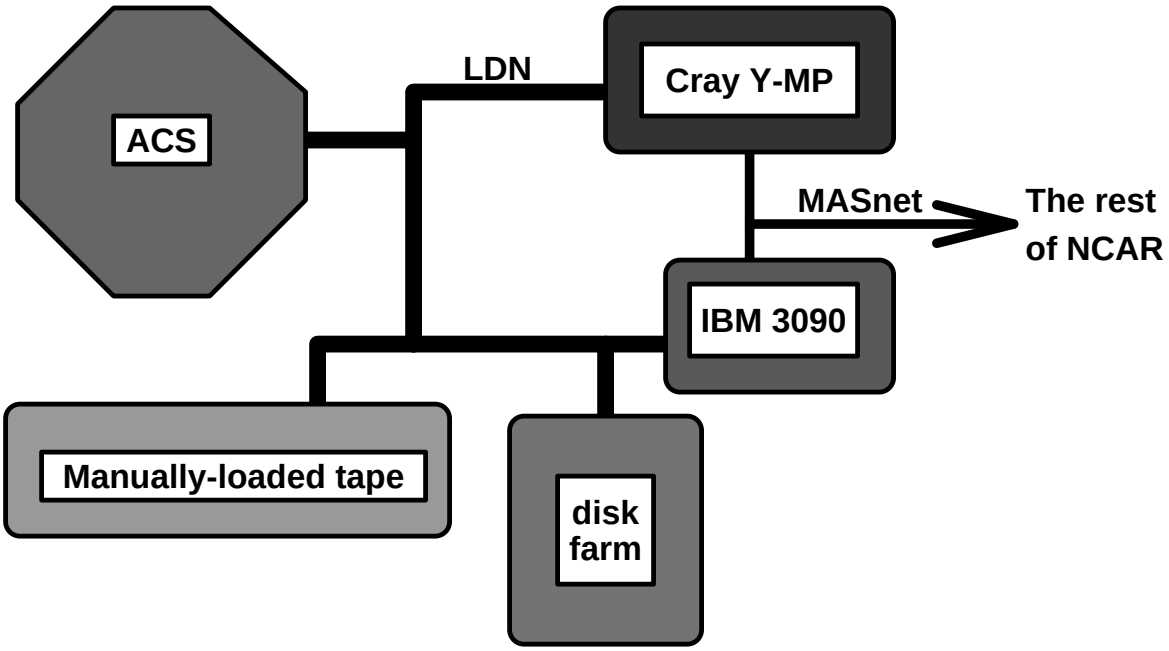

Figure 2: Network connections to the MSS at NCAR.

3.1 Hardware Configuration

The CPU in the study was a Cray Y-MP 8/864 (`shavano.ucar.edu`), with 8 CPUs and 64 MWords* of main memory. Each CPU has a 6 ns cycle time. `Shavano`, like other Cray Y-MPs, has several 100 MB/sec connections to its local disks and two 1 GB/sec connections to a solid state disk (SSD). There are about 56 GB of disks attached directly to the Cray; 47 GB of this space is reserved for application scratch space and files over a few days old are purged from it regularly.

The mass storage system (MSS) at NCAR is composed of an IBM 3090—used as a bitfile server—with 100 GB of online disk on IBM 3380s, a StorageTek 4400 Automated Cartridge System with 6000 200 MB IBM 3480-style cartridges, and approximately 15 TB of data in shelved tape. The MSS tries to keep all files under 30 MB on the 3090 disks, and immediately sends all files over 30 MB to tape. Usually, the tapes written are those in the cartridge silo. Files on the MSS are limited to 200 MB in length, since a file cannot span multiple tapes. While the Cray supports much larger files on its local disks, they must be broken up before they can be written to the MSS.

*Each Cray word is 8 bytes long.

The MSS at NCAR is shared by the entire NCAR computing environment, which includes the Cray Y-MP, an IBM 3090 which runs the MSS, several VAXen, and many Sun-type workstations. Figure 2 shows the network connections between the various machines at NCAR. The disks and tape drives attached to the MSS processor have direct connections to the Crays, providing a high-speed data path. All machines connected to the MSS (including the Crays) are connected to the 3090 by a hyperchannel-based network called the MASnet. Data going out over the MASnet must pass through the 3090's main memory, so it is a slower path than the direct connection the Crays have. All of the mainframes, and a few workstations also have connections to the MASnet, which is a custom high-speed network. The few workstations with connections act as gateways to the networks which connect the rest of the workstations at NCAR. These gateways are also usually the fileservers for the local networks. Many of these smaller machines have their own local lower-speed disks, about 5.5 GB of which are mounted by the Cray via NFS (Network File System). According to the monthly report published by the NCAR systems group [Wal91], `shavano` puts more data on the network than any other node, but several other nodes receive more data. In particular, several of the Sun workstations receive comparable amounts of data. It is likely that these workstations, which are the gateways to internal networks of desktop workstations, are receiving a large amount of image traffic.

3.2 System Software

The Cray Y-MP is primarily used for climate simulations— both the extensive number crunching necessary to generate the data, and the less computationally-intensive processing used in visualizing it. The Cray has two primary modes of operation; it can either run in primarily interactive mode, where programs are short and run as the user requests them, or in batch mode, where jobs are queued up and run when space and CPU time are available. There is no explicit switch between operating modes, but short interactive jobs typically have