Mass Storage at the PSC

Phil Andrews *Manager, Data Intensive Systems*

Pittsburgh Supercomputing Center, 4400 Fifth Ave, Pittsburgh Pa 15213, USA

EMail:andrews@psc.edu

Last modified: Mon May 12 18:03:43 EDT

ABSTRACT:

The Achival system at the Pittsburgh Supercomputing has been altered to use several hybrid systems, including IBM Magstar drives in an STK silo and DFS on top of the DMF file systems.

KEYWORDS:

Archival, File Systems

Other Personnel: Janet Brown, Susan Straub, Bruce Collier, Vidya Dinamani, Rob Pennington.

This talk is online at http://www.psc.edu/~andrews

Introduction

Until April '97 the Pittsburgh Supercomputing Center was one of 4 NSF-funded Supercomputing Centers. Unfortunately, in the PACI recompete it was not selected for continued funding.

The main machines are all Cray products; some of them the first of their kind.

Figure 1 is the Cray C90, 16 processors

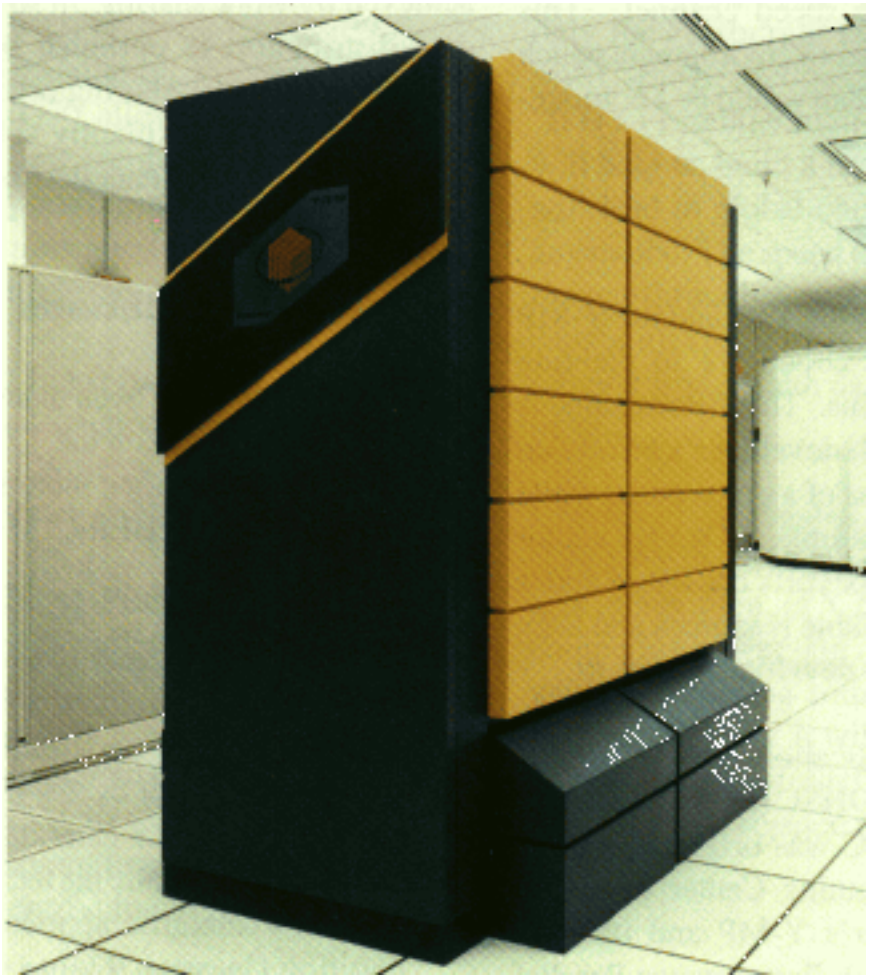


Figure 2 is the Cray T3D, 512 processors

Figure 3 is the Cray T3E, 512 processors

Figure 4 shows 3 Cray J90s with 8,8,10 processors

Archival needs: commonly run dedicated machine jobs, need to get large files in and out of storage quickly. Shelf operations unacceptable.

Overall strategy:

- Use Cray J90 (Golem) as file server
- Golem's rotating storage = disk cache
- Systems read/write from/to the Golem disks
- DMF manages cache misses, tape migration

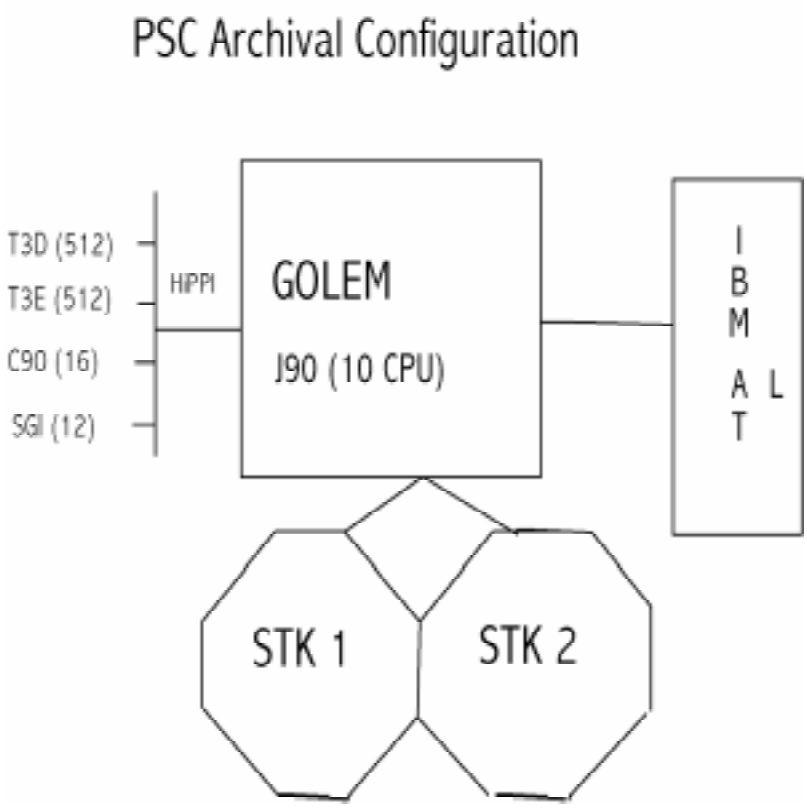


Figure 5 is a schematic of the file server layout

Hardware Configuration:

File Server/File Systems:

Golem has 10 CPUs, 11 IOSs available for Archive use. Low memory (64MW-512MB).

Eight identical user file systems because:

- Need to reduce no. of INodes per File system.
- Prepare for parallel tape I/O.

Each file system has:

- 1x9GB Disk primary (mirrored), INodes and small files. Optimized for safety.
- 4x9GB Disk secondary (4-way striped), larger files. Optimized for speed.

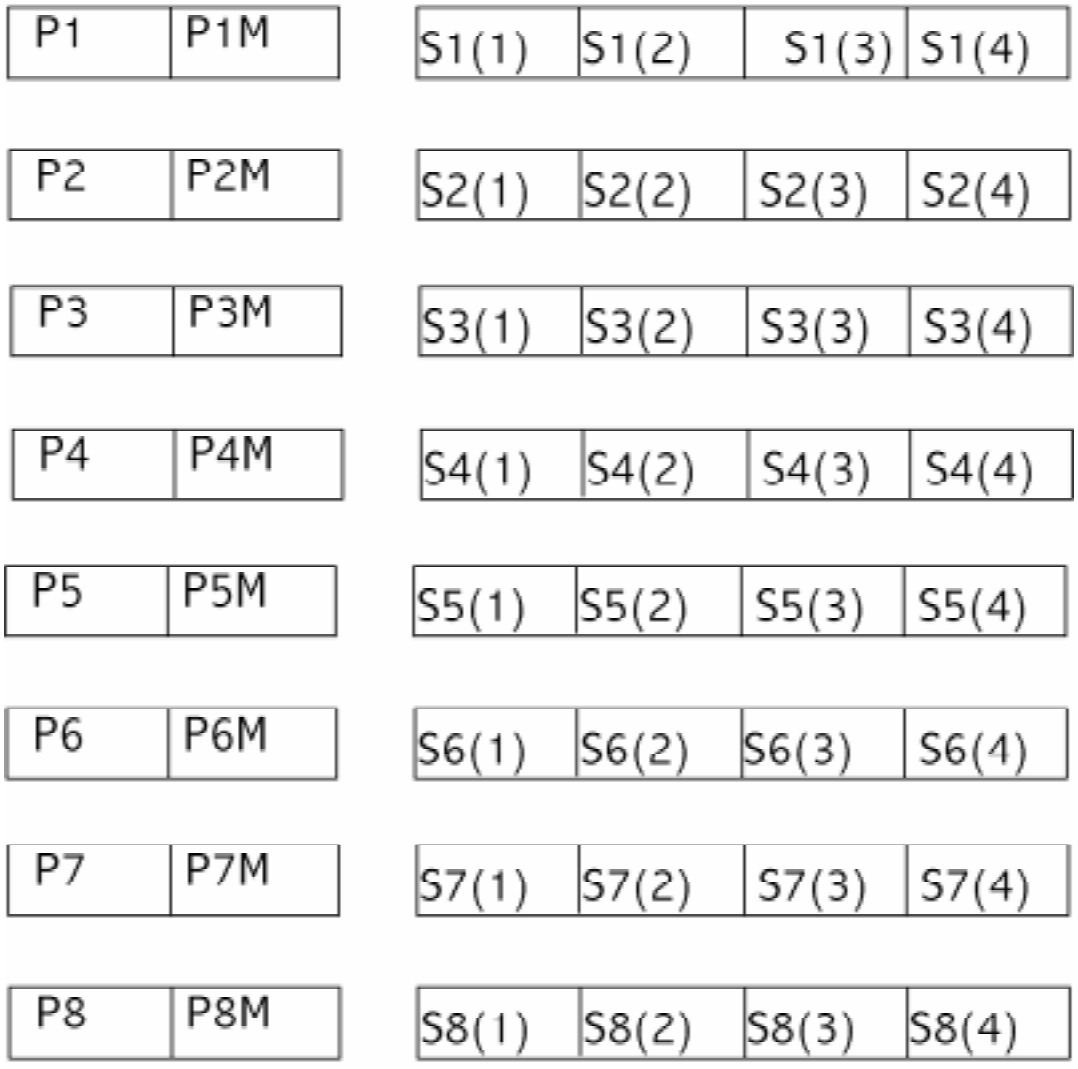


Figure 6 is the eight-way file system layout

Each pair of file systems gets an MSP, each MSP can grab 2 tape drives

Tape Systems:

Timings:

- Disks: 6.7 MB/s each
- SCSI Chain: 16 MB/s each
- IOS: 35MB/s each

Secondaries run with 2 Disks per SCSI chain, 2SCSI chains per SI3 adapter, 2 SI3 adapters per IOS. File systems deliver 26 MB/s.

Tape Systems:

IBM ATL:

7 Cabinets, 8 IBM Magstar drives, space for 2400 tapes. Run one drive per SCSI chain.

Timings, etc,:

- Drives rated at ~10 MB/s
- See 10-12 MB/s
- Average compression ratio: 1.4
- Tape capacity 10 GB (uncompressed)

STK Silos:

The two STK silos in the PSC machine room at Westinghouse.

Two connected silos, 6,000 slots each

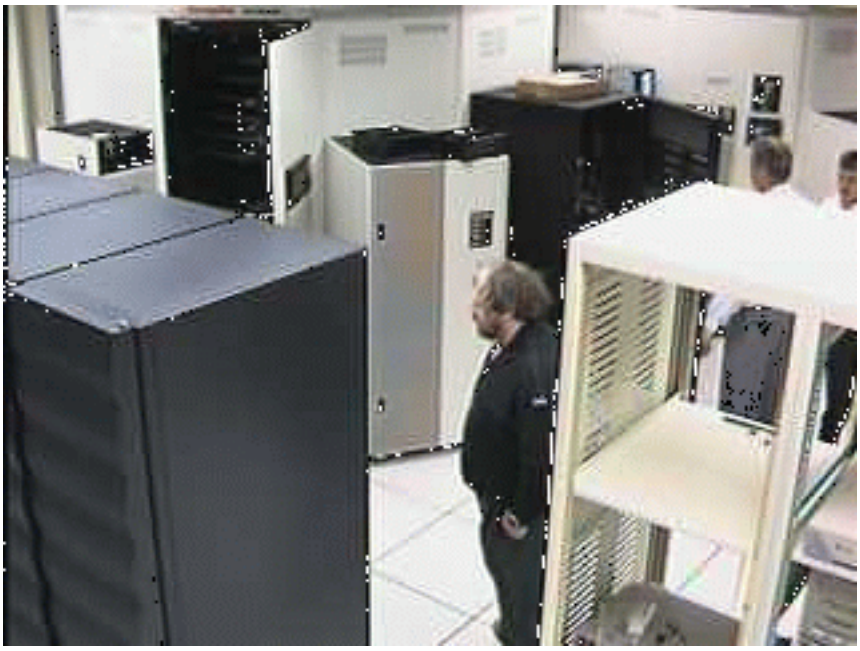
Tape drives:

- Originally: 4X4 STK drives
- Now: 2X4 STK drives, 1X4 IBM Magstar drives.

Running mixed media, both older STK tapes and IBM 3590 tapes in silos. Plus shelf operation for STK tapes.

Conversion is via IBM C-12 frame, bolts to STK Silo, holds 4 IBM Magstar drives. No problems so far.

An MPEG animation of the actual drive frame installation.

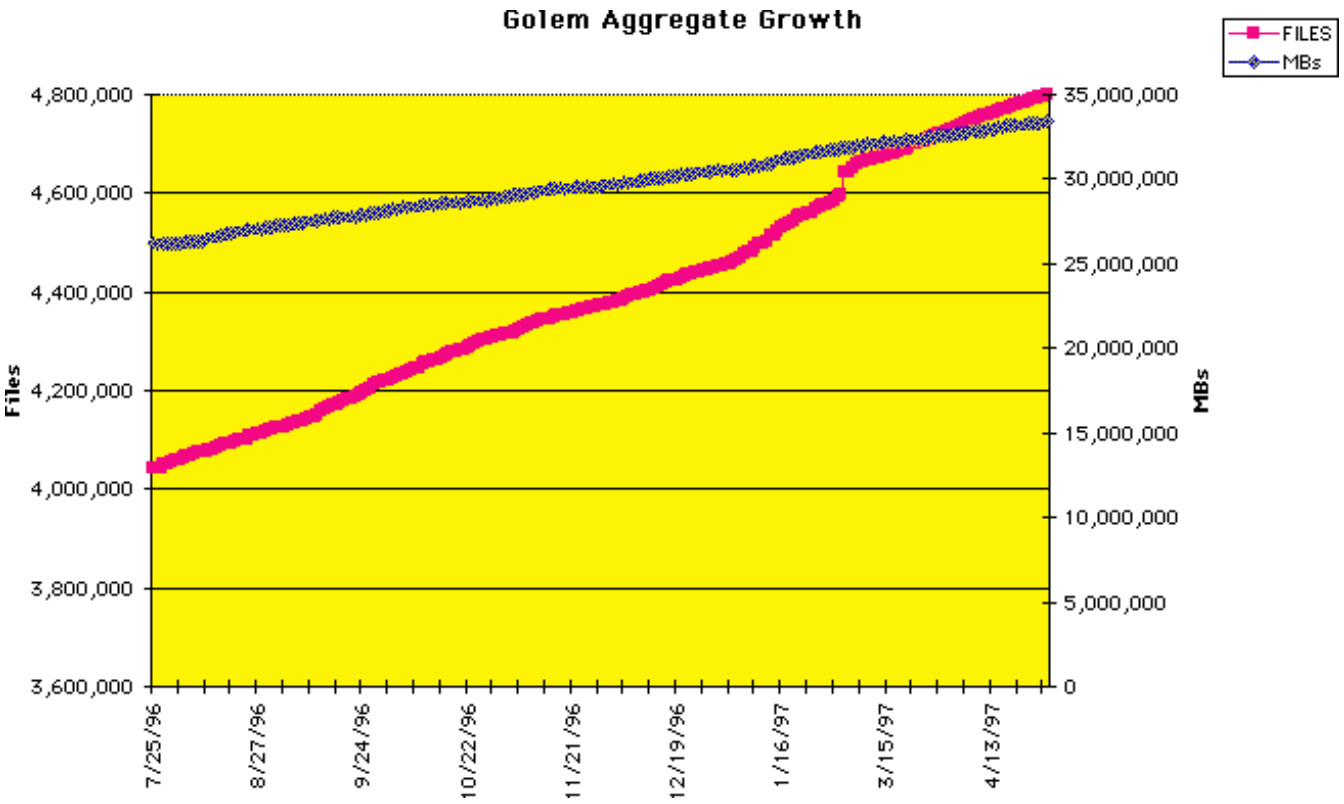


View after the drive frame installation.

Total automated tape capacity now 14,400 tapes, approximately 200 TB. Double length tapes (later this year) would bring it to ~400 TB.

Archive growth

The new (Golem-based) archive now now totals approximately 34TB, growing at about 2TB/month. There is an almost identical amount of data in the pre-1996 archive.



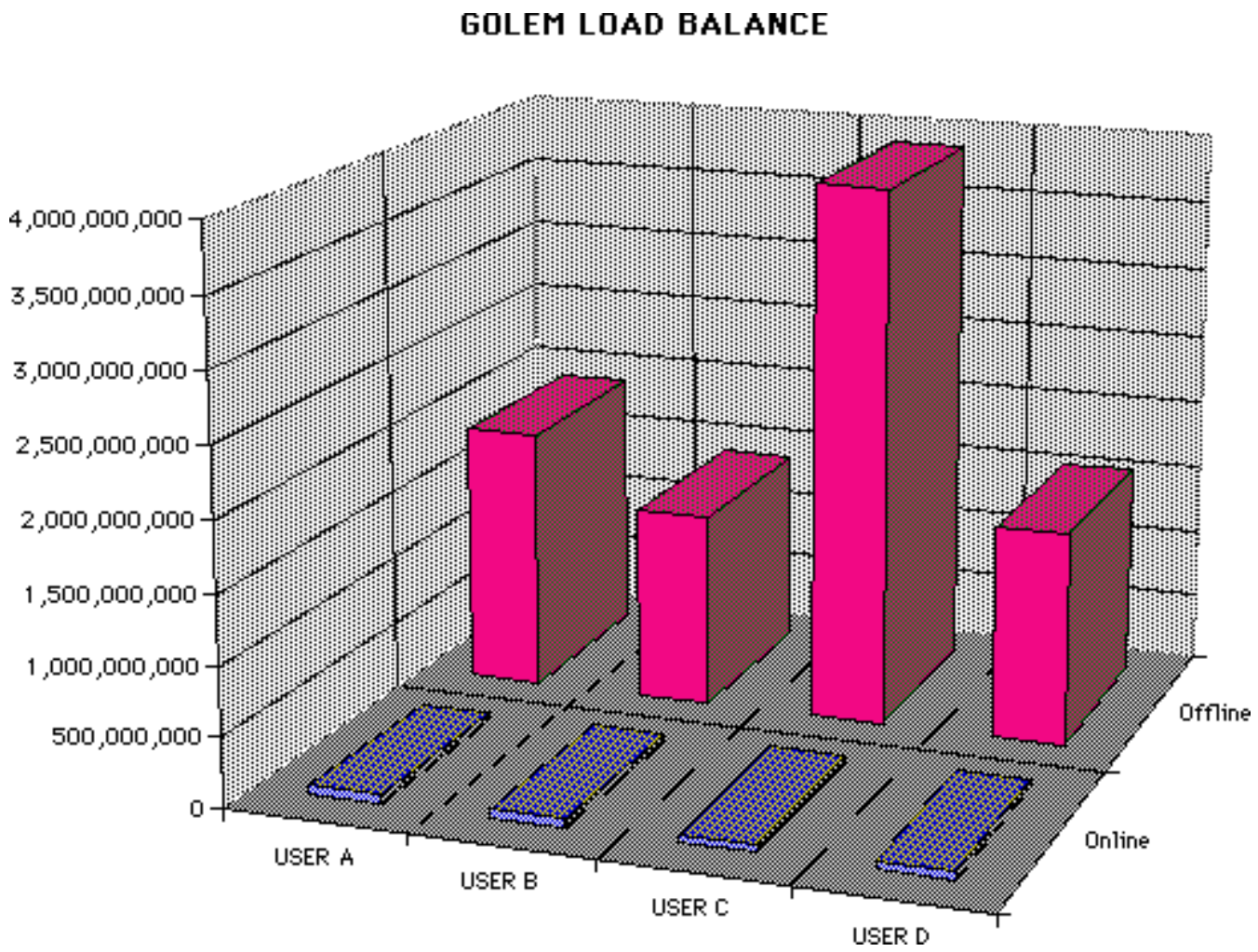
Growth rate of the new archive in total data and files.

User access

Initial user access was by FAR (home grown archiving software) and FTP to Golem file systems.

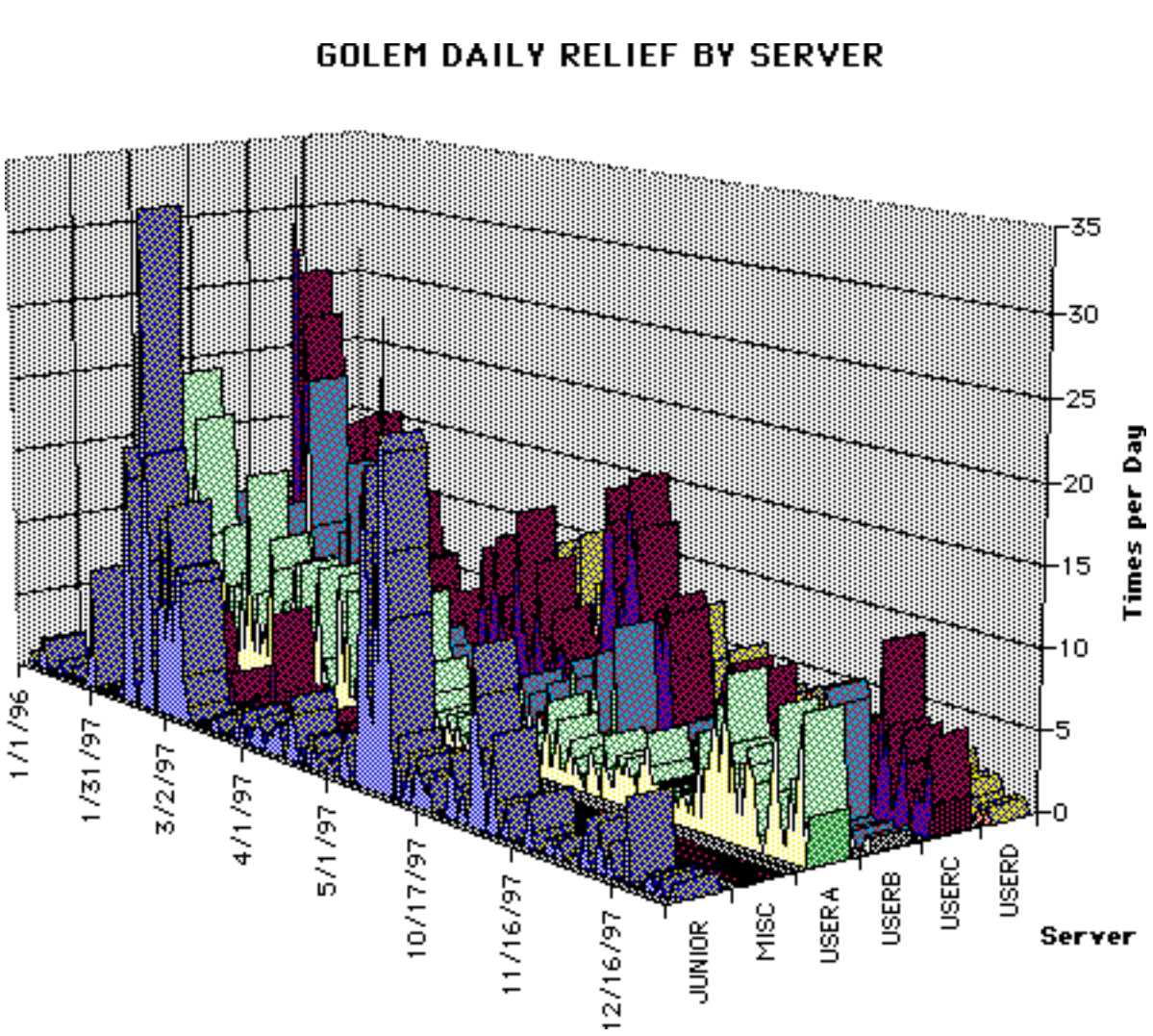
Archival measures:

Each pair of file systems gets a Media Specific Process; DMF counts files and data on an MSP basis.



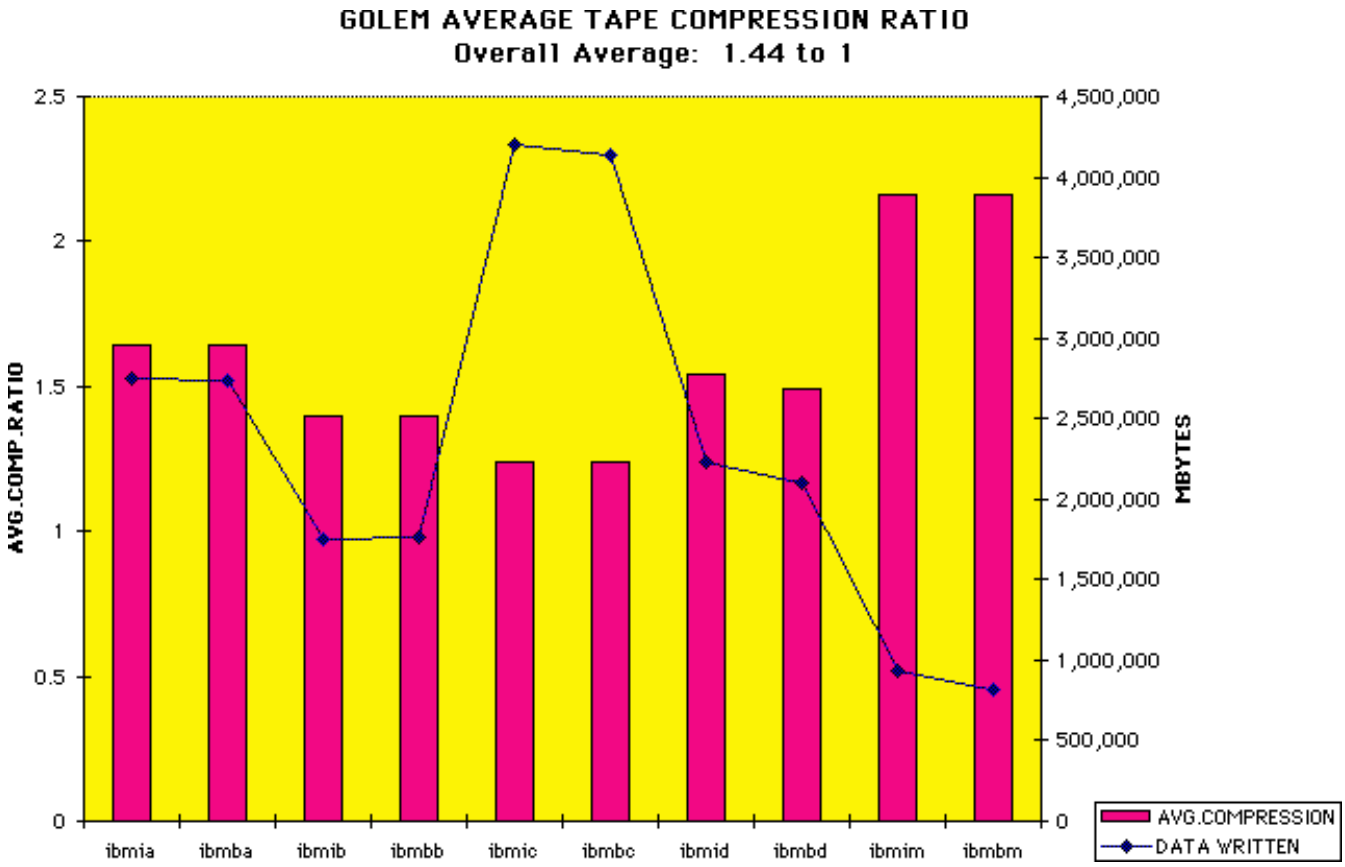
Load Balance, bytes stored by MSP

Whenever the disk occupancy on an MSP reaches a "warning threshold", DMF performs a "relief" operation; deleting premigrated files and premigrating more files to tape.



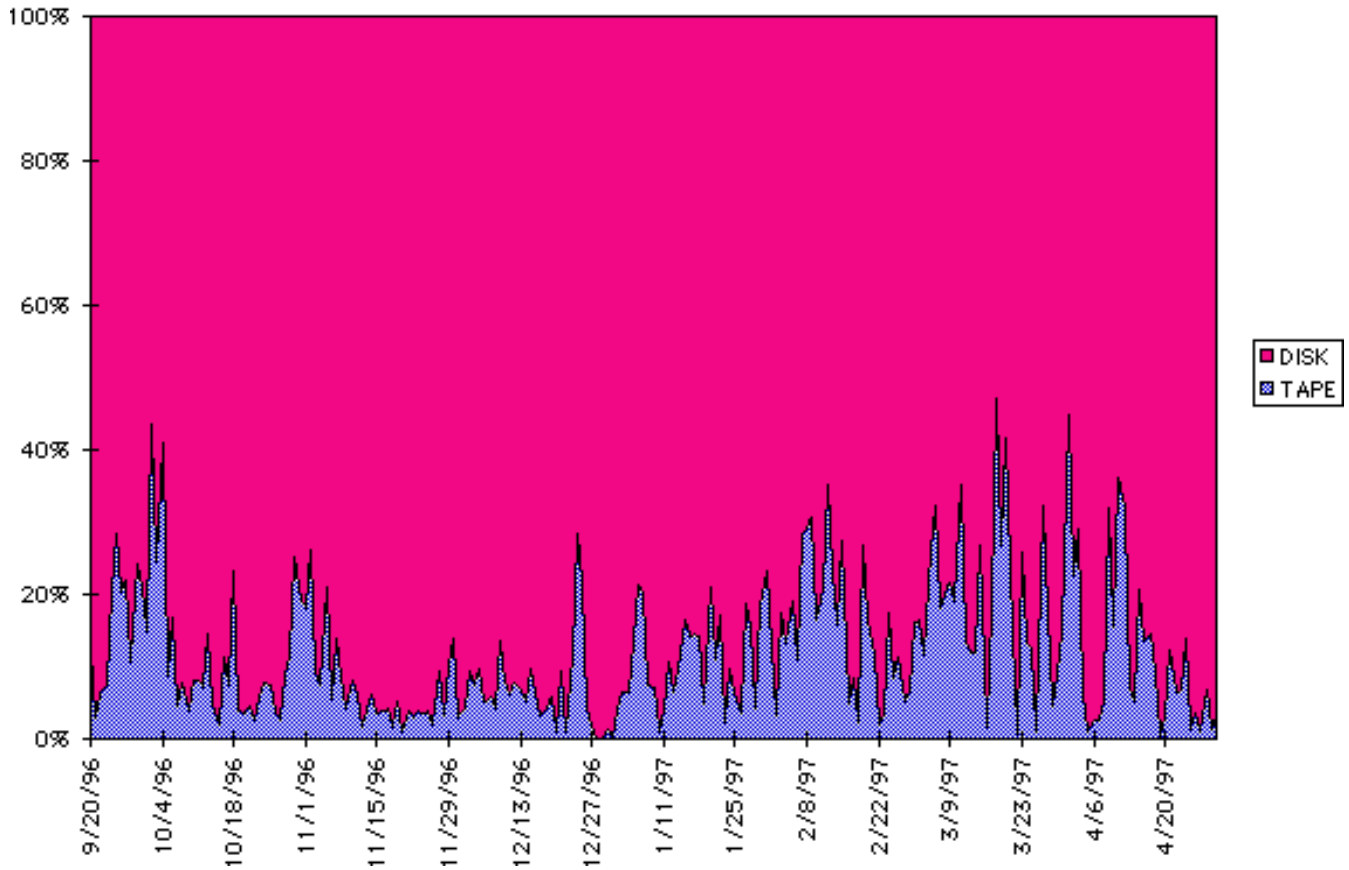
Relief from warning threshold

The IBM drives perform hardware compression; we track this on an MSP basis:

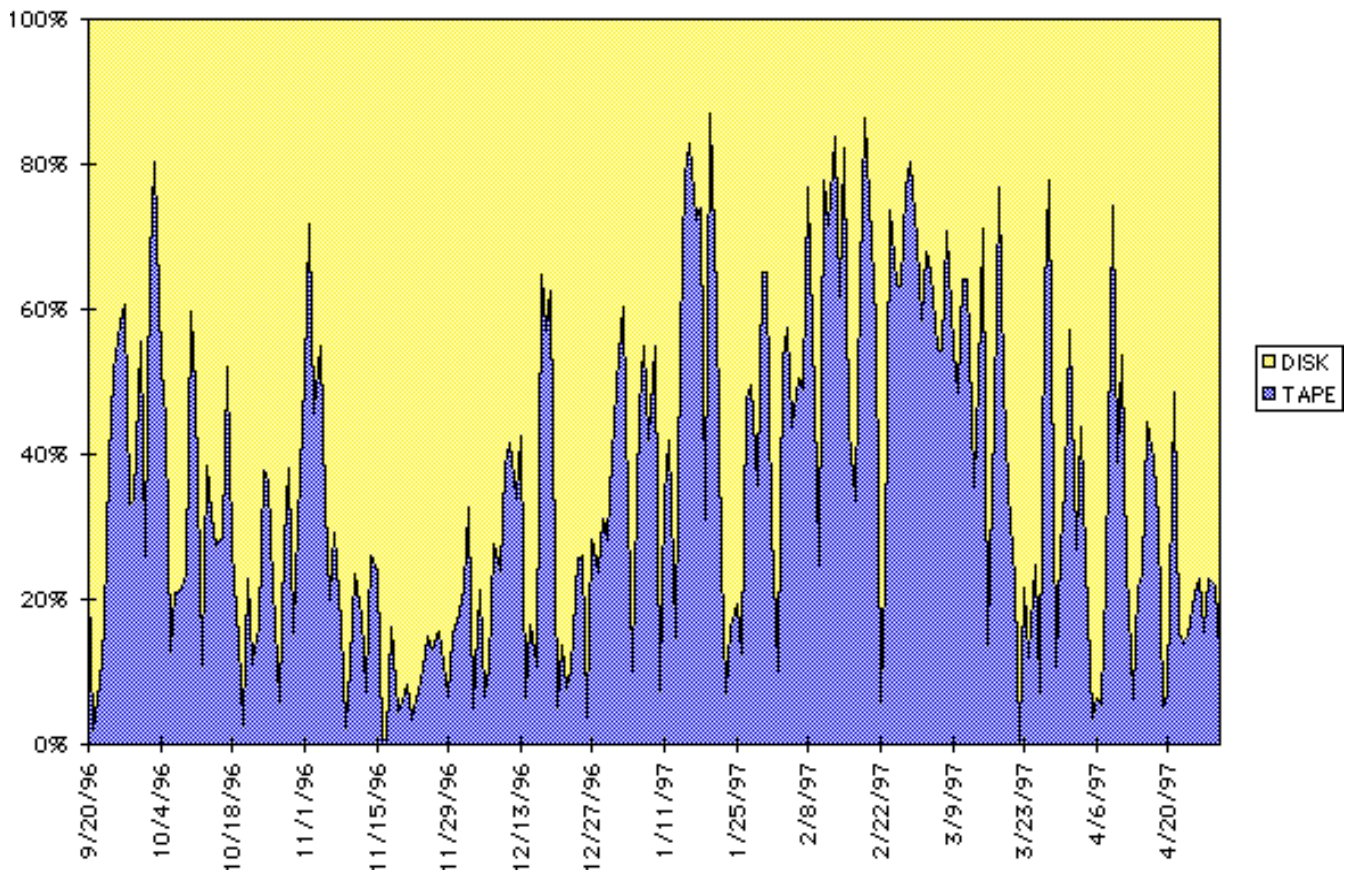


Compression ratio by MSP

We attempt to keep the "hit" ratio on file requests as high as possible; the following figures measure it on a daily basis as a function of both file number and data.

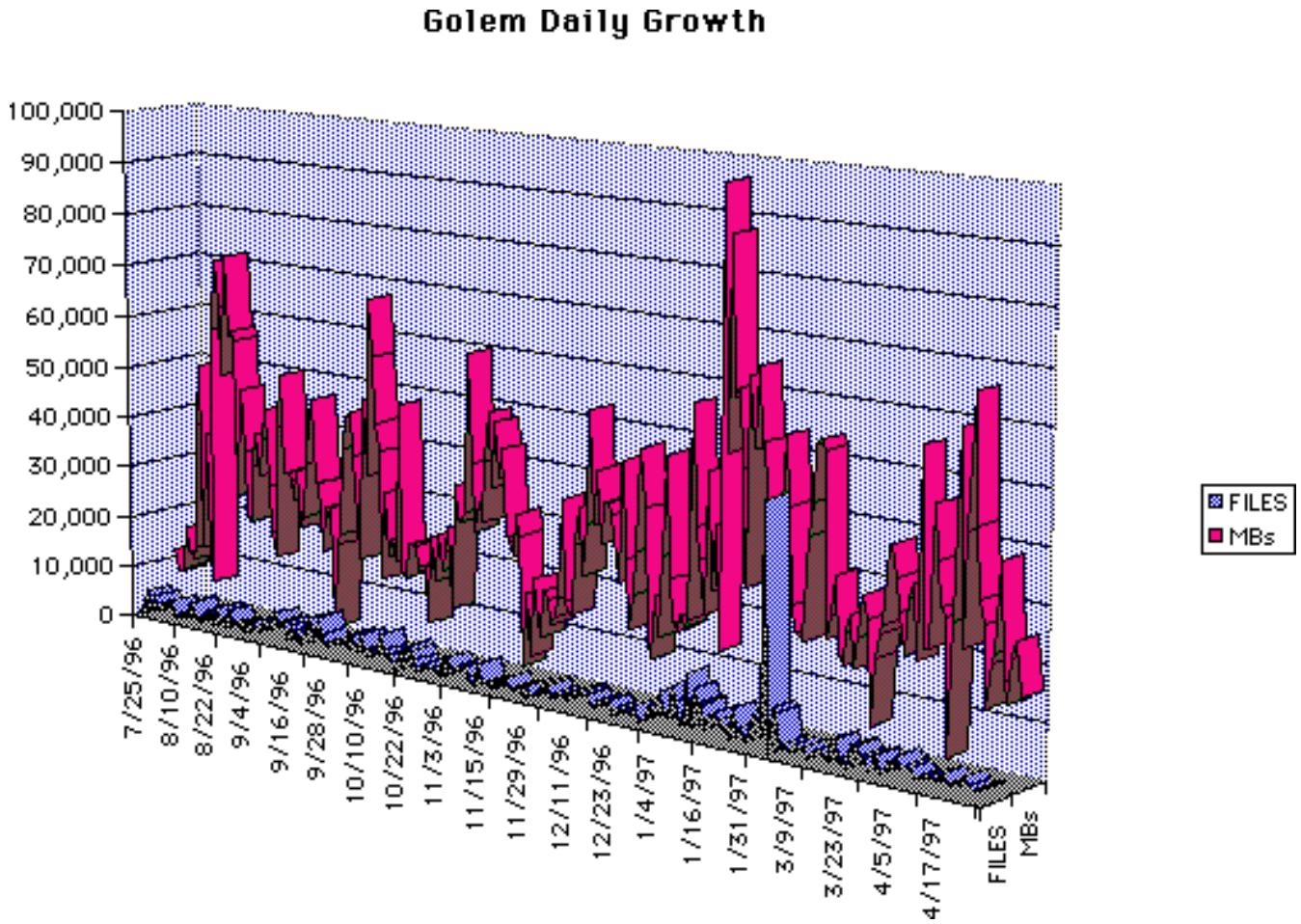


"Get" Hit Rate by files



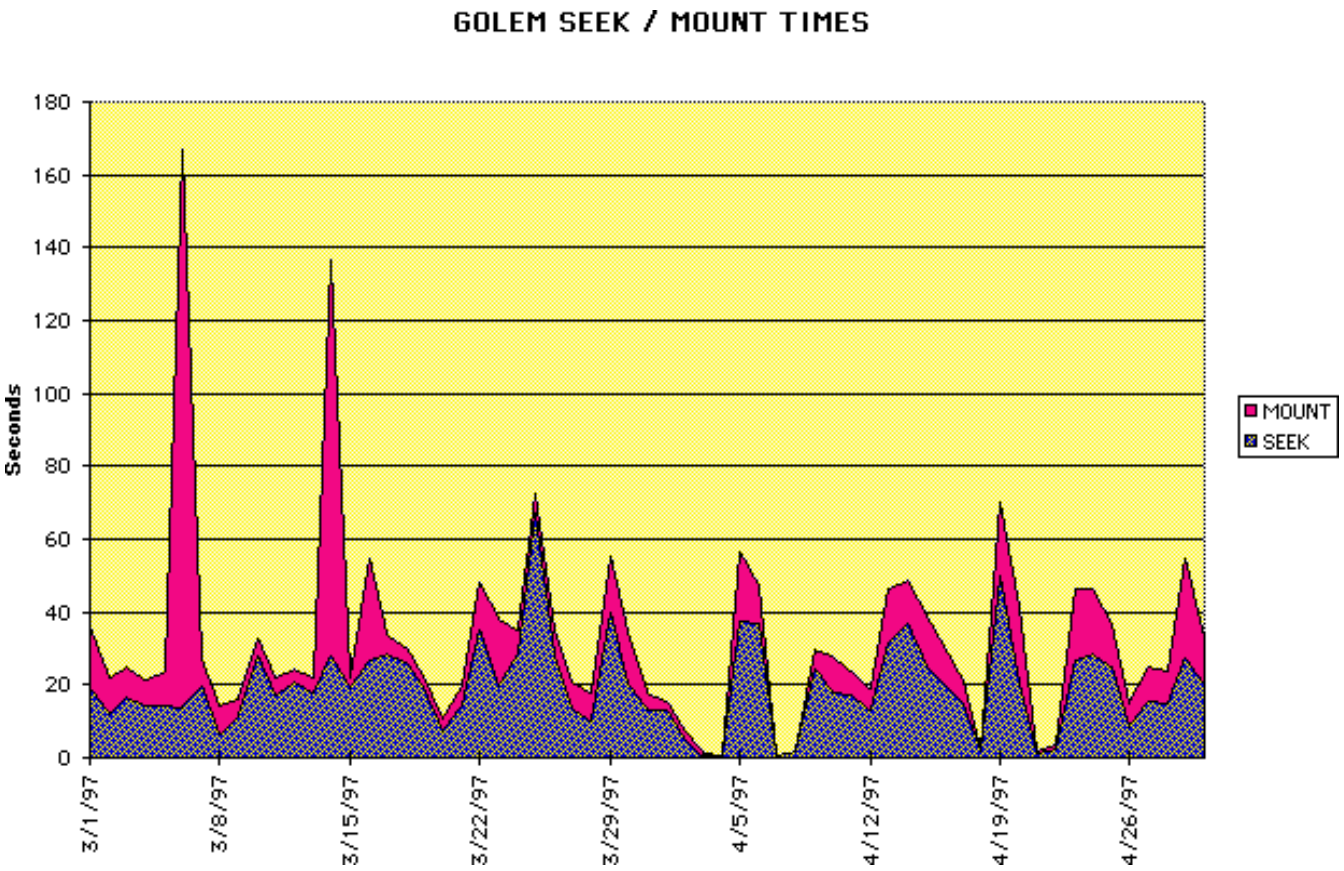
"Get" Hit Rate by data

The daily growth rate is spiky, but without long-term trends.



Daily growth rate, files & data

We graph both the seek times (time to find tape and insert into drive), and mount times (time to start reading data)

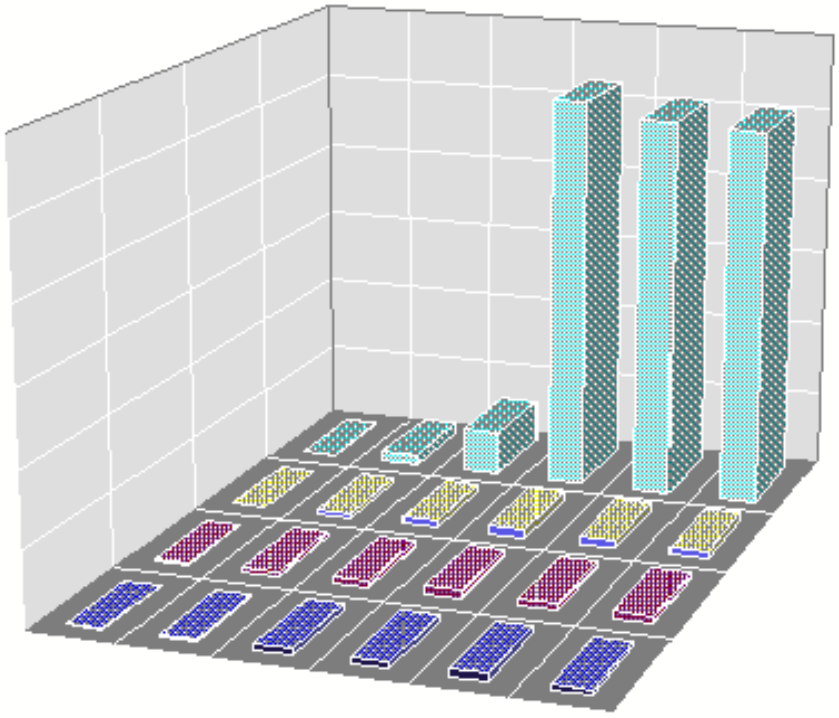


Seek and mount times

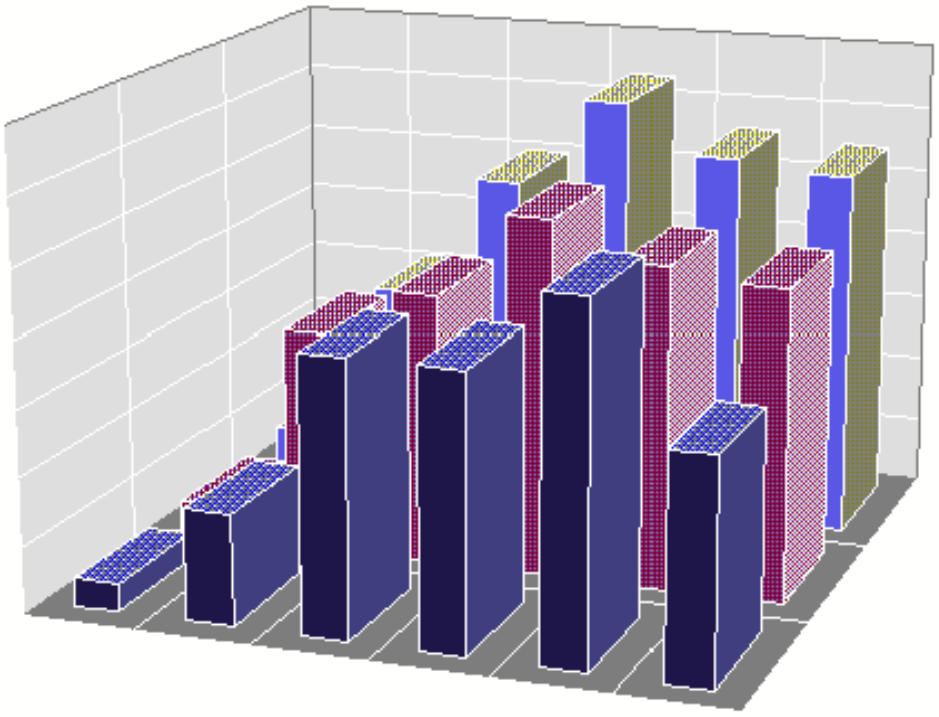
Distributed File System

Now running DFS server on Golem, exporting all 8 Golem user file systems. Clients running on SGI, Sun, J90, IBM/AIX. Windows NT, Apple Macintosh, C90 coming. Early days yet.

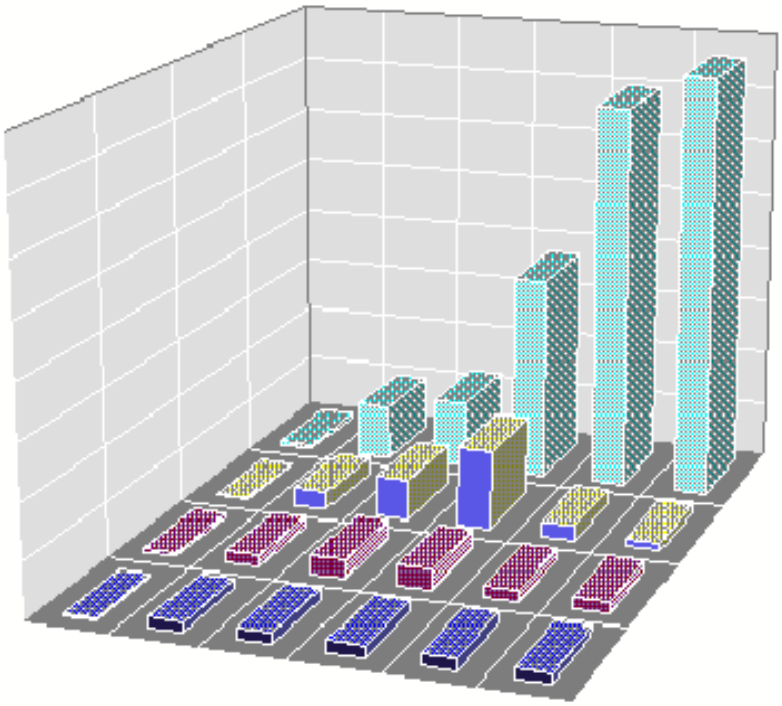
DFS performance, all operations to/from the archival file server (Golem). We present raw data, we have not yet had time for evaluation.



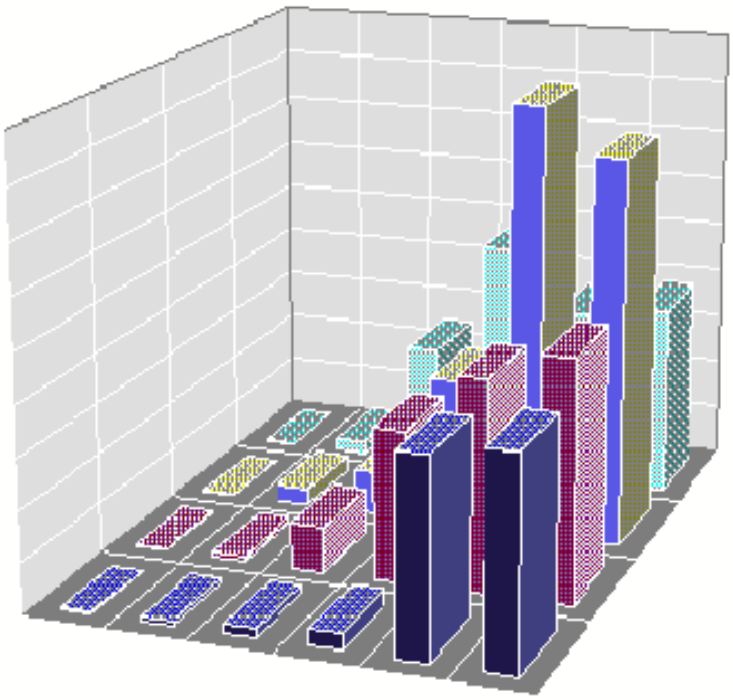
Client Write Comparison



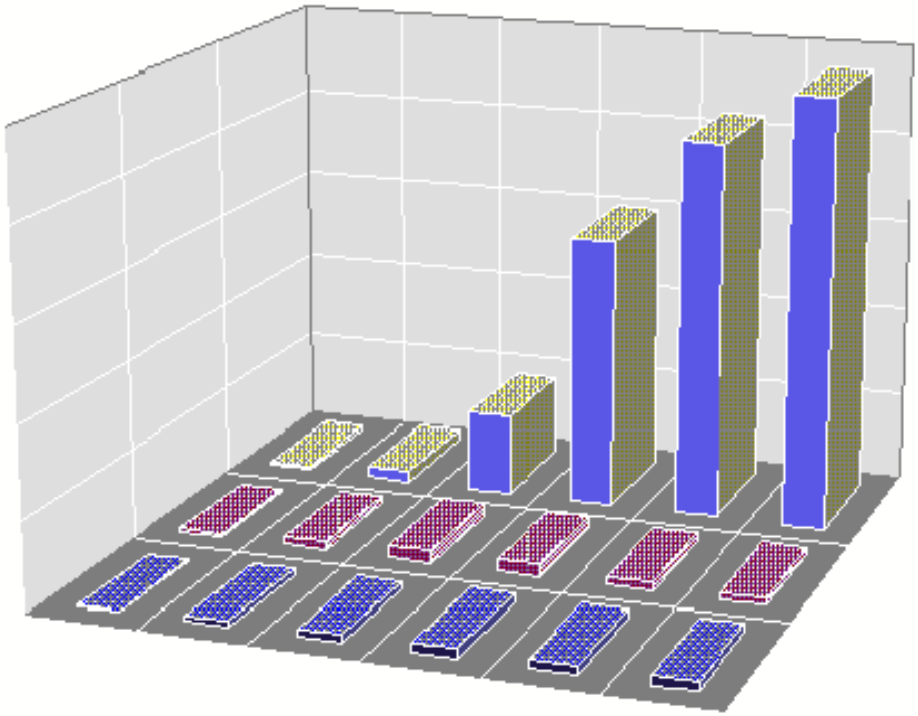
Client Write Comparison(2)



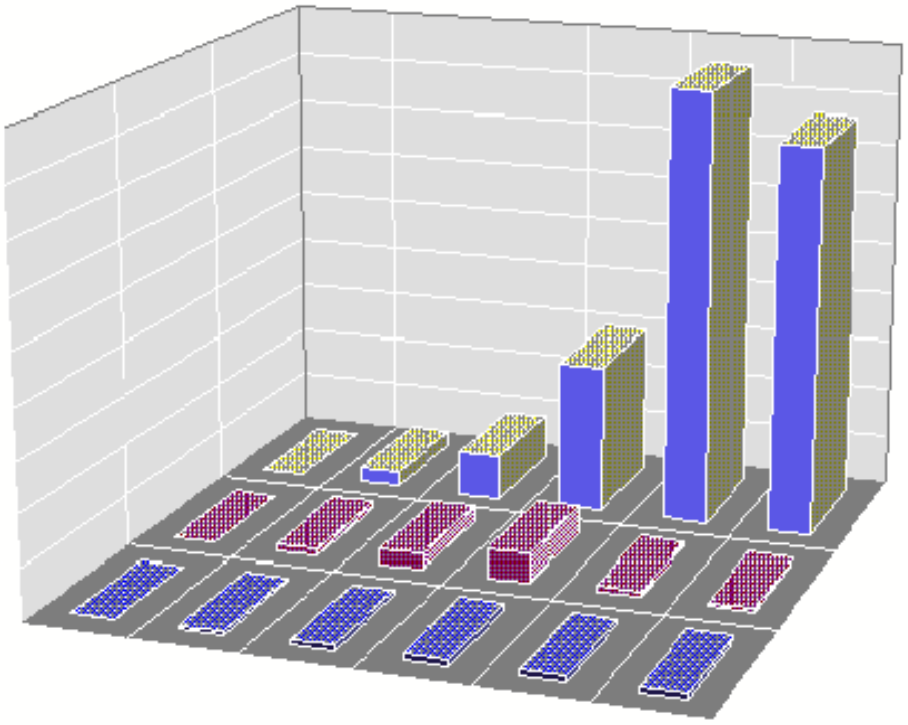
Client Full Read (Complete File) Comparison



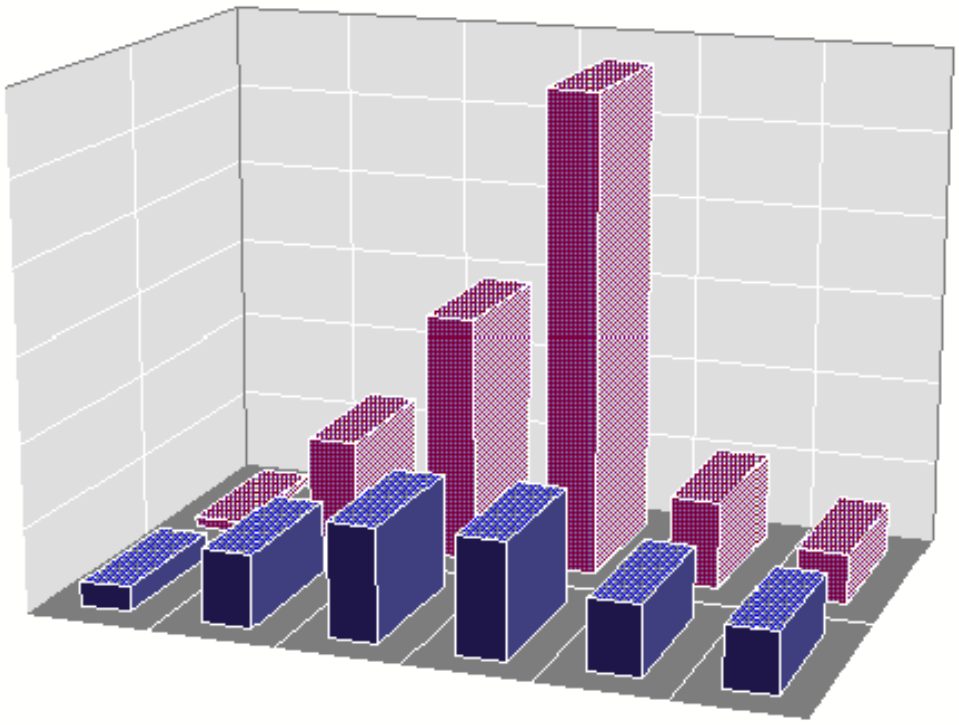
Client Partial Read (40 bytes from file center) Comparison<.h5>



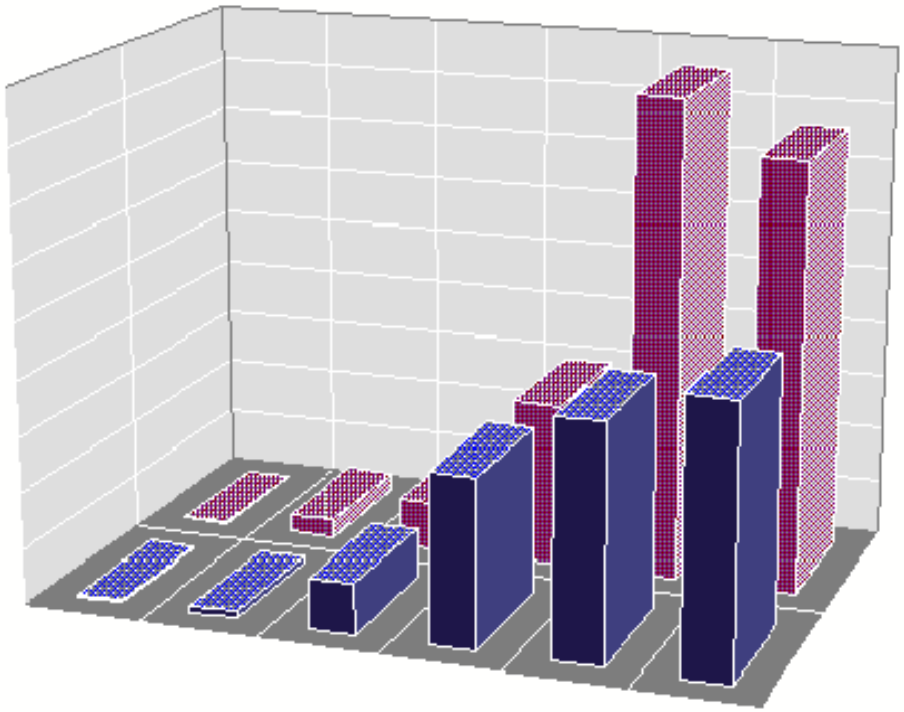
SGI Client R/W 10MB Cache



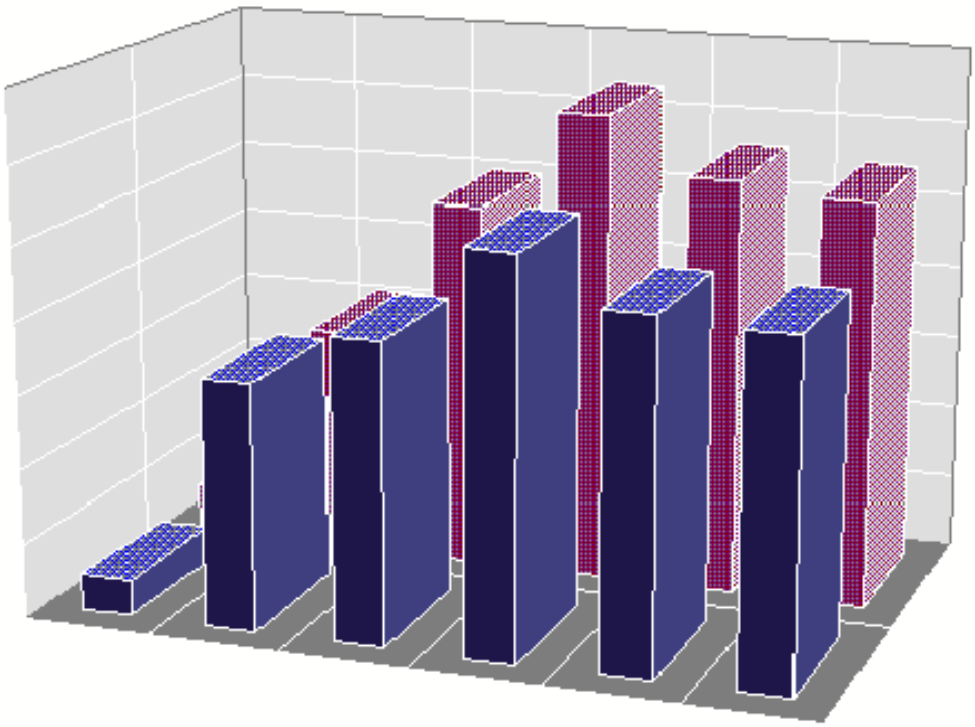
SGI Client R/W 100MB Cache



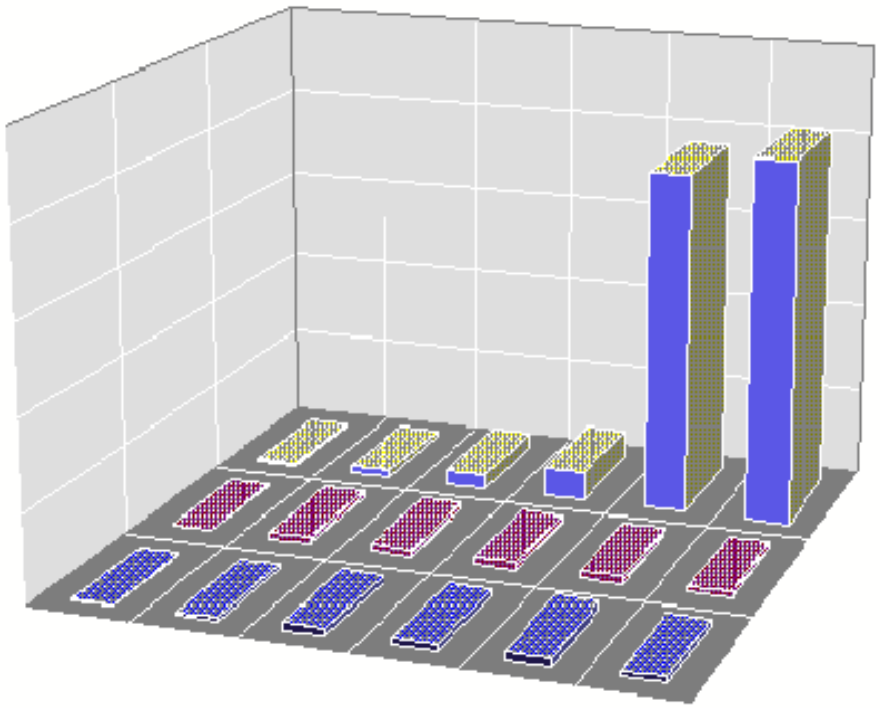
SGI Client Full Read Cache Comparison



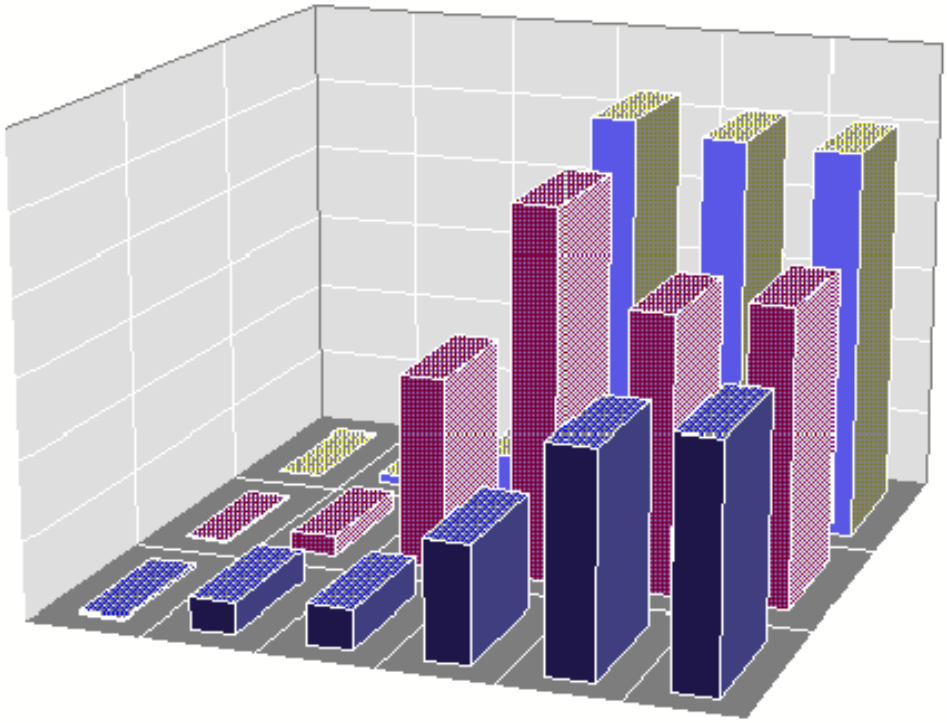
SGI Client Partial Read Cache Comparison



SGI Client Write Cache Comparison



Sun Client R/W 40MB Cache



J90 Client R/W 64MB cache

DMAPI Application on SGI XFS

We have used the DMAPI (Data Management API) implementation available with the XFS filesystem under IRIX 6.2 to create a user level DMAPI application that successfully managed a large striped filesystem on an SGI Power Challenge machine. The goal of this work was to build a prototype system that was capable of providing a high-performance storage management system for the SGI based on currently available standard technologies that provided transparent access to the file archiving system resident on the CRI J90 system at the PSC, the File Archiver. This is a follow-on project to the Multi-Resident AFS work at the PSC and was intended to examine the DMAPI technology that is beginning to become available from several of the vendors (HP/Convex and SGI, in particular).

The software prototype had to meet the following requirements:

- Behave like a normal filesystem. Users should be able to perform all normal file operations on files and directories on the managed partition with existing, unmodified tools.
- Be able to manage at least two residencies for the files, one on the primary storage server and one (or more) on a back-end machine or tape drive.
- Exhibit reasonable performance characteristics. Discounting the time necessary to retrieve the file from tape, the amount of overhead incurred by the use of the DMAPI application must remain small.
- Does not require source code changes to the OS or the filesystems on the machine on which it is running.
- Provide similar namespaces on the machines involved to enable access to the managed files through either the DMAPI managed filesystem on the SGI or directly on the FAR system. Permissions and ownership on the two systems must be consistent.
- The DMAPI prototype system on the SGI cannot negatively impact the production archiving system running on the CRI machine.

Performance tests were performed early on to ensure that the overhead of using the DMAPI was acceptable. Of particular concern was the rate at which the filesystem could be scanned to locate candidates for migration. If this rate

is too low, then it would not be possible to keep up with a reasonable rate of creation of new files on the filesystem. The measured rate of roughly 13,000 files/second on the SGI 4xR1000 CPU Power Challenge machine was deemed to be sufficiently fast for this work.

The prototype software system consisted of four daemons on the SGI, a parent to control three child processes. The children are: an archiver to move files between the data stores, a purge daemon to maintain the amount of freespace by clearing dual-state files and the event loop process which responds to operations on the files (read/write/truncate/delete

We have successfully managed files on a large striped logical volume, maintaining specified level of free space using DMAPI application to handle file events, moving and retrieving files from backend storage on a local tape drive and across HiPPI to one or more CRI systems as archiving back-ends, including the File Archiving machine.