Mass Storage at the Pittsburgh Supercomputing Center

Phil Andrews, Pittsburgh Supercomputing Center

In production now for one month on a J90 based archival system

Plan of attack:

1) Ensure good performance and reliability from present system

2) Plan for significant expansion

3) Develop/Install a seamless user interface

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

Previous Archival Hardware

File Server:

Cray YMP-EL. Cray's DMF (Data Migration Facility) software is used to manage files.

Rotating Storage

Four User File systems

The disks hold all large files until a minimum empty space requirement is reached, then they are written to tape.

Tape Storage

Two StorageTek Silos with robots and pass-thru port. 6,000 tapes each capacity. Additional shelf capacity

Current Archival Hardware

File Server:

Cray J90, 10 CPUs, 11 IOSs available for Archive use, 32 MW (256MB) memory. Cray's DMF (Data Migration Facility) software is used to manage files.

Rotating Storage

Eight User File systems; each with a single mirrored primary disk (9GB) and a 4-way striped secondary (4 X 9GB)

The primary disk stores all INodes for the file system and small files below a certain size threshold. With mirroring, it can be immediately replaced in the case of disk failure.

The secondary disks hold all large files until a minimum empty space requirement is reached, then they are written to tape.

Tape Storage

IBM 3494 Automated Tape Library with 8 IBM 3590 Magstar drives, one robot and space for approx 2,400 tapes. A linear system, quite different from STK Silos

Nominal tape capacity is 10GB (after compression).We are averaging a compression ratio for 2.3

Presently have 2,000 tapes

Design Motivation

Main machine will be 512 processor (presently 256) T3E with >1TB of local fiber-channel disk. *Must* be able to move large files in and out quickly as we anticipate large number of dedicated machine jobs. Performance is vital.

Must also be able to deal with large archives of images, etc., that must be available to WWW access

Currently have 2 HPPI interfaces (2 X 40 MB/s), plan to improve in future when new technology comes available.

Performance

From Disk:

Nominal single disk performance is 6.7 MB/s. We see up to 16MB/s on a single SCSI chain, so we chain disks two deep.

We see 34.8 MB/s per IOS and run two file system secondaries per IOS with 4-way striping.

Peak performance (> 10MB file) is 26 MB/s, with typical performance 20-24 MB/s

From Tape:

Nominal tape performance is 9 MB/s, so running only one deep SCSI chain, 4 drives per IOS.

Actually see 9-12 MB/s, presumably because of hardware compression.

Usage

Current FIle access is either via FAR, a home grown interactive system based on ACP (a locally modified version of RCP) or via FTP.

Growth in first month of operation was 2TB

Whole of original archive became the "dead" file system, about 15 TB. Read Only system not optimized for access (no disk striping).

Rapid access to tape archive files has lead to more "casual" archiving of large (>200MB) files.

Largest "real" files are about 6-12 GB

Future Expansion

IBM tape improvements

(see your local IBM rep)

Parallel Tape I/O:

One reason to have 8 user file systems is to allow tape striping with up to 8 tape drives, plan to experiment with this for rapid support of dedicated T3E jobs (not for routine use).

Adapting existing STK silos:

Plan to be beta test site for insertion of IBM MagStar drives into our STK silo

"Seamless Access"

With rapid access even to tape-resident files, we would like to provide transparent access to all files without losing performance. We are experimenting with two different approches.

DFS:

The Distributed File System is part of the OSF Distributed Computing Environment (DCE) and is the successor to AFS.

We are running DFS in an experimental mode on top of a DMF file system on one of our other J90s and during test time on our J90 file server. A DFS client on other machines then sees the whole archival system as part of the local file system. The client can even run diskless.

One advantage of DFS is that it is widely (but not universally) available and has extensive security measures as part of DEC.

DFS makes it very easy to setup a WWW server that has instant access to the archive for, e.g., the stroage and provision of very large and numerous image files.

DMAPI:

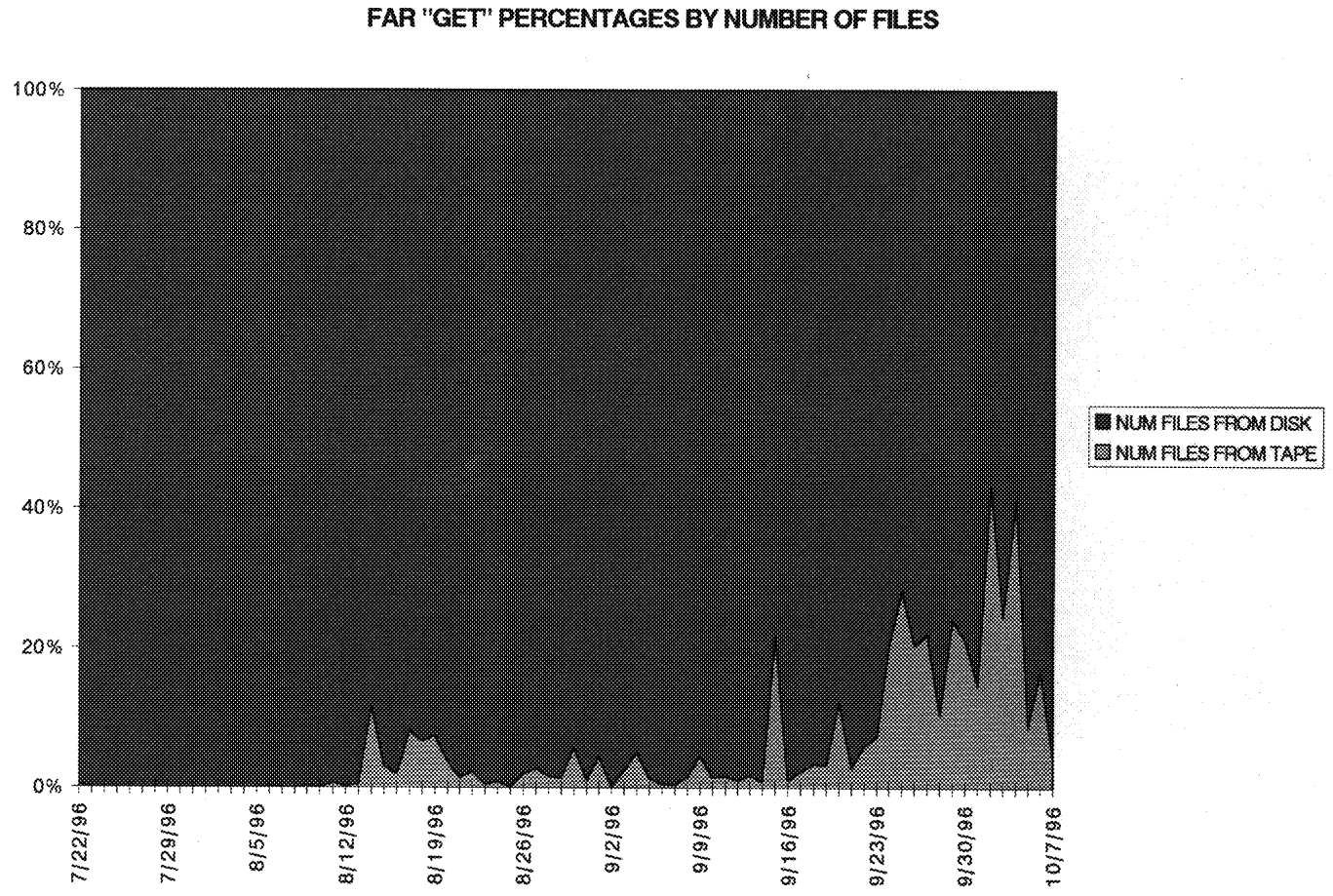
The Data Migration Application Programming Interface is a proposed standard to allow user-level code (no kernel intrusion required) to satisfy file access requests. We are using the DMAPI implementation available with the SGI IRIX 6.2 release.

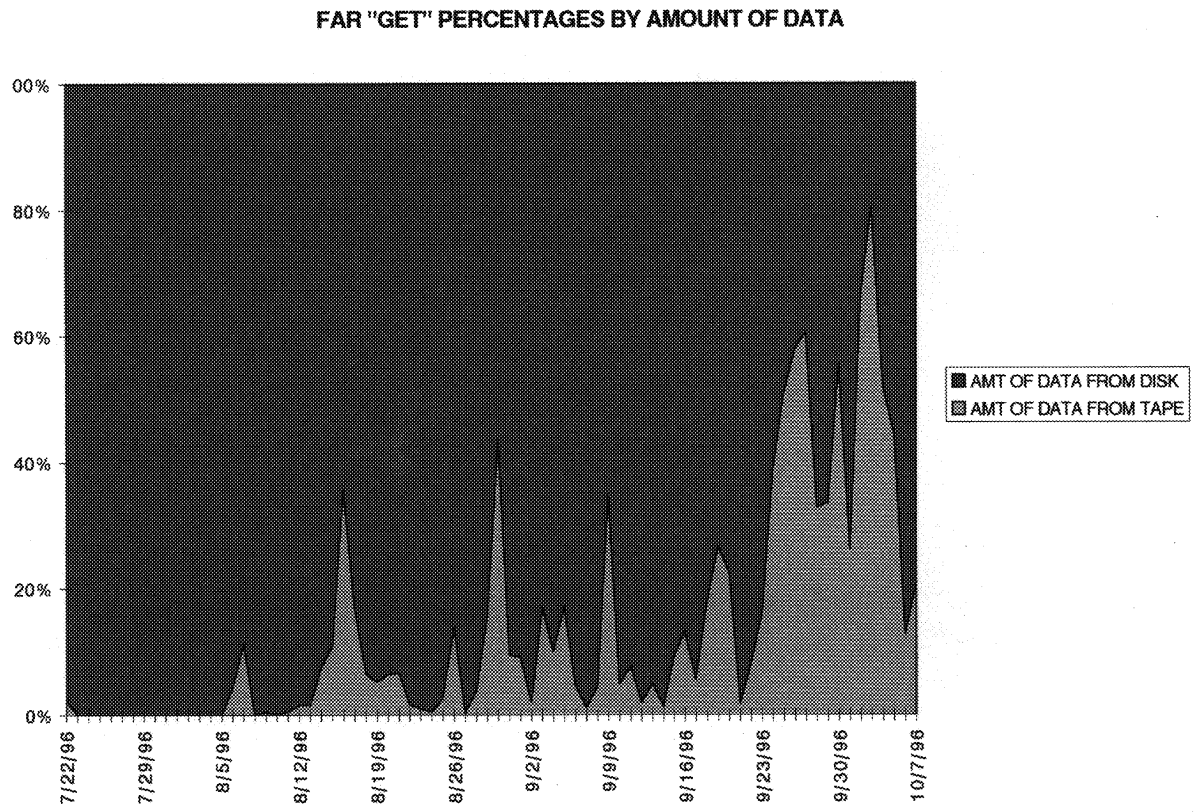
Using the DMAPI it is possible to determine your own caching, etc., rules and take advantage of any optimization possibilities that are locally available.

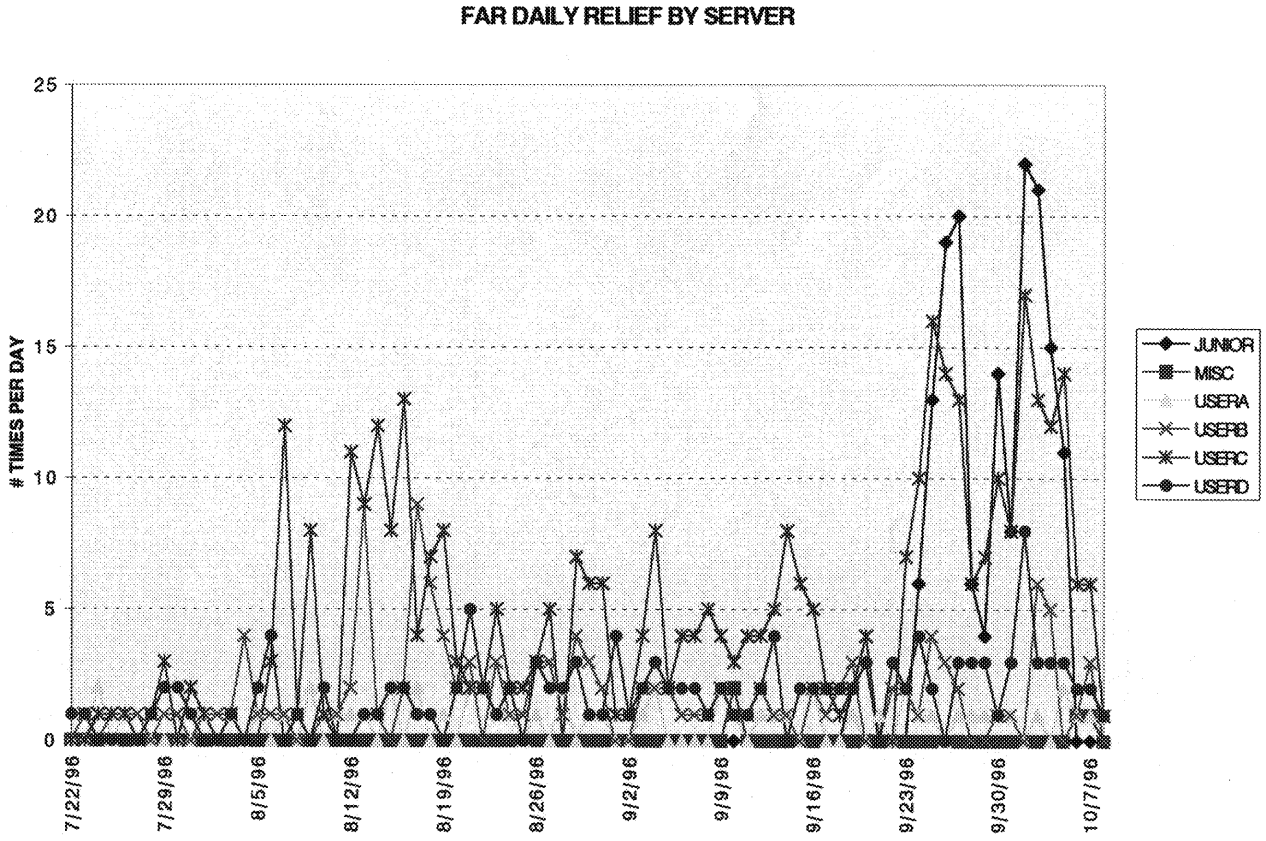
At the PSC we have created DM Apps that allow us to experiment with optimization techniques in accessing the archive.

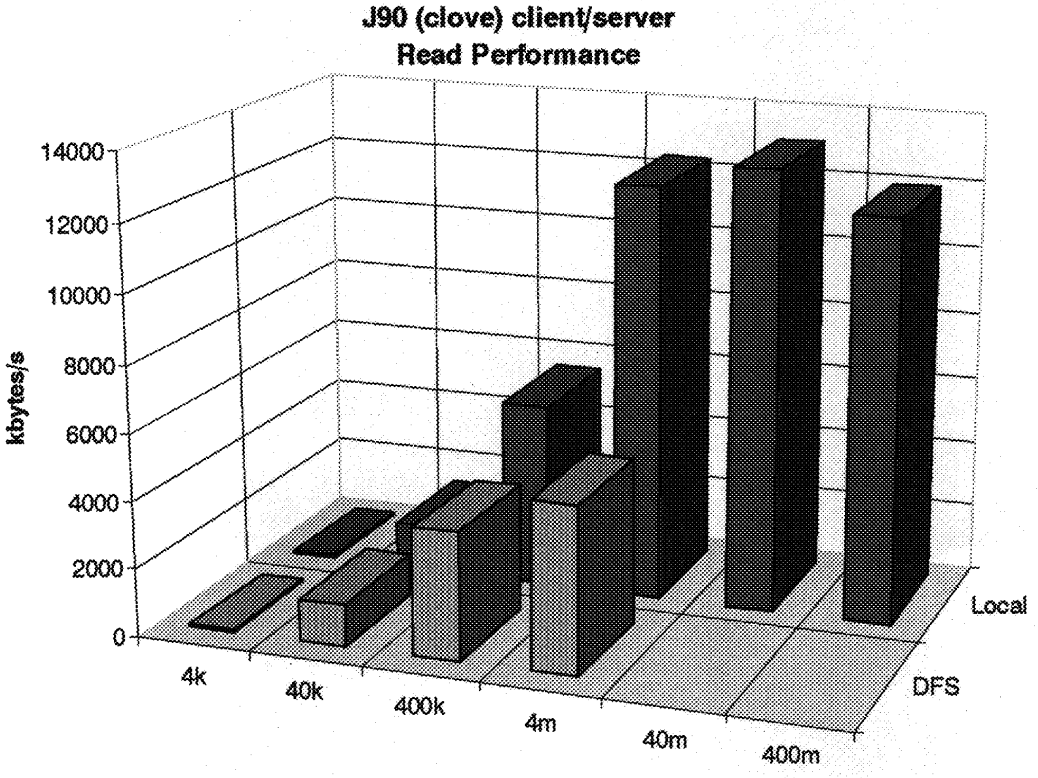
Preliminary results see Ethernet accesses at 850 KB/s and FDDI (T3 limited) at 5.2 MB/s

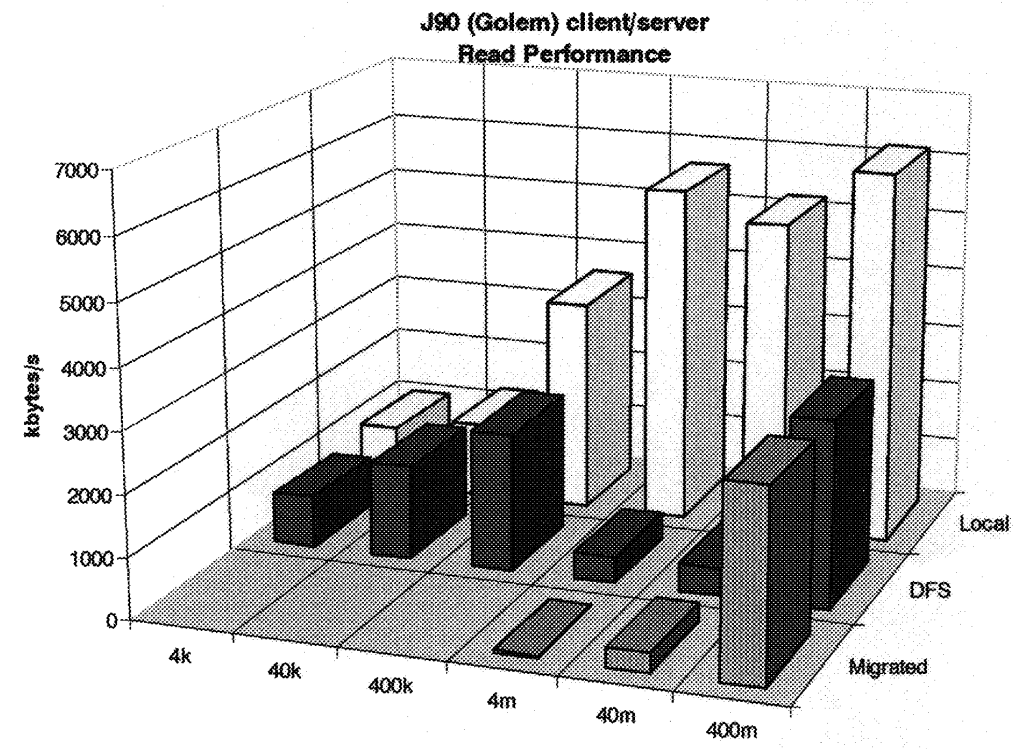
It is likely that there is a place for both technologies.

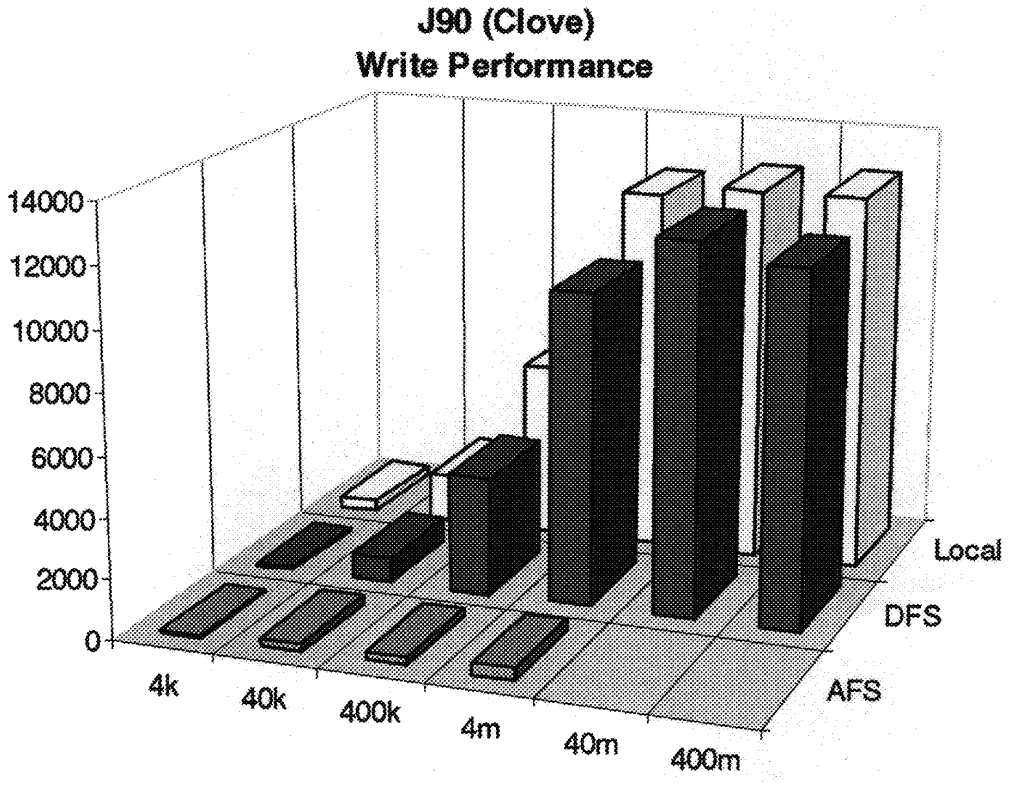



FAR "GET" PERCENTAGES BY AMOUNT OF DATA
00%
80%
60%
40%
20%
0%
7/22/96
7/29/96
8/5/96
8/12/96
8/19/96
8/26/96
9/2/96
9/9/96
9/16/96
9/23/96
9/30/96
10/7/96
AMT OF DATA FROM DISK
AMT OF DATA FROM TAPE


FAR DAILY RELIEF BY SERVER
TIMES PER DAY
25
20
15
10
5
0
7/22/96
7/29/96
8/5/96
8/12/96
8/19/96
8/26/96
9/2/96
9/9/96
9/16/96
9/23/96
9/30/96
10/7/96
JUNIOR
MISC
USERA
USERB
USERC
USERD


J90 (clove) client/server
Read Performance
kbytes/s
14000
12000
10000
8000
6000
4000
2000
0
4k
40k
400k
4m
40m
400m
Local
DFS


J90 (Golem) client/server
Read Performance
kbytes/s
7000
6000
5000
4000
3000
2000
1000
0
4k
40k
400k
4m
40m
400m
Local
DFS
Migrated


J90 (Clove)
Write Performance
14000
12000
10000
8000
6000
4000
2000
0
4k
40k
400k
4m
40m
400m
Local
DFS
AFS