The Growth of the UniTree Mass Storage System at the NASA Center for Computational Sciences

Adina Tarshish
NASA/GSFC Code 931
Greenbelt, MD 20771
(301) 286-6592

Ellen Salmon
Hughes STX Corporation
NASA/GSFC Code 931
Greenbelt, MD 20771
(301) 286-7705
XREMS@CHARNEY.GSFC.NASA.GOV

ABSTRACT

In October 1992, the NASA Center for Computational Sciences made its Convex-based UniTree system generally available to users. The ensuing months saw the growth of near-online data from nil to nearly three terabytes, a doubling of the number of CPUs on the facility's Cray Y-MP (the primary data source for UniTree), and the necessity for an aggressive regimen for repacking sparse tapes and hierarchical "vaulting" of old files to freestanding tape. Connectivity was enhanced as well with the addition of UltraNet HiPPI. This paper describes the increasing demands placed on the storage system's performance and throughput that resulted from the significant augmentation of compute-server processor power and network speed.

I Introduction of UniTree at GSFC

The NASA Center for Computational Sciences (NCCS) is a scientific computing center serving more than 1200 users with a range of needs from supercomputing to data analysis. The UniTree file storage management system first arrived at the NCCS on July 6, 1992. As UniTree was to be the primary system for mass storage management, the Convex C220 was upgraded to a C3240 with four CPUs, 512 megabytes of memory, and 110 gigabytes of disk. Also included in this initial configuration were 2.4 terabytes of nearline robotic storage provided by two StorageTek 4400 silos. Although UniTree supported both NFS and ftp as access methods, access to UniTree was permitted only through ftp in order to meet the throughput demands of our Cray Y-MP (UniTree's primary storage client), IBM 9000 users, and workstation clients.

The mass storage contract under which Convex/UniTree was obtained required that it be able to handle 32 concurrent transfers while 132 other sessions supported users. The size of the transfers done in testing was realistically large, about 200 megabytes each. UniTree ultimately showed itself able to manage this workload, and by the third week in September it had passed acceptance.

In the following sections we describe the extensive efforts of the NCCS in supporting the initial configuration to bring UniTree to a robust production-level file storage system.

II Getting Ultranet Access

The NCCS computing environment supports an UltraNet network configuration between seven buildings serving more than 750 scientists at the Goddard Space Flight Center. This network

provides gigabit-per-second access to the Cray Y-MP and the IBM 9000. At acceptance time, the UniTree system had not yet been compiled and tested under ConvexOS version 10.x, requiring that we backlevel the operating system down to 9.1. ConvexOS UltraNet support, however, required a minimum level of 10.0. When 10.x-compatible UniTree executables were finally available in early October, we upgraded directly to 10.1 and installed the new UniTree routines. A month later, the HiPPI UltraNet hardware interface arrived, followed in a few weeks by the beta UltraNet 4.0 software. The next few weeks were spent stress-testing the system, ultimately uncovering the same bug that had been reported by other beta-test sites, i.e. the native Ultra path (-u) could only handle a maximum of sixteen concurrent transfers, refusing to connect the seventeenth at all. A fix for this problem had been released by Ultra and had been proven to cause the Convex to crash. Crashes also occurred when too many concurrent transfers over the host stack (-uh) path were attempted. By early January of 1993 an Ultra microcode fix was finally available which managed to avoid this problem. The fix allowed up to 28 simultaneous transfers to take place, and Ultra access to UniTree was now enabled for users on the same port used for UniTree Ethernet transfers.

Within a week, we discovered that when Ethernet transfers used the same executables as UltraNet transfers (as intended by the developers), all Ethernet communication throughout the machine would go into a hang state. To overcome this problem, Ultra access via the default UniTree ftp port was removed and enabled via a different port, employing locally-written software to enforce that Ethernet transfers could not be started up on the UltraNet port. Although patches have been applied to address the original problem, the local software has remained in place pending stress-testing of the patches. UltraNet access to UniTree has since enjoyed relative stability in this configuration.

III Usage Rises, Functionality Increases

Once stability on UltraNet was realized, overall demand for UniTree increased sharply. In February alone users added nearly as much data to the UniTree system as they had added in November, December, and January *combined* (fig. 1).

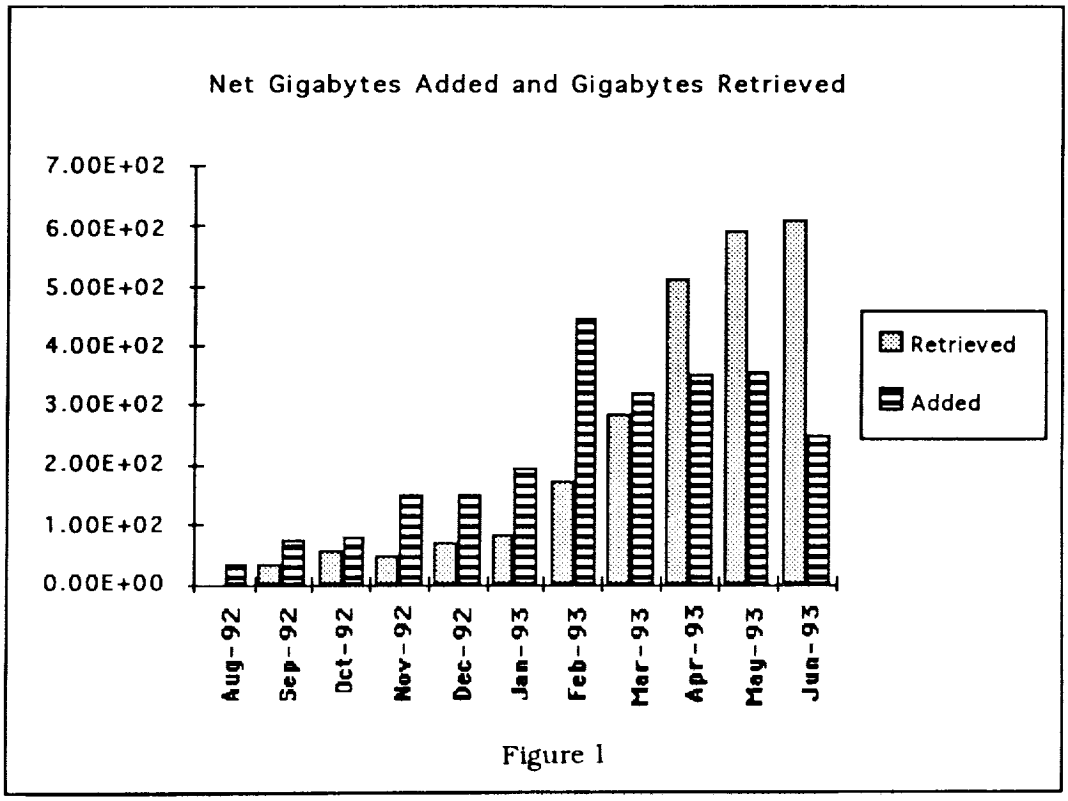

Figure 1

By the end of February, more than 7500 silo tapes out of an available total of 10,000 had been filled with UniTree data (fig. 2).

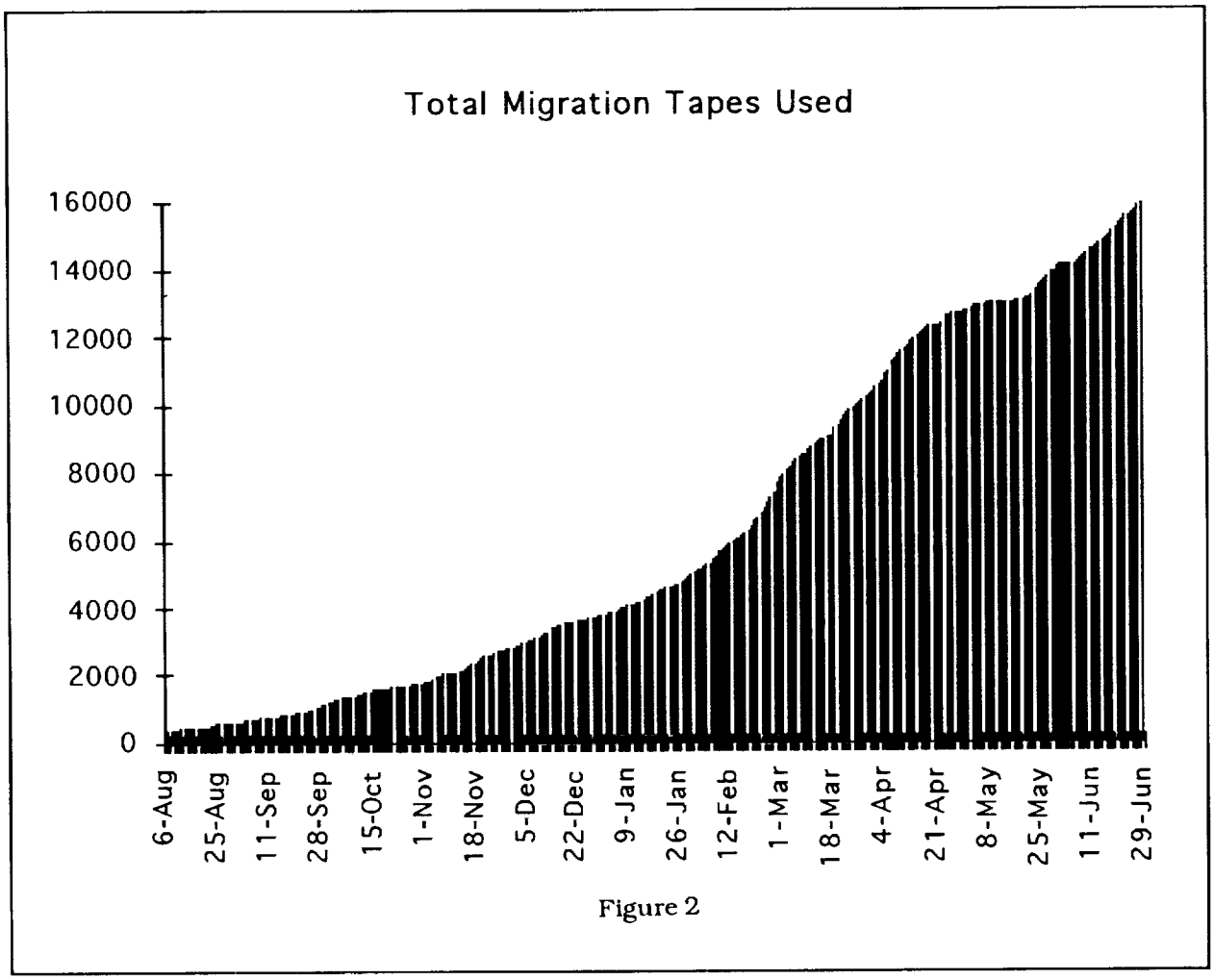

Figure 2

But UniTree's growing popularity soon placed us in a potentially disastrous situation - we were quickly running out of storage. The reason for this was that UniTree 1.5, the only production-level version of Convex/UniTree that existed at that time, did not allow for more than 10,000 tapes to be managed by the system. Not until early March was Convex able to install a modification to allow for up to 100,000 tapes, 18,000 of them for nearline storage and the rest for vaulting, or deep archive.

UniTree vaulting and repacking remained a concern. Our version of UniTree 1.5 did include executables for repacking, or removing the "holes" from tapes caused by deleted files, as well as those for vaulting, or the copying of little-used files onto free-standing tape, for deep archive, but neither of these worked properly at our site. We soon realized that the additional 8000 nearline tapes that could be accommodated by the software would not last for more than a couple of months, and that even if they lasted longer, without repacking or vaulting, we would not be solving our storage crisis but merely postponing it. Another catastrophe we were facing at that time was that both of ours silos were nearly full. Without vaulting, most of the additional 8000 tapes for nearline storage would actually have to be offline, mounted by operators. On busy days, that would amount to hundreds of tape mounts a day. We did not have the operations staff necessary for such an undertaking, nor did we want to slow UniTree down while humans located and mounted the tapes. For these reasons, we found ourselves clamoring

for repacking and vaulting executables that worked, so that we would have a measure of control over the number of free silo tapes.

By April 5 we finally had a working repacker with UniTree 1.5. We began immediately to repack in earnest, freeing hundreds of tapes for new data. By April 22 we had also succeeded in vaulting to free-standing tapes. Working with Convex, we assembled utilities that operators could invoke to place a UniTree label on new free-standing tapes, so that they could be used for vaulting. Both repacking and vaulting are now fully operational and running in a production mode.

By April 14 we had obtained a second modification to UniTree 1.5 that allowed up to 36,000 tapes to be used for nearline storage. In early May we were able to acquire a third silo for UniTree, and as of this writing, we have already used over 3500 of the tapes stored within it.

IV Current Trends in Usage

Recently we have observed that the total number of retrieves for a given period of time outstrip the number of files stored (fig. 1). Since we have discouraged the use of UniTree as a "black hole" from the beginning, we are encouraged by this finding to believe that users are making use of the data they have stored. Accordingly, we have begun to see a considerable increase in the number of silo tape mounts, many of which are done for the purpose of retrieving data from archive (fig. 3).

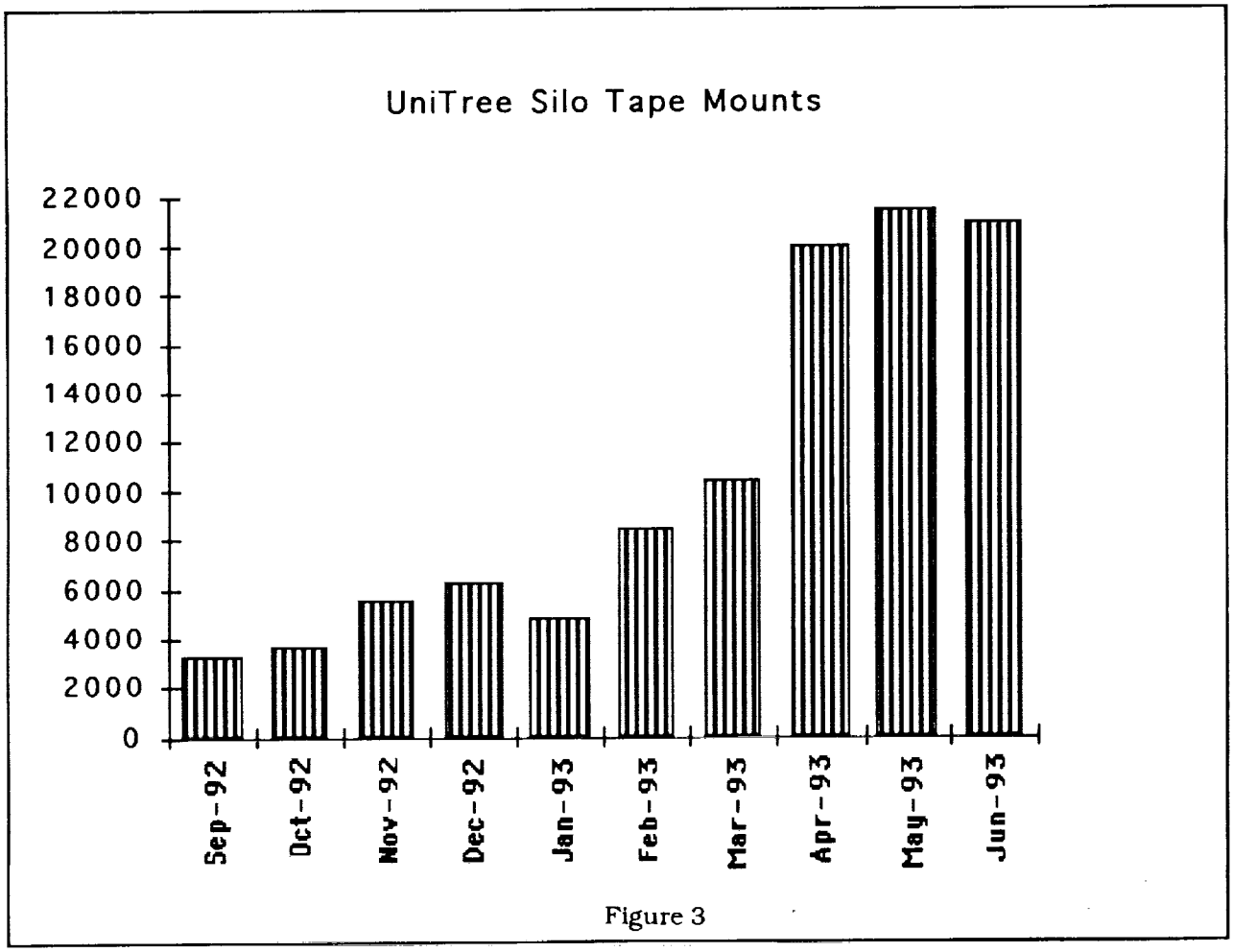

Figure 3

We are also gratified to find that users are making increasing use of the UltraNet interface where available (fig. 4), which frees the heavily-used Ethernet for telnet sessions and transfers from machines lacking UltraNet hardware.

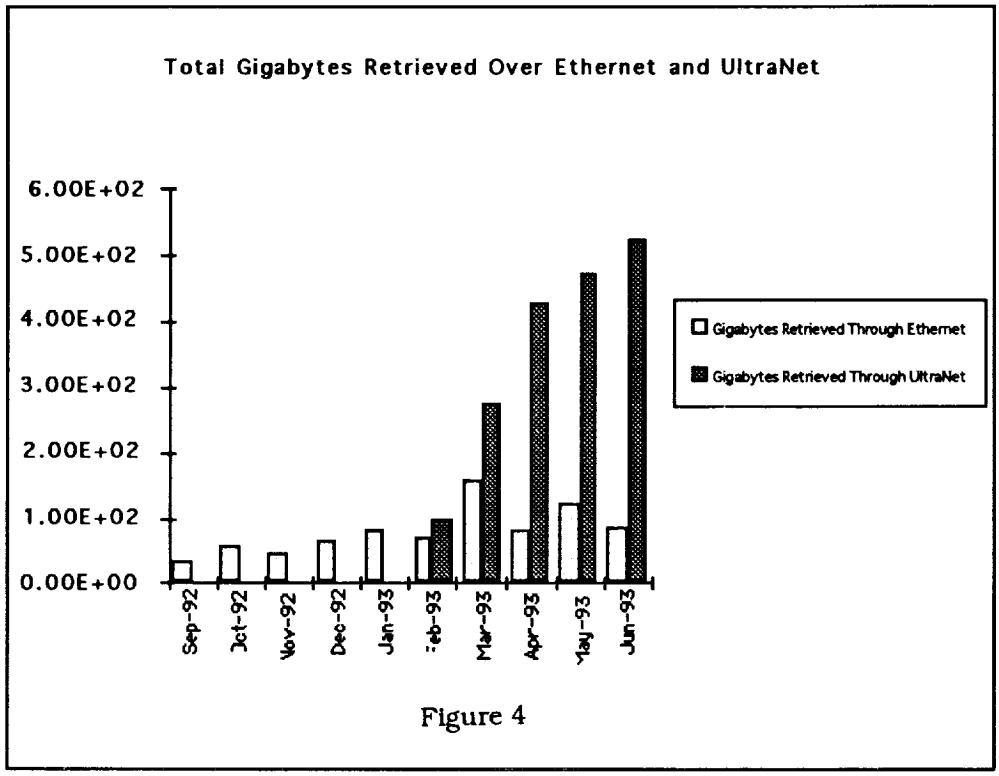

Figure 4

At this time, UniTree contains nearly two-and-a half terabytes worth of data (figure 5).

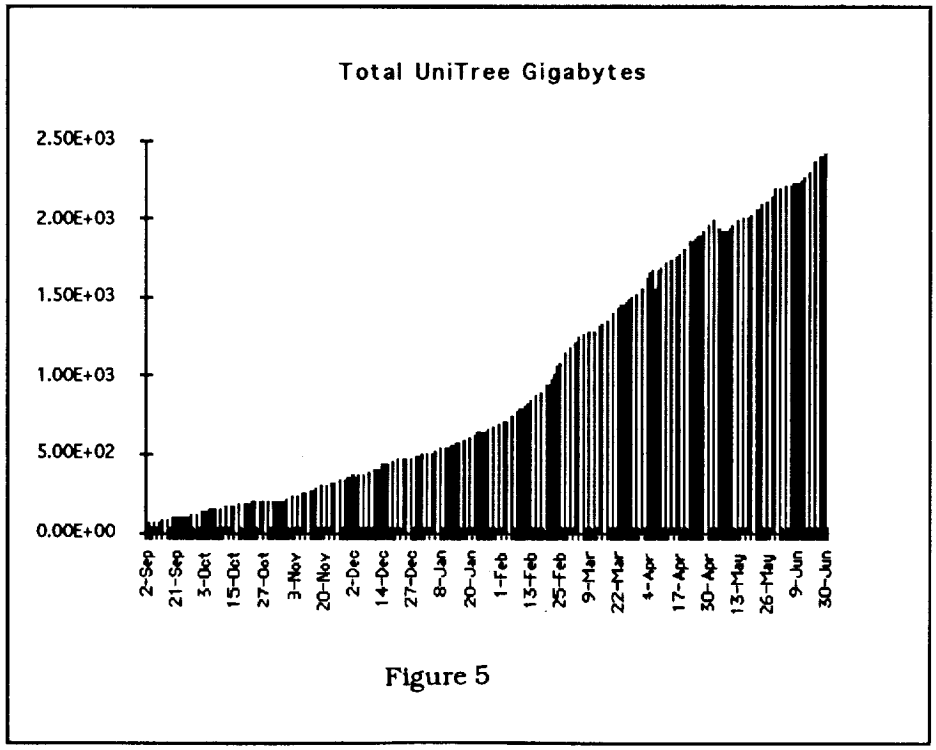

Figure 5

IV Near-Term Challenges

Since late February, our disk cache has been 77.5 gigabytes in size. We have seen many days since then when the cache has been so busy that its contents have completely turned over within 24 hours (fig. 6).

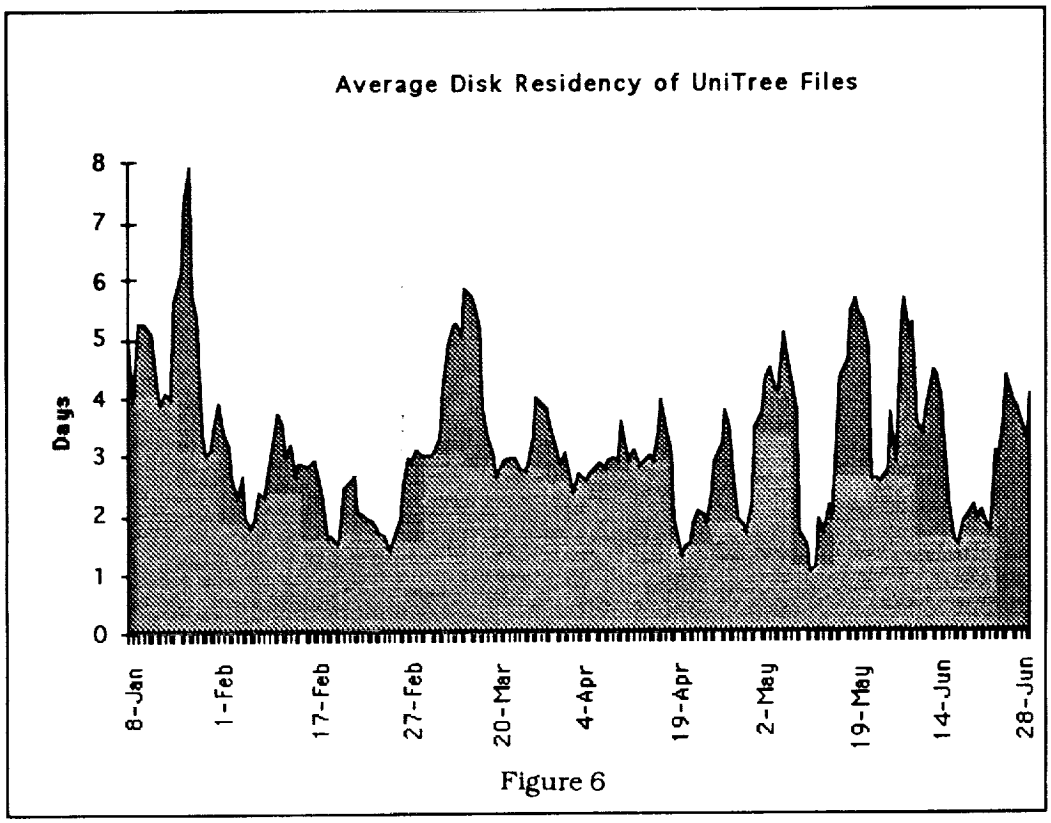

Figure 6

It is very inconvenient for users to find that their data which was stored only the day before must now be retrieved from tape, and it puts a considerably greater load on the UniTree tape processes. To address this problem we ordered an additional 40 GB of disk, bringing our total disk capacity to 150 GB (fig. 7).

However, the inherent difficulty of recovering from disk crashes under UniTree 1.5 prompted us to allocate the newly acquired disk for RAID use and for hot spares instead of using it to enlarge our disk cache, opting for reliability over performance. We have yet to determine the effects RAID has on our UniTree system.

Our real concern of late has been UniTree's current inability to migrate to more than a single tape at a time. In our experience, migration has never been able to proceed faster than one 3480 cartridge tape every six minutes. If migration performed at that peak rate round-the-clock, we would have no more than 240 tapes filled by the end of a 24-hour period. For a cartridge tape with a 200-megabyte capacity, this would mean no more than 48 gigabytes could be migrated each day. There have already been individual days when more than 35 gigabytes have been added to UniTree. The arrival of a new Cray C98 in late August will likely mean a three-fold increase in data production at our site, and if migration to tape cannot keep pace with the arrival of new data, UniTree will crash, irrespective of the size of the disk cache.

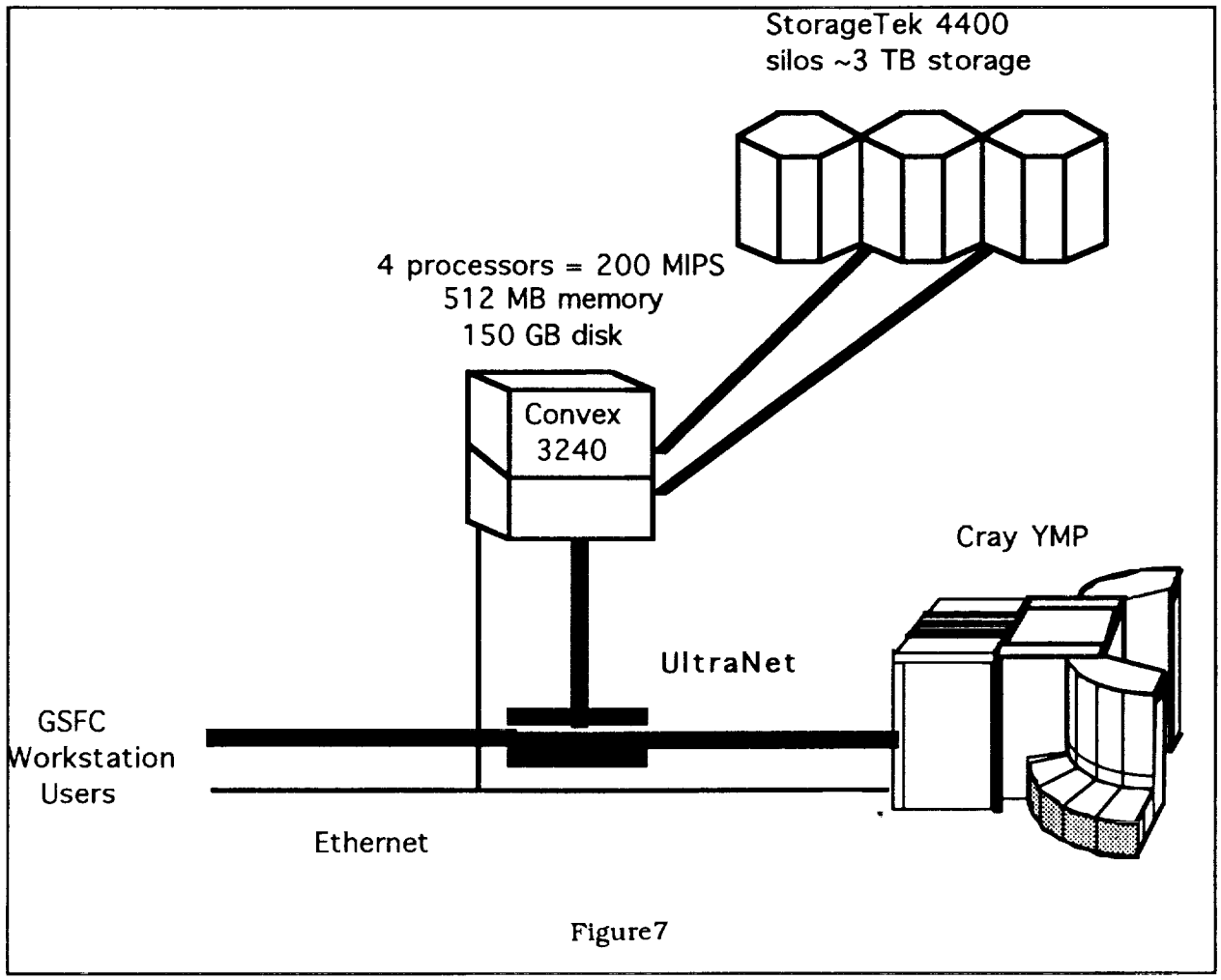

Figure7

For this and other reasons, we are eagerly looking ahead to future releases of UniTree. Convex/UniTree 1.7, released just recently, includes a new feature known as *family of files*, which will allow selected files to be migrated directly to offline tape, bypassing robotic storage entirely. In the same vein, large files could be automatically selected for denser archival media. By the end of the year, a performance release of UniTree 1.7 is expected that we hope will include faster writing to tape and will allow us to accommodate the storage rate anticipated from the new Cray.

V SUMMARY

User acceptance of UniTree has been high, as evidenced by the rapid turnover of our disk cache (figure 6). We have had no complaints about the integrity of the data stored. Although users have found UniTree's instability to be frustrating, we believe that with time UniTree will prove to be the valuable and reliable storage system that mass storage sites have anticipated.