Mass Storage System Upgrades at the NASA Center for Computational Sciences

Adina Tarshish
NASA Center for Computational Sciences, Code 931
NASA/Goddard Space Flight Center
Greenbelt, MD 20771
Adina.Tarshish@gsfc.nasa.gov
Tel +1-301-286-6592
Fax +1-301-286-1634

Ellen Salmon
NASA Center for Computational Sciences, Code 931
NASA/Goddard Space Flight Center
Greenbelt, MD 20771
Ellen.Salmon@gsfc.nasa.gov
Tel +1-301-286-7705
Fax +1-301-286-1634

Medora Macie
NASA Center for Computational Sciences, Code 931
NASA/Goddard Space Flight Center
Greenbelt, MD 20771
Medora.Macie@gsfc.nasa.gov
Tel +1-301-286-3812
Fax +1-301-286-1634

Marty Saletta
NASA Center for Computational Sciences, Code 931 (Raytheon)
NASA/Goddard Space Flight Center
Greenbelt, MD 20771
Marty.Saletta@gsfc.nasa.gov
Tel +1-301-286-9810
Fax +1-301-286-1634

Abstract
The NASA Center for Computational Sciences (NCCS) provides supercomputing and mass storage services to over 1200 Earth and space scientists. During the past two years, the mass storage system at the NCCS went through a great deal of changes both major and minor. Tape drives, silo control software, and the mass storage software itself were upgraded, and the mass storage platform was upgraded twice. Some of these upgrades were aimed at achieving year-2000 compliance, while others were simply upgrades to newer and better technologies. In this paper we will describe these upgrades.

1 Introduction

UniTree first arrived at the NCCS in July of 1992, when it was installed on a Convex C3240. At the time it was attached to 8 3480 tape drives in 2 StorageTek silos, 110 GB of disk, and its main client was a Cray Y-MP. Our UniTree system now runs on a Sun E10000 and is connected to 56 tape drives in 7 silos and an IBM 3494 robotic library, as well as 4 freestanding Timberline drives. It has a disk cache of 1.5 TB, and its main clients are SGI/Cray J932se machines. Figure 1 below shows our current configuration.

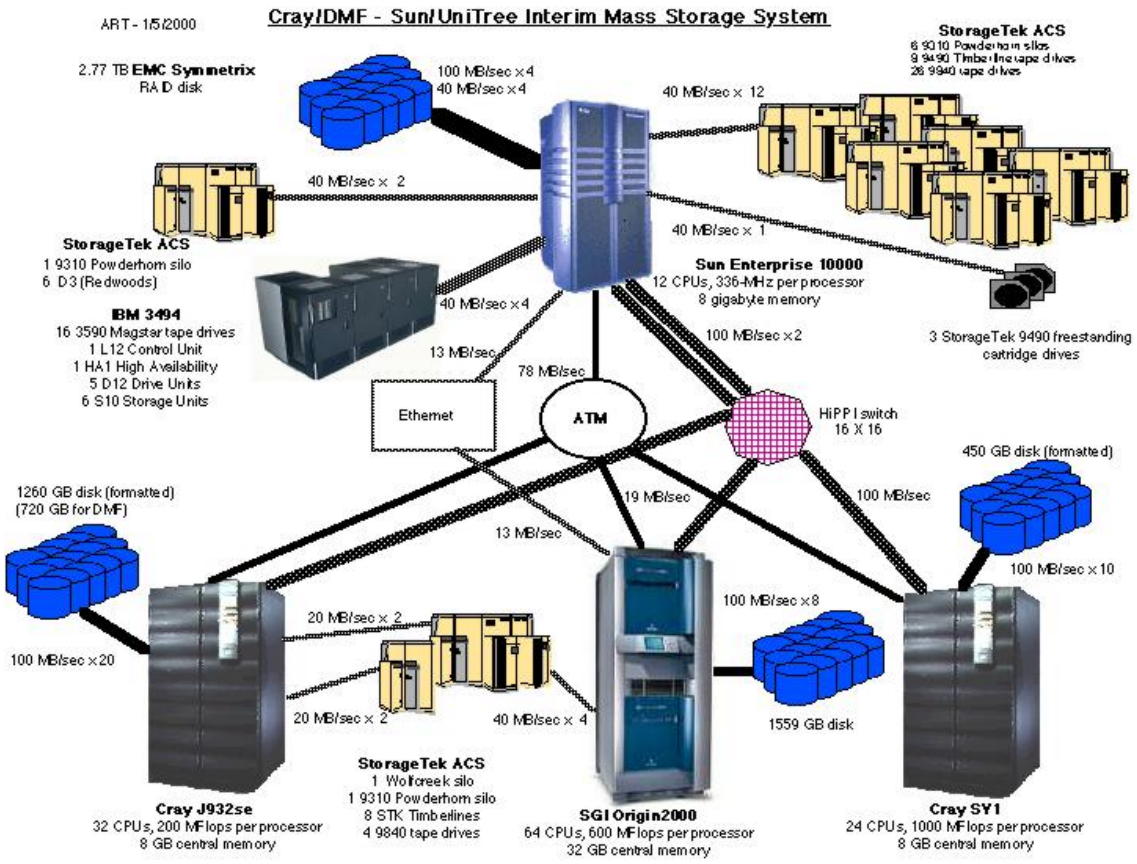

Figure 1. NCCS supercomputing/mass storage configuration

As of November 1, 1999 there were nearly 69 TB of unique data under UniTree's control, plus 35.6 TB of duplicated data stored in a remote facility. Figure 2 shows the breakdown of data by category as of that date.

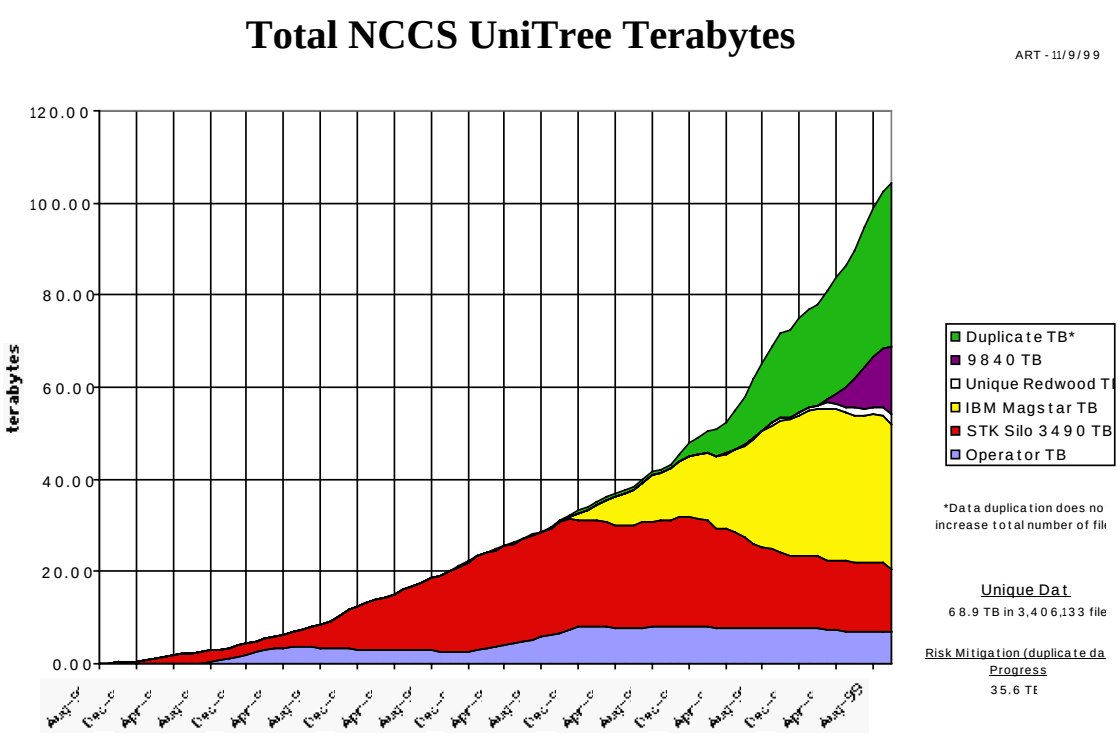

Figure 2. breakdown of data under NCCS UniTree control as of 11/1/99

NCCS users have done as much as 300 GB of network traffic in a single peak day. Figure 3 shows the weekly network traffic to and from UniTree for the past two years.

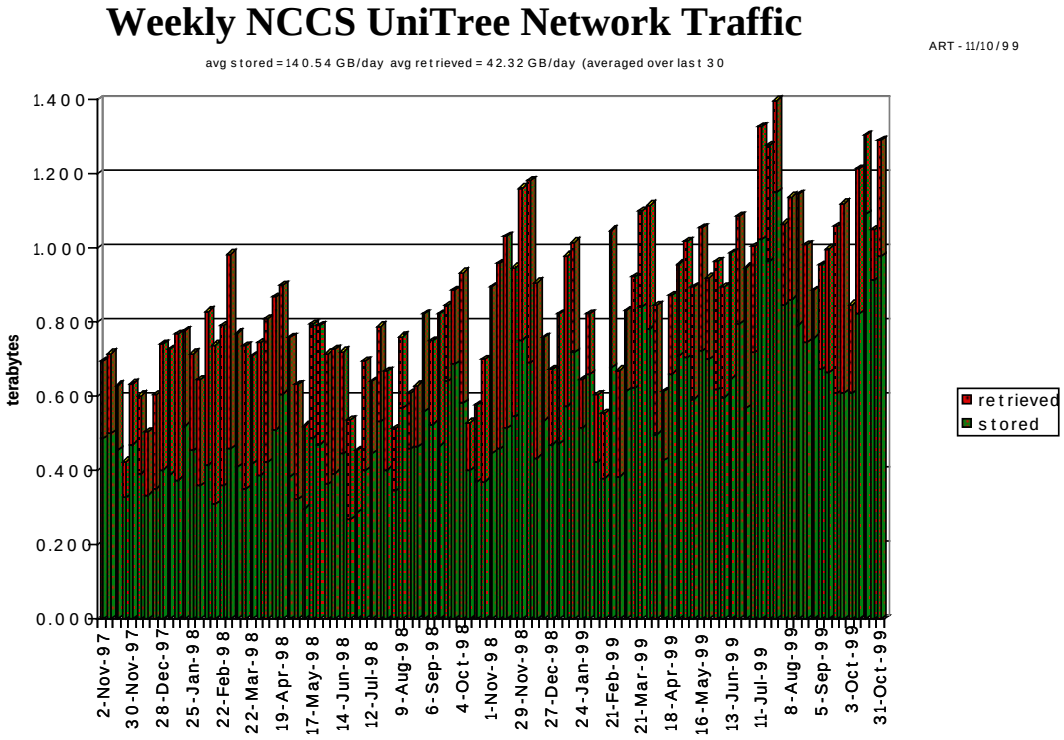

Figure 3. Weekly NCCS UniTree network traffic

The retrieval pattern of the NCCS user community is shown in figure 4 below.

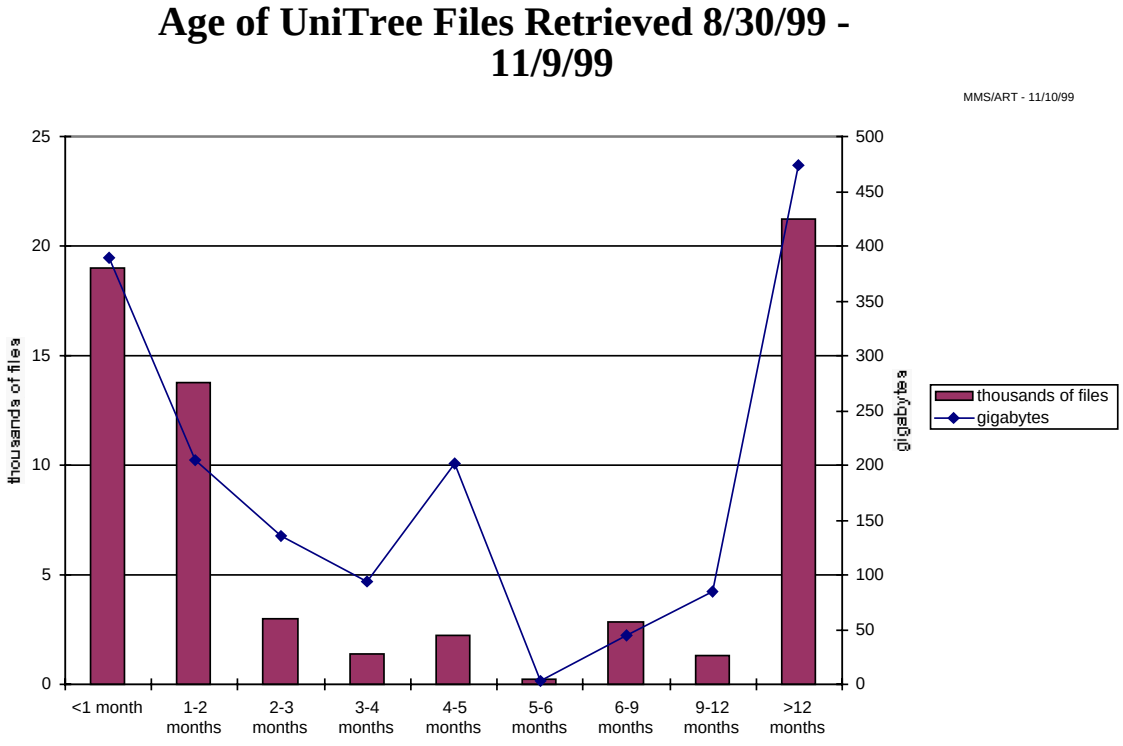

Figure 4. Age of UniTree files retrieved over a 2-month period

The early history of the NCCS UniTree+ mass storage system can be found in the proceedings of the third, fourth, and fifth Goddard conferences on mass storage [1,2,3].

2 Platform/software upgrade

Since September of 1993, the NCCS had run HP's UniTree+ mass storage software on an HP/Convex C3830 machine. By the end of 1997, however, we had decided to survey the market, seeking a newer system that might be less expensive to maintain. In the early spring of 1998 HP informed us that the C3830 was not year-2000 compliant and would not be supported past 9/30/99. Support for UniTree+ was being dropped as well, and they recommended that we convert to UniTree Software, Inc.'s (UTSI) UniTree and run it on a V-class HP machine. With the short amount of lead time we were given to find something compliant, we decided to look for an interim year-2000 solution that would be able to read our UniTree+-written tapes outright, without requiring UniTree+ to remain running as a "middleman". At the time we found only two possible software candidates: LSC's SAM-FS, and UTSI UniTree.

SAM-FS ran only on a Sun platform. UTSI UniTree was supported on HP, Sun, DEC, and SGI machines. Using provisions of NASA's SEWP contract, the four hardware vendors provided us loaner machines with which to "test drive" UTSI UniTree and (for Sun) SAM-FS. By July of 1998 we had four test platforms installed: a Sun Ultra E6000

with A3000 and A5000 RAID disk arrays, an SGI Origin2000 with a Clariion RAID fibre disk array, an HP V2250 with an EMC fibre disk array, and a DEC Alpha 4000 with a StorageWorks RAID Ultra SCSI disk array. Our intention was to test SAM-FS on the Sun E6000 and to test UTSI UniTree on all four. However, the deadline for making the final decision was mid-August. The massive effort of learning and testing two new mass storage systems and four separate flavors of Unix while simultaneously supporting the production UniTree+ mass storage system ultimately proved overwhelming for the limited staff we had. Something needed to be taken out of the equation. We therefore decided that the interim solution would be UTSI UniTree for common-sense reasons: it involved the shortest learning curve, was available on several platforms, and was successfully running at DKRZ, a site which regularly experienced more than three times our network traffic load.

With hardware platform and disk array decisions still to be made, we threw ourselves into configuring the machines and disk for optimum I/O performance. Generally, the disk vendors recommended creating a small number of large disk stripes while UniTree support recommended splitting disk arrays into as many small luns, as possible. In our I/O tests several of the disk arrays had problems handling simultaneous reads and writes to the same lun; writes would wait for the reads to finish before getting started. We concluded that this must be a configuration issue within the machine or the RAID array itself and did not allow it to affect our decision. We ultimately decided to purchase EMC disk and StorageTek Clariion disk because they were reasonably priced and able to connect to various platform types, preventing us from being locked into a particular vendor.

An important component of our testing was network-related. Our largest UniTree clients have always been the Cray systems, to which the UniTree platform has long been connected via a HiPPI switch. The switch being used in production was a Netstar with only copper interfaces, while the test machines required fiber interfaces. We had a new Gigalabs HiPPI switch waiting in the wings to be tested, so a fiber "blade" was ordered for that switch as well as a HiPPI modem for the Netstar switch. Complications arose when certain ftp retrieves over HiPPI hung; this was eventually traced to the HiPPI modem. Transfers over HiPPI that did not hang were nevertheless much slower than expected. ODS, the manufacturer of the faulty modem, decided to give us a loaner HiPPI switch, confident that we would see much greater HiPPI performance than we had in the past.

Meanwhile, our deadline for making our decision arrived, and we were forced to use the data we had at that point. Network performance was obviously inconclusive. However, Sun was the most cost-effective solution, and we were comforted by the fact that it was UTSI's primary UniTree port and that DKRZ was running UniTree on a Sun with far heavier loads than we expected at the time. We decided to purchase a Sun Ultra E6500 as our interim year-2000-compliant mass storage platform. At the same time we purchased 1.3 TB of EMC disk, 900 GB of StorageTek Clariion disk, 22 StorageTek 9840 tape drives, and 4 freestanding Timberline tape drives. Since the 9840 tapes drives

were not yet available, we were to receive Timberlines to use temporarily in their place. The upgrade to 9840 tape drives will be discussed in the next section.

The new Sun E6500 arrived, and we promptly rolled up our sleeves and got to work. For our 24 IBM Magstar tape drives, we obtained both the Sun Magstar driver as well as the IBM Magstar driver. Testing quickly showed that with the Sun Magstar driver we were unable to append to previously-written tapes, making our decision to go with the IBM driver a simple one. UTSI, meanwhile, was busily developing an interface for UniTree to communicate with the IBM 3494 robotic library. When this was completed, and we had tinkered with a test system long enough to feel reasonably comfortable with how it worked, we converted the existing UniTree+ 3.0 test system we had worked with on the HP/Convex C3830. This was an excellent exercise and brought to light many issues we would have to deal with when conversion of the production UniTree+ system would occur. Among the most important was the discovery that tape format "A" written under versions of UniTree+ prior to 3.0 was actually different than the same format "A" written under UniTree+ 3.0. This discovery surprised us greatly. In the course of our attempt to upgrade to UniTree+ 3.0 months before, we had converted to and reverted from version 3.0 several times, until the major problems were fixed, and we had never seen 2.0 show any difficulty reading 3.0-formatted tapes, nor did we see 3.0 show any difficulty reading 2.0-formatted tapes. UTSI UniTree, however, needed separate tape types defined for the separate tape formats. Mark Saake of UTSI, who had made this startling discovery, proceeded to analyze all of our production tapes. He then listed for us those that had been written since the upgrade to version 3.0 as well as those that were being written at conversion and reversion times, invariably written partly in one format and partly in the other. He advised us to repack those mixed-format tapes before conversion to UTSI UniTree, advice we carefully followed.

On Wednesday January 27, 1999, we halted user activity on the UniTree+ system running on the C3830 and allowed migration to complete. We copied the databases to theSun E6500, and Mark began the conversion process. The Sun assumed the IP address of the C3830. Tape drives were uncabled from the C3830 and connected to the Sun. By about 10 PM that evening, users were permitted access. Mark remained logged into our new system through the entire night, and virtually no problems were seen. The UniTree we began running that evening was version 1.9.1 into which was backported several 2.x features such as Y2K support, support for up to 64 tape drives, and support for raw disk cache devices greater than 2 GB. Additionally, it had 3494 robotic support, which had not been previously available with UTSI UniTree.

3 Tape Drive Upgrades

9840s were a technology we had been awaiting a long time. Nearly half our UniTree files were under 1 MB in size, effectively ruling out helical drives for anything but duplicate copies. However, we were very interested in denser linear technologies. Back when StorageTek’s densest linear technology was 3490E tape, storing something over 1 GB a cartridge, we had made the decision to purchase an IBM 3494 tape library with 8 Magstar drives. At 10 GB per cartridge uncompressed and 9 MB/s transfer rate, this

technology served us very well. When it was available, we purchased additional Magstar tape drives in C12 cabinets and installed them in two of our StorageTek silos, which increased our silo data capacity tenfold. By the time StorageTek 9840s became available, we had 16 IBM Magstar drives in the silos and another 8 Magstar drives in the IBM 3494 robotic library, and we had been using Magstars to store all new data coming into UniTree for nearly two years with great success.

9840s had a capacity of 20 GB uncompressed, and were therefore very desirable. However, their arrival would mean the silo Magstars would have to go. The Library Management Unit microcode level that was required for the 9840s disallowed Magstar cartridges with "J" letters on them. Without "J"s on cleaning cartridges, we could not have automatic cleaning enabled in the silos, because our silos were mixed-media – they had Timberlines as well as Magstars. Disabling automatic cleaning was not an option in silos mounting hundreds of tapes a day. Since 9840s and Magstars could therefore not coexist in the same Automated Cartridge System, we had 2 choices: either to move the Magstars out of the silos entirely, or to create a separate ACS with back-level LMU microcode for the Magstar drives and install the 9840s in the other ACS. The first choice was a much better one, especially since we had a 3494 robotic library that was already home to 8 Magstar drives. Our next step was to purchase additional drive cabinets for the 3494, with the intention of moving the 16 silo Magstars into them. Sometime after this 3494 upgrade took place, however, IBM informed us that in a SCSI-connected robotic library we could only fit 16 Magstar drives total. They offered to take back the extra 8 silo Magstars, as well as the 4 silo cabinets in which they had been housed. In return, they would provide certain items that were of interest to us.

In preparation for the move of 8 silo Magstars into the 3494 robotic library, the silo Magstar tapes had been gradually transferred to the 3494. A few weeks after the UniTree conversion, the Magstar tape drives were moved out of the silos. This paved the way for the upgrades required for STK 9840 support. First the Library Management Unit microcode was upgraded, then the hands of the robots themselves, then the Automated Cartridge System Library Software that manages tape access. Finally, the 9840s themselves were installed and tested. By the end of March all new data being stored into UniTree was being written to 9840 tapes.

The following month, IBM announced their E1A product, a 256-track Magstar drive that could write double the original density to the same Magstar media. At the time of this writing, 7 out of our 16 Magstar tape drives have been upgraded from 128-track to 256-track, and data on older 3490-type media is being rewritten onto Magstars with these new drives.

4 Further upgrades

The Sun E6500 we had purchased was fully adequate for supporting our current load. However, we were told that some of our users expected their storage needs to increase considerably. To support this increase, UniTree would require more disk; however, the E6500 had almost no room for additional peripherals. Ultimately, we replaced the E6500

with an E10000 as the UniTree server. With nearly 5 times the I/O bandwidth, 4 additional CPUs, 11 additional SCSI adaptor slots, and room for 4 additional system boards, the E10000 gave us the ability to scale up to whatever the near term might require. We are successfully running UniTree on the Sun E10000 today.

At the time of the original upgrade to UTSI UniTree, 2.0 had just been released, but we were advised to upgrade to the relatively stable 1.9.1 instead, with support for y2k, 64 tape drives, and raw disk partitions greater than 2 GB backported into our version. UniTree 2.1, however, offered considerable performance increases as well as the ability to use up to 256 disk cache partitions, where 1.9.1 allowed for only 110. Our first attempt at upgrading to 2.1 uncovered some bugs which were quickly fixed. Our second attempt shortly thereafter uncovered a Solaris bug which UTSI was eventually able to work around. On Monday August 30 we successfully upgraded to 2.1 and never looked back.

In the past year, some of our users have informed us that their data requirements are expected to increase drastically over the next several years. To accommodate this anticipated increase we decided to purchase more disk for UniTree. After testing several brands and reviewing offers from several vendors, we purchased an additional 4 TB of 10,000 RPM disk from EMC. Adding this new disk to UniTree's cache will require the repartitioning of the EMC disk we are currently using, since UniTree at present can use only 256 disk cache partitions.

5 Future Work

Within the next several months (as of this writing) 4.5 TB of High Performance Computing data, currently stored under DMF on a Cray T3E, are expected to be moved into UniTree using a DMF FTP Media Specific Process (MSP). This will allow current T3E DMF users to continue retrieving their files transparently from DMF during and after the move. The MSP defaults to one flat UniTree directory per DMF user, which is causing us some concern. There will be some very large UniTree directories created in this fashion, one of which is expected to contain over 200,000 files. Testing of this situation and how it might affect UniTree performance for other users is ongoing, and namesrvr tuning advice is being solicited from UniTree support. The silo on which these DMF files have been stored, along with its 4 Redwood drives, will be attached to the existing remote silo to add to UniTree's duplicate-copy capacity.

Also ongoing is the effort to duplicate existing UniTree data in our remote silo. Duplication of new data coming into UniTree has been automatic since November 1997. As of this writing 33.2 TB of older UniTree data remains to be duplicated. A UniTree 2.1 facility allows an existing file to be marked "dirty" so that it is rewritten to tape and a second copy generated at the same time. This will have the added benefit of rewriting data on older 3490-type media to new 20 GB media.

In the long term, we plan to do a more thorough survey of the HSM market, the survey we would have done had Y2K not become a pressing deadline. Many new HSM

products have come to market within the past few years, and we are interested in keeping abreast of these new developments.

References

[1] A. Tarshish and E. Salmon, "The Growth of the UniTree Mass Storage System at the NASA Center for Computational Sciences," *Proceedings of the Third Goddard Conference on Mass Storage Systems and Technologies*, (1993) 179-185.

[2] A. Tarshish and E. Salmon, "The Growth of the UniTree Mass Storage System at the NASA Center for Computational Sciences: Some Lessons Learned," *Proceedings of the Fourth Goddard Conference on Mass Storage Systems and Technologies*, (1995) 345-357.

[3] E. Salmon, "Storage and Network Bandwidth Requirements Through the Year 2000 for the NASA Center for Computational Sciences," *Proceedings of the Fifth Goddard Conference on Mass Storage Systems and Technologies*, (1996) 273-286.