Mass Storage Upgrades at the NASA Center for Computational Sciences

Adina Tarshish (adina.tarshish@gsfc.nasa.gov)
Ellen Salmon (ellen.salmon@gsfc.nasa.gov)
Medora Macie (medora.macie@gsfc.nasa.gov)
Marty Saletta/Raytheon (marty.saletta@gsfc.nasa.gov)

Earth and Space Data Computing Division
Earth Sciences Directorate
NASA Goddard Space Flight Center
Who We Are

- Central computing facility providing high-end computing and mass storage services for NASA-funded Earth and Space science researchers
- Part of Earth and Space Data Computing Division
- Cray SV1s, Origin 2000, Cray T3E compute servers (plus user desktop workstations) feed data into Sun-based UniTree mass storage system
We’ve Come a Long Way with UniTree

- We’ve run some form of UniTree since July ‘92
- Started out on a Convex C3240
  - Two STK silos with eight 18-track tape drives
  - Eight freestanding 18-track tape drives
  - 110 GB disk cache
- Currently on a Sun E10000
  - Seven STK silos
    - Eight 36-track Timberlines
    - 26 9840 tape drives
    - Two Redwood SD-3 tape drives
  - One IBM 3494 library with sixteen Magstar E1A tape drives
  - Four freestanding Timberlines
  - Close to 4 TB disk cache
Mass Storage System Upgrades at the NASA Center for Computational Sciences
NCCS  Total UniTree Terabytes - 3/16/00

*Data duplication does not increase total number of files

Unique Data: 77.17 TB in 3,905,165 files
Risk Mitigation (Duplicate Data):
Progress: 47.41 TB
avg stored = 168.25 GB/day  avg retrieved = 56.37 GB/day  (averaged over last 30 days)
Mass Storage System Upgrades at the NASA Center for Computational Sciences

Spring of ‘98

- We upgraded our HP UniTree+ software from Rel. 2.0 to 3.0
  - Final upgrade occurred after 3 test upgrades
  - Gave us the ability to duplicate existing files
- HP informed us that the C3830 that hosted our UniTree+ system was not Y2K-compliant
- Very short lead time
  - No time to adequately evaluate alternative HSM software
  - Needed to find a system that could read the TB written already in UniTree+ format
  - Wanted the shortest learning curve possible
- We decided that UniTree Software Inc.’s UniTree Central File Manager would be our choice, at least for the interim
  - Still needed to decide on platform and disk
On What Machine?

- At the time, UniTree Software, Inc.’s (UTSI) UniTree Central File Manager (UCFM) was supported on Sun, HP, DEC, and SGI
- Sun offered us an E6000 through SEWP as a test machine; HP, SGI, and DEC also provided machines for testing
- Eventually, we found ourselves with the following test platforms:
  - Sun E6000 with A3000 and A5000 RAID disk arrays
  - SGI Origin2000 with Clariion RAID fibre disk array
  - HP V2250 with EMC RAID fibre disk array
  - DEC Alpha 4000 with DEC StorageWorks RAID USCSI disk array
- Each machine was allocated a silo Timberline and a silo Magstar drive
Testing Begins, Summer of ‘98

- dd-testing to all disks except for EMC showed poor performance for simultaneous reads and writes
  - reads would be stuck waiting until writes completed
- HiPPI gave us big headaches, but was required to get adequate transfer performance with Cray J90s
  - Sun, SGI, and DEC machines had serial (fiber) HiPPI interfaces, while our production Netstar switch was all parallel (copper)
  - We managed to get hold of both an ODS HiPPI modem for the Netstar switch and a fiber blade for our Gigalabs switch that was still in test mode
  - HiPPI modem induced hangs for retrieves over a certain size
Out of Time

- By August ‘98, time had run out, and we were forced to decide—a Sun E6500 was selected
- Many factors were considered, including
  - Reasonable cost
  - Performance
  - Other UniTree sites with a greater load than ours were running successfully on a Sun
  - Solaris was the initial port for UTSI releases
- We chose RAID disk arrays that were attachable to other machines, in case long-term follow-on system was not a Sun
  - 1.3 TB EMC disk
  - 900 GB Clariion fibre disk (purchased from STK)
Tape Drives

- At the same time, we also purchased Storage Tek 9840 tape drives
  - Until they were ready, we received Timberlines in their place
- We had 16 IBM Magstars in our silos
  - Were working well for us, until...
  - STK’s next version of LMU microcode disallowed “J”s on cartridges, which we needed Magstar cleaning carts to have if we wanted automatic cleaning
  - We remained on back-level microcode so we could continue to run
  - 9840s, though, required newest microcode
Mass Storage System Upgrades at the NASA Center for Computational Sciences

Tape Drives (cont’d)

• IBM 3494 robotic library already had 8 Magstar drives
• We purchased 4 more D12 cabinets to hold the 16 drives that were in the silos, so that 9840s could be installed
• IBM afterwards told us that with SCSI drives in our library we couldn’t install any more than the 8 drives we already had
• More research indicated that with new feature code we could install 8 more, leaving us with 8 silo drives in limbo
Tape Drives (cont’d)

- After negotiating with both IBM and STK, we decided to give the 8 extra Magstars back to IBM
- In return IBM agreed to
  - apply whatever feature codes were necessary to the 3494 so it could run with 16 SCSI-attached drives
  - move the 8 drives down into the 3494 (silos were upstairs)
  - give us certain upgrades to the drives that were soon to be available
- This tape drive move-and-return wound up not happening until after conversion to UTSI UniTree (more on this later)
UniTree Tapes, Fall of ‘98

- Test conversion of a UniTree+ 3.0 test system to UTSI UniTree went fine
- No problems storing and migrating new data
  - Sun driver for Magstars wouldn’t let us append to current write tapes, but IBM driver worked well
- Trouble encountered retrieving some UniTree+ written files, but not others
- UTSI discovered that many of our tapes that we thought had the identical media type actually differed in block size
UniTree Tapes (cont’d)

• It was determined, based on the tapes with the strange new media types, that UT+ 3.0 formatted its tapes differently than UT+ 2.0
• UT+ 2.0 easily read UT+ 3.0-written tapes, and vice versa
  – we knew that because we had converted and reverted to and from 3.0 3 times until we stayed on 3.0 for good
UniTree Tapes (cont’d)

- Each conversion and reversion to and from UT+ 3.0 had created tapes half-written in 2.0 format, half-written in 3.0 format
- Mark Saake of UniTree Software Inc. analyzed our UniTree+ logs and determined the tapes that were totally 3.0-written and the tapes that were mixed-format
- 3.0-written tapes needed a separate media type for UTSI UniTree
- Mixed-format tapes needed to be repacked under UniTree+
  - considerable manual intervention would be needed for mixed-format tapes to be readable under UTSI UniTree, resulting in delays for users
More UniTree Testing

- Test system had very many mixed-format tapes, making it difficult to work with
- A read-only copy of the production UniTree+ system could provide a large-scale, realistic test
- Once production UniTree+ copy 0 mixed-format tapes were repacked, UTSI converted a snapshot of production UniTree+ databases for the new test system
- Pilot users selected and given access
- Some minor bugs were found and quickly fixed
Actual Conversion, Jan. ‘99

- User traffic was halted on UniTree+ while migration completed
- Databases ftped to Sun and converted
- Critical files moved
- Sun acquired IP address of the C3830
- All tape devices uncabled from C3830 and cabled to Sun
- Success!
Next Hurdle

• IBM had greatly delayed the move of the silo Magstars into the 3494
  – We could not upgrade the necessary silo components to install 9840s until this happened
• Finally, after conversion to UTSI, this move occurred
• Dual-active-accessor also installed
  – allowed both robots (one used to be just a hot backup) to actually work at the same time
• With Magstar tapes and drives no longer in silos, STK C.E. upgraded our Library Management Unit, LMU microcode, and the robot hands (also required for 9840 support)
• We then upgraded the silo control software (ACSLS), and finally installed the 9840 tape drives
More Tape Drive Upgrades

- IBM then announced their 256-track Magstar tape drive, the E1A
- Over the next several months, we gradually upgraded all 16 of our Magstars to 256-track
- Currently rewriting older data on 3490s to the 256-track Magstar tapes
Platform Upgrade

- With 16 Magstars, 26 9840s, 6 Redwoods, 4 freestanding tape drives, and 1.5 TB of disk, we were nearly out of I/O slots
- Users were already indicating increasing requirements within the next few years
- In June of '99 we upgraded the UniTree server from the E6500 to an E10000
  - 5 times the I/O bandwidth
  - 11 additional USCSI adapters
  - 4 additional CPUs
  - Room for 4 additional system boards with I/O slots
UniTree Upgrade, Aug. ‘99

- UniTree 2.1 was now available
  - Some of its features, such as Y2K-compliance and support for 64 tape drives had been backported to 1.9.1 for us
- 2 initial upgrade attempts made, then third attempt succeeded
  - As first site to upgrade with 56 tape drives, we brought to light some previously unknown bugs
More Disk

- Under UniTree 1.9.1 we had 817.5 GB for UniTree disk cache
- Couldn’t add much more because 1.9.1 was limited to 110 partitions
- 2.1 extended that limit to 256, so after upgrade to 2.1 we gradually added 775 GB of the Clariion disk we had purchased
- For several months we had a total of 1.56 TB of disk cache
More Disk (cont’d)

• Clariion’s inability to failover or load-balance caused concern
  – ATF, the product provided for failover, interfered with the installation of totally unrelated patches, as well as with the EMC disk devices, so we could not run it
  – When one of the fibre interfaces failed, UniTree took a hard hit
• Eventually we purchased another 3.5 TB of EMC fibre-attached disk and decommissioned the Clariion
  – EMC’s PowerPath had been working well for us for both load-balancing and failover
  – Currently have nearly 4 TB of UniTree disk cache, all EMC
Redwoods

- Since Nov. ‘97 we had been duplicating all new UniTree data to Redwood tape stored in a remote silo
- StorageTek recently announced the end of Redwood support as of 9/30/2002
- We moved one of our 6 local silos with its 4 9840 drives to the remote location, to take over the duplication task from the Redwoods
- We have more than 40 TB of duplicate UniTree data on Redwood media to rewrite by then
Near-term

- Users have indicated massively increased requirements for the next several years
- Working on a mass acquisition to fill those requirements
Earth Science Computing Center
Mass Storage Requirements
2000-2004
Earth Science Computing Center
Projected Total Data Stored Requirements

Projections from "Refined Earth Science Computing Requirements" 10/21/1999
Earth Science Computing Center Unique* Data
Observed and Projected** Terabytes by Media Type

* Does not include risk mitigation duplicates
** Simplistic media scenario; total stored based on 10/99 user estimate
Earth Sciences Computing Center Unique** Data
Estimated-Observed and Projected-Optimistic Media Usage

*1 "Silo" = 5743 Cartridges   ** Does not include risk mitigation duplicates
Earth Science Computing Center Unique** Data
Estimated-Observed and Projected-Pessimistic Media Usage

*1 "Silo" = 5743 Cartridges   ** Does not include risk mitigation

* * *
FY 2004: 5 PB total, 2.7 TB/day new

- Implications and questions:
  - 5 PB at $4/GB = $21M for media alone
    - $7.9M if $1.50/GB
  - Media density: 7 silos per building or 27?
  - ~130 million files (4 million files today), performance for metadata operations?
  - Avg. network 55 MB/s (437 Mbps) sustained (assumes retrieve traffic ~1.8 TB/day)