53-82
83202

Storage and Network Bandwidth Requirements Through the Year 2000 for the NASA Center for Computational Sciences

Ellen Salmon
Science Computing Branch
NASA Center for Computational Sciences
NASA/Goddard Space Flight Center, Code 931
Greenbelt MD 20771
xrems@dirac.gsfc.nasa.gov
phone: 301-286-7705
fax:301-286-1634

Abstract

The data storage and retrieval demands of space and Earth sciences researchers have made the NASA Center for Computational Sciences (NCCS) Mass Data Storage and Delivery System (MDSDS) one of the world's most active Convex UniTree systems. Science researchers formed the NCCS's Computer Environments and Research Requirements Committee (CERRC) to relate their projected supercomputing and mass storage requirements through the year 2000. Using the CERRC guidelines and observations of current usage, some detailed projections of requirements for MDSDS network bandwidth and mass storage capacity and performance are presented.

Introduction

The mission of the NASA Center for Computational Sciences is to enable advanced scientific research and modeling for NASA-sponsored space and Earth science researchers by providing a high performance scientific computing, mass storage and data analysis environment. Science efforts supported by NCCS resources include climate data assimilation and other atmospheric and oceanographic sciences, orbit determination, solid-earth and solar-terrestrial interactions, magneto hydrodynamics, and astrophysics. The NCCS is a part of NASA/Goddard's Earth and Space Data Computing Division (ESDCD).

The NCCS has been directed to focus on use of Commercial Off-The-Shelf (COTS) products and "open" software that runs on more than one vendor's hardware platform. The Mass Data Storage and Delivery System currently uses UniTree software for hierarchical file storage management.

MDSDS Environment and Configuration

The MDSDS UniTree system became operational at the NCCS in October 1992. Since that time, MDSDS robotic storage has grown from the initial two StorageTek 4400 silos with 2.4 TB total capacity (200-MB 3480 cartridge tapes) to 6 StorageTek silos with 28.8 TB capacity uncompressed (800-MB 3490E cartridges; in operation the MDSDS stores about 1 GB compressed data per cartridge). The MDSDS is currently adding an IBM 3494 robotic library with an additional 24 TB (10-GB 3590 tape cartridges).

The MDSDS UniTree software runs on a Convex C3830. While UniTree supports both NFS and ftp, access is limited to ftp to support throughput demands from the NCCS's supercomputers.

Local high-speed network connections join the MDSDS with the NCCS supercomputers, which are the primary sources and sinks for MDSDS data. The original Ethernet was augmented by an UltraNet/HiPPI connection in early 1993. By late 1994 the UltraNet/HiPPI connection had been replaced with two HiPPI/TCP connections, and FDDI was deployed to provide higher speed access to the rest of the NASA/Goddard campus.

The NCCS Supercomputers have undergone significant augmentation from the 4-processor Cray YMP (1.2 GFLOPs) in place October 1992. The current supercomputers are 3 Cray J90 systems, at present configured with 68 processors (13.3 GFLOPs) and to be upgraded to 96 processors (~19 GFLOPs) by spring of 1997.

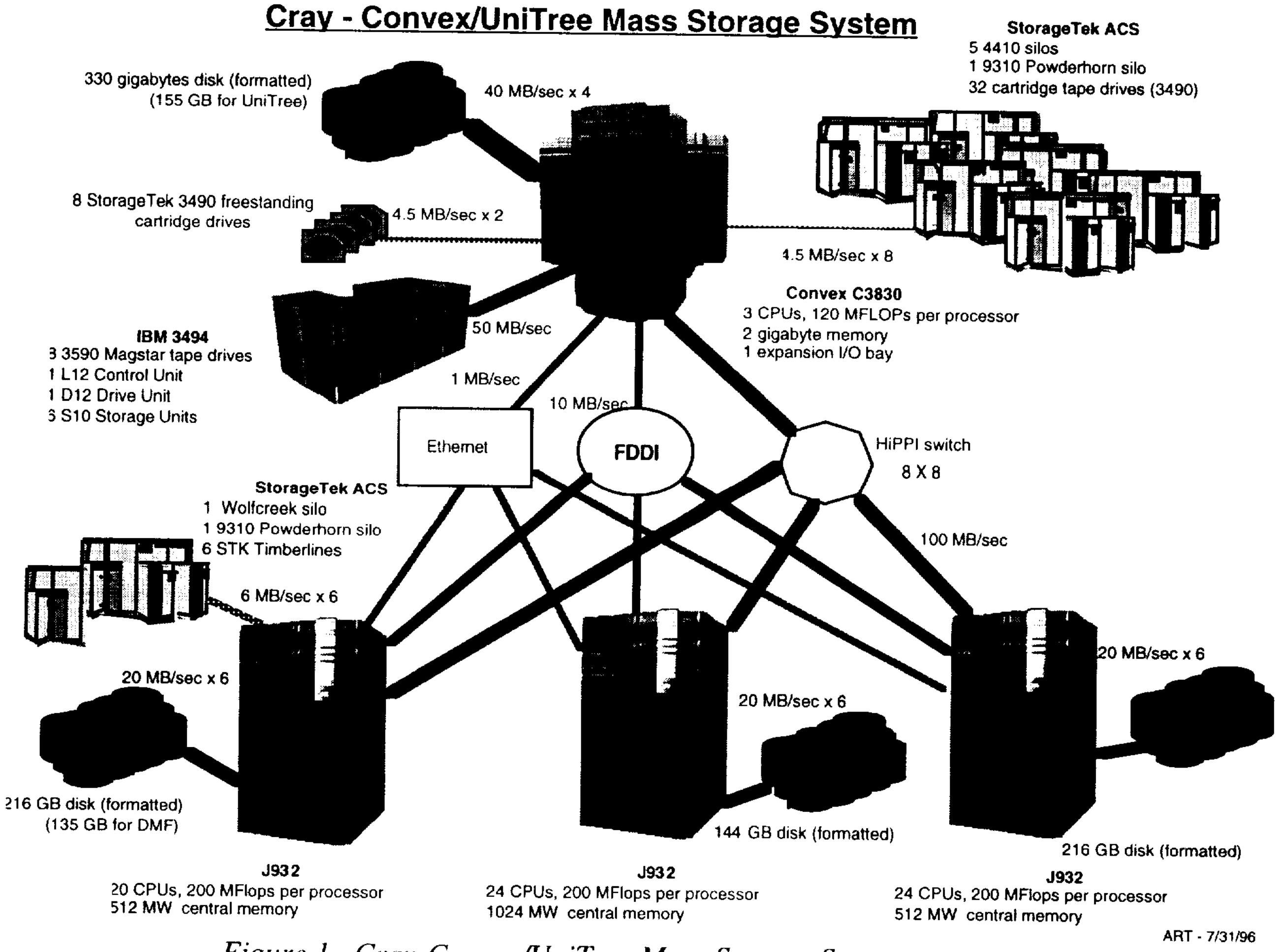


Figure 1. Cray-Convex/UniTree Mass Storage System

Requirements Input: Current MDSDS Usage Characteristics

Pentakalos et al. [1, 2] have studied and modeled the behavior of the NCCS UniTree system and Tarshish et al. [3, 4] have reported on its growth. In addition, ESDCD and NCCS staff maintain ongoing statistics to measure MDSDS usage and performance characteristics. The following characteristics are derived from ongoing observations and from some related studies.

Network Load

Much like a water plumbing system, a high-end storage and delivery system's performance must accommodate bursts that can be an order of magnitude or greater than rates averaged over time. For example, between June and August 1996, MDSDS aggregate network rates in excess of 17 MB/s have been observed in daily usage, whereas MDSDS average daily network traffic for June 1996 (71 GB/day) evenly distributed over an entire day would amount to 0.84 MB/s.

The middle of the working day is when the peak MDSDS request loads usually occur (cf. Pentakalos [2]) a characteristic common to other storage facilities (Behnke et al.[5]). It is especially typical for MDSDS traffic from sources other than NCCS supercomputers to peak during working hours. The non-supercomputer traffic to the MDSDS is significant (about 20-50 GB/weekday) but represents only about 23 % of all MDSDS traffic (*Figure 2*).

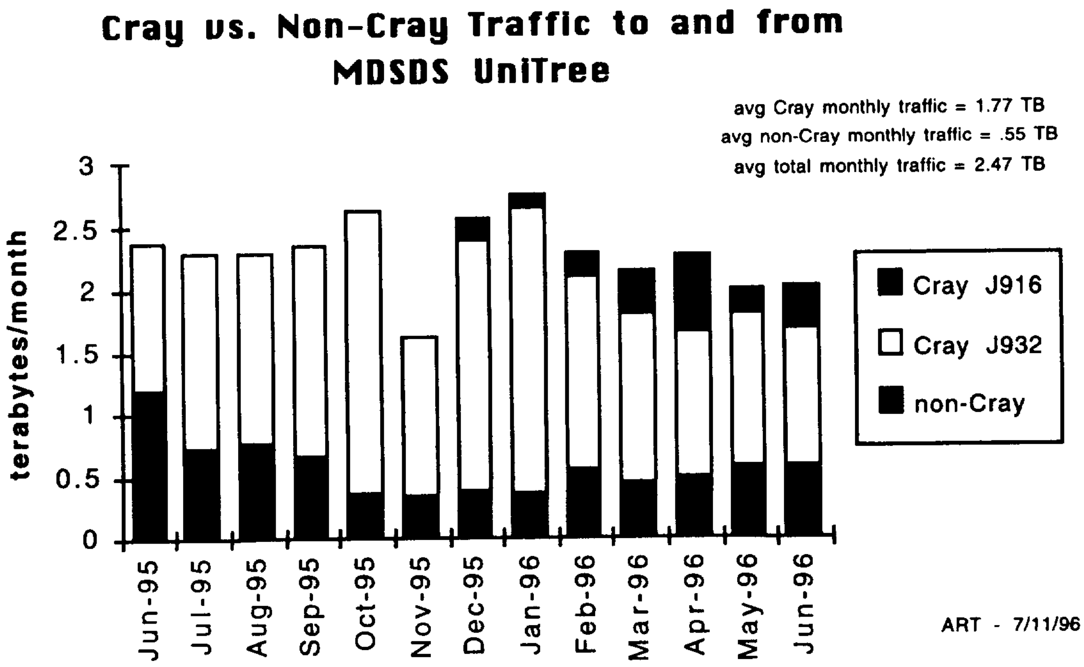


Figure 2. Cray vs. Non-Cray Traffic to and from MDSDS UniTree.

Relationship Between NCCS Supercomputing Resources Used and MDSDS Traffic

Historically, MDSDS transfer traffic has scaled approximately linearly with the increase in NCCS supercomputing CPU power once the user community has had time to adjust to new supercomputing technologies (*Figure 3*):

- In November 1993, shortly after delivery of the 6-processor Cray C90 (2.5 times the CPU power of previous 8-processor Cray YMP), a 30-day average for MDSDS traffic was ~36 GB/day.

- In late January 1996, Cray C90 usage was at its peak, and MDSDS traffic 30-day average reached ~101 GB/day.

The replacement of the 6 Cray C90 processors (1 GFLOP each) with 48 Cray J90 processors that were slower (0.2 GFLOP each) but much more plentiful resulted in some decrease in MDSDS traffic as users modified codes to improve throughput in the more parallel supercomputer environment. At this writing the third Cray J90 (20 additional processors) has been in place only a short time, but there are indications that MDSDS - supercomputer traffic is on the increase.

In addition, storage growth rates have tended to increase even if supercomputer CPU capacity stays the same for an extended period once users become accustomed to new supercomputing paradigms. As long as there are sufficient resources (both in available supercomputing capacity and in budgetary support), researchers tend to increase the resolution or complexity of models. In addition, the largest-volume users of the supercomputers tend to make ongoing refinements to code to maximize throughput.

*FY96 H2 is projected from first four months of the half

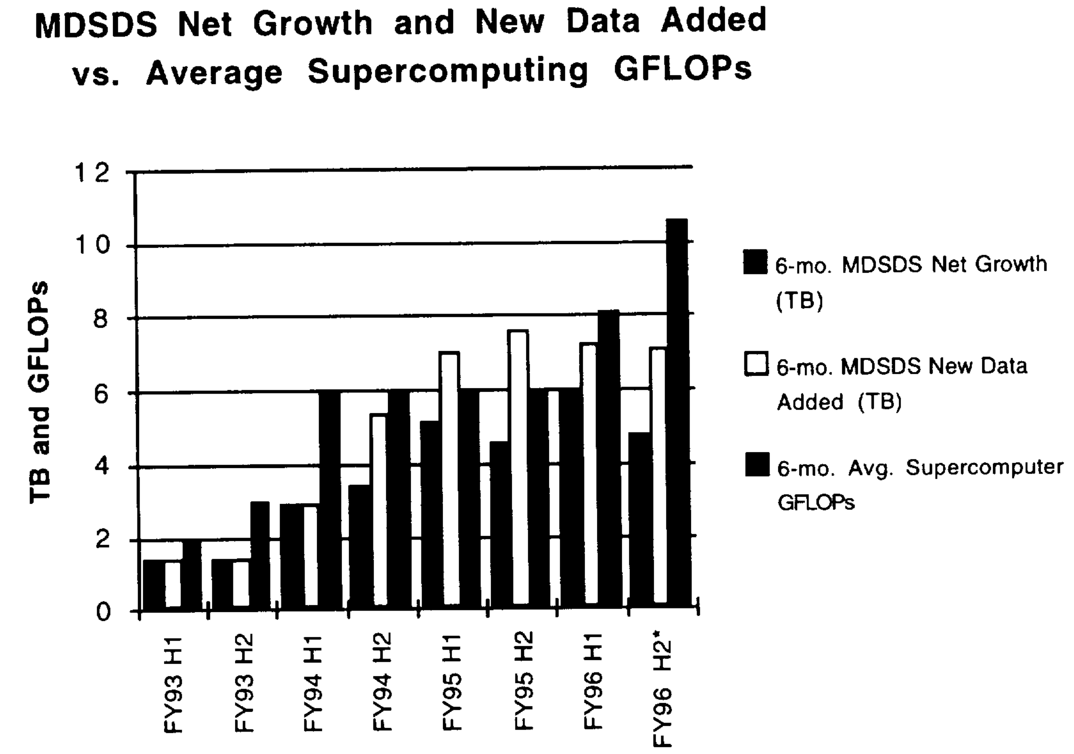


Figure 3. MDSDS Net Growth and New Data Added vs. Average Supercomputing GFLOPs.

Not a Black Hole for Storage: Retrieval Traffic Is Significant

As illustrated in the figures below, retrieve traffic is significant, so the MDSDS must be able to retrieve files quickly and efficiently. Clearly, the NCCS cannot afford to tune the MDSDS to optimize writing/storing at the expense of poorer performance for reading/retrieving files.

- NCCS users retrieve old files, not just recently created files (*Figure 4*).

- MDSDS users retrieve nearly as much data as they store (*Figure 5*). About 1.5 million MDSDS files were transferred between August 1995 and July 1996. 53% of the bytes transferred were stored, and 47% of the bytes transferred were retrieved. However, 62% of the *files* transferred were newly stored, and 38% were retrieved. This implies that on average, files retrieved are larger than files stored (averaging 24.8 MB and 17.1 MB respectively) and suggests that smaller files are somewhat less likely to be retrieved.

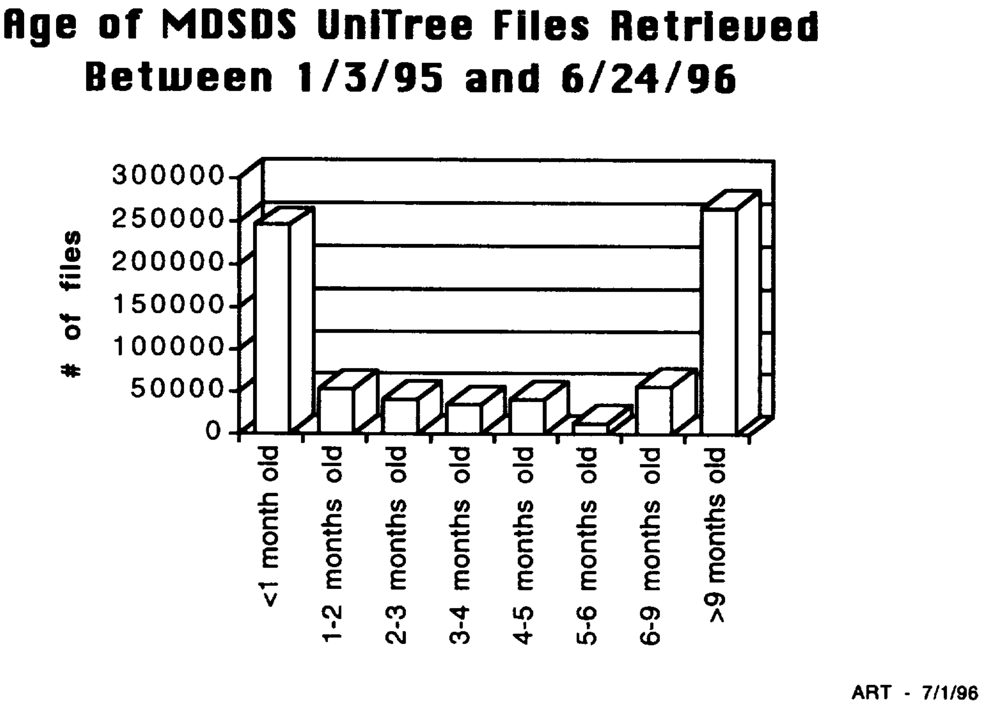


Figure 4. Age of MDSDS UniTree Files Retrieved Between 1/3/95 and 6/24/95.

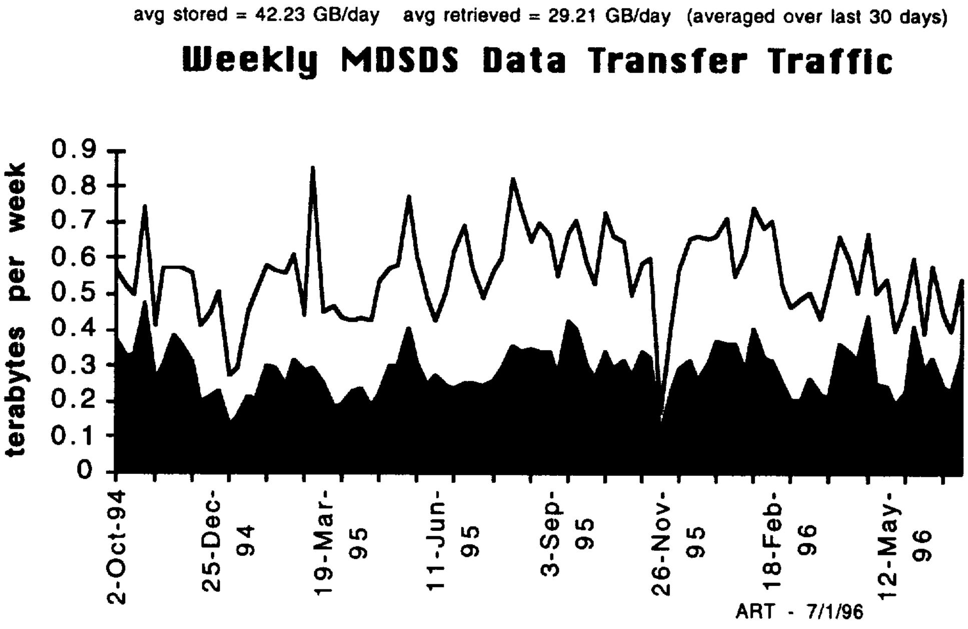


Figure 5. Weekly MDSDS Data Transfer Traffic.

Working Set and Temporal Locality

The 3 months of MDSDS activity studied by Pentakalos [2] exhibited little temporal locality in the working set of the MDSDS. Six months during the period July 1995 through April 1996 were examined for the current study. The one-month working set during this period averaged about 1.8 TB, while total MDSDS traffic over one month averaged 2.6 TB.

- Of the 1.34 TB new data created in a month, on average only 0.25 TB (less than 1/5) would be retrieved in that same month; the remaining 0.48 TB of unique data retrieved would be more than 1 month old.

- On average 1.3 TB would be retrieved in a month; of that, 0.73 TB would be unique, so on average a given byte retrieved in a month would have been retrieved twice.

- However, 2/3 of the 0.73 TB unique data retrieved was retrieved only once, so the remaining 0.24 TB was responsible for all the repeated retrieves (0.86 TB of traffic). *Figure 6* shows the average proportions of bytes and files retrieved repeatedly over a 1-month period.

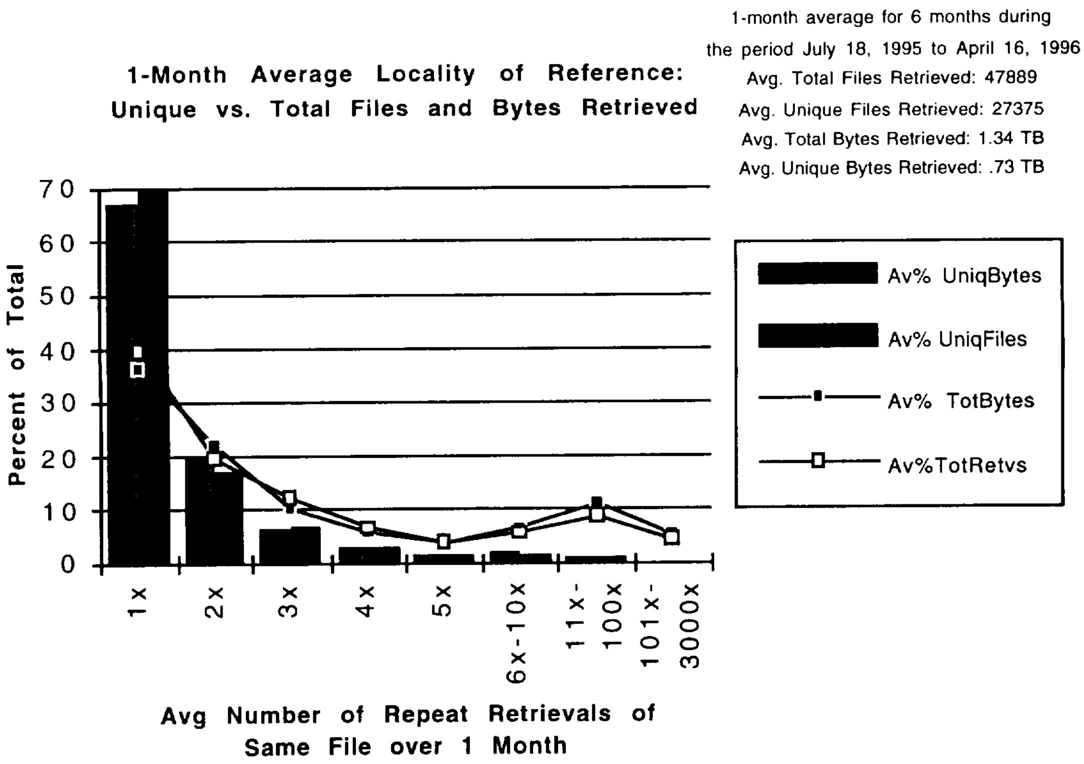


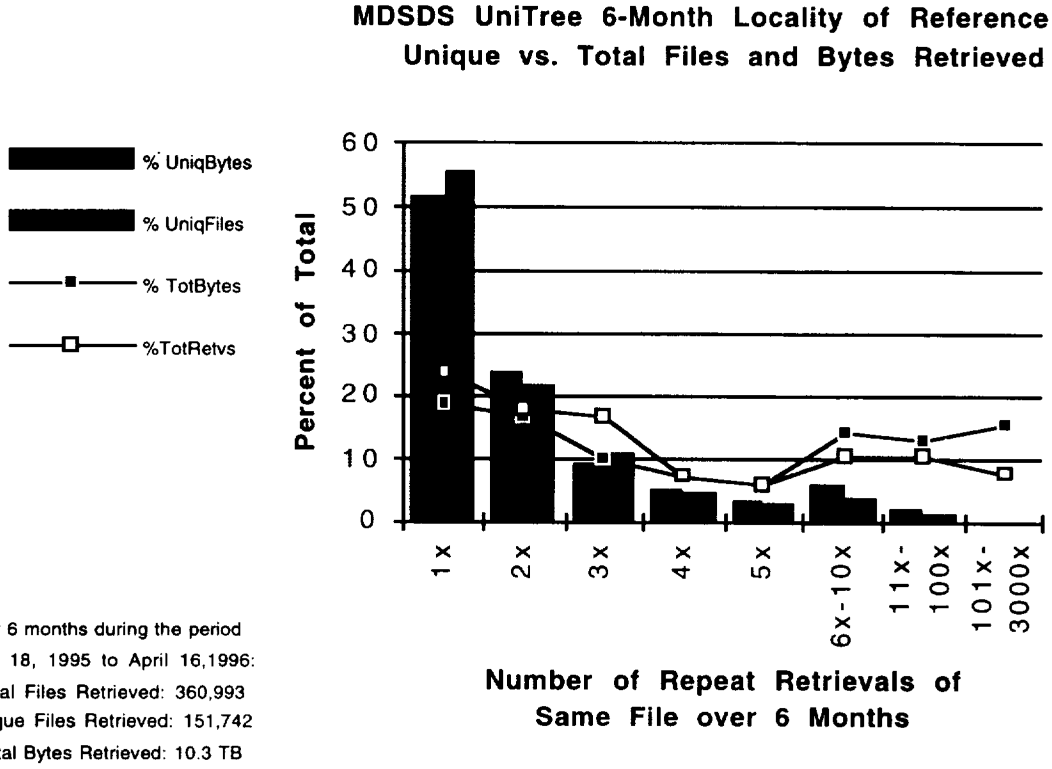
Figure 6. MDSDS 1-Month Average Locality of Reference: Unique vs. Total Files and Bytes Retrieved.

The likelihood that a file had been retrieved more than once increased when the time examined is expanded to six months: of the 10.3 TB retrieved during 6 months between

August 1995 and April 1996, 3.6 TB was unique data. The entire working set over 6 months was 9.5 TB. On average, at the time of retrieval:

- 1.2 TB (33%) of unique data was one month old or younger.
- 2.1 TB (58%) of unique data was 6 months old or younger; 1.5 TB (42%) of unique data was older than 6 months.

Figure 7 shows relative proportions of files and bytes retrieved repeatedly over the 6 months examined.



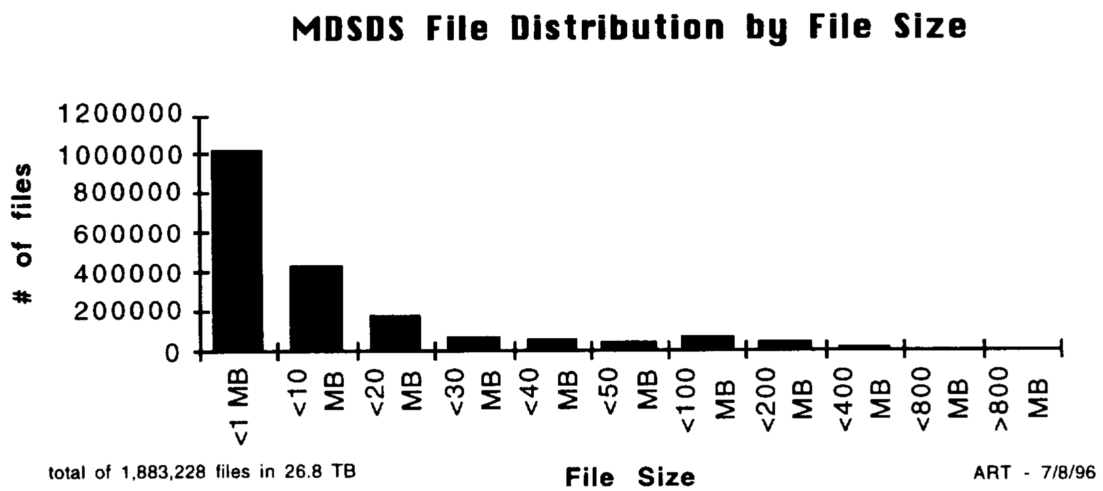
For 6 months during the period July 18, 1995 to April 16,1996:
Total Files Retrieved: 360,993
Unique Files Retrieved: 151,742
Total Bytes Retrieved: 10.3 TB

Figure 7. MDSDS UniTree 6-Month Locality of Reference: Unique vs. Total Files and Bytes Retrieved.

The MDSDS UniTree system's disk cache is configured to favor retaining most recently used files on disk, but at present there are no other reasonable means to set aside portions of the disk cache for special purposes (e.g., areas for files retrieved vs. files stored vs. files being moved to different tapes to consolidate free space on tapes). As a result, the cost to accommodate a RAIDed disk cache that holds six months' or even one month's working set would be prohibitive for the current machine architecture in the current budget climate. However, the re-use patterns may bear further to study to explore whether the near-online storage could be organized to optimize retrieval of the most frequently used data.

Distribution of File Sizes: Number of Files vs. Number of Bytes

While nearly a million *files* stored in the MDSDS system are 1 MB in size or smaller, the vast majority of the *data* stored in the MDSDS is in files of 50 MB and larger (*Figure 8, Figure 9*). This suggests that storage media with poor stop/start performance would not be suitable for the large number of small files, unless the file management software can compensate for small files, e.g., by grouping many small files together when writing, or by automatically directing small files to different media.

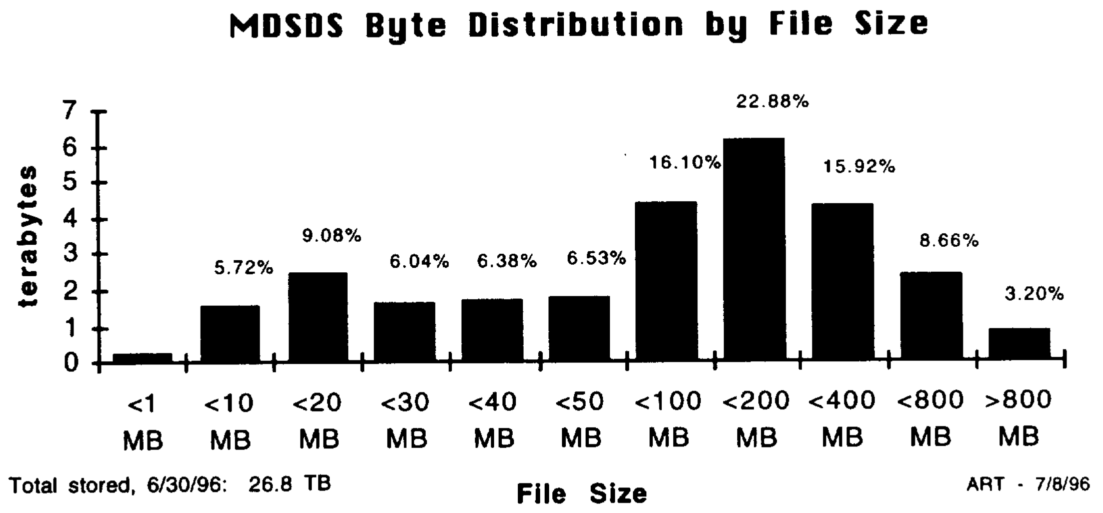


MDSDS File Distribution by File Size

total of 1,883,228 files in 26.8 TB

ART - 7/8/96

Figure 8. MDSDS File Distribution by File Size.



MDSDS Byte Distribution by File Size

Total stored, 6/30/96: 26.8 TB

ART - 7/8/96

Figure 9. MDSDS Byte Distribution by File Size.

Repacking Is a Way of Life

MDSDS users delete nearly half a terabyte of data per month, an amount that approaches 1/3 the quantity of new data stored (*Figure 10*). Consequently, repacking (consolidating files from partly empty tapes to free those tapes for re-use) is a crucial activity The I/O load from this tape repacking is significant and must be factored into the performance of the MDSDS system. Under the MDSDS's current release of UniTree software (Convex UniTree+ 2.0), at least 4 bytes must be moved for each byte copied from a tape being freed by repacking.

In addition, repacking is also the mechanism used to move MDSDS files to new storage media. This evolution to newer, more dense media is essential in order to (1) accommodate increasing volumes of new data and (2) ensure that older files continue to be readable as storage technology advances. In the 4 years since the MDSDS was deployed, there have already been 3 rounds of repacking to migrate files from 3480 cartridges (200 MB each) to 3490 cartridges (400 MB) to 3490E cartridges (800 MB). With the recent arrival of IBM Magstar tape drives, a new round of repacking will move some MDSDS data to the 10-GB IBM 3590 cartridges.

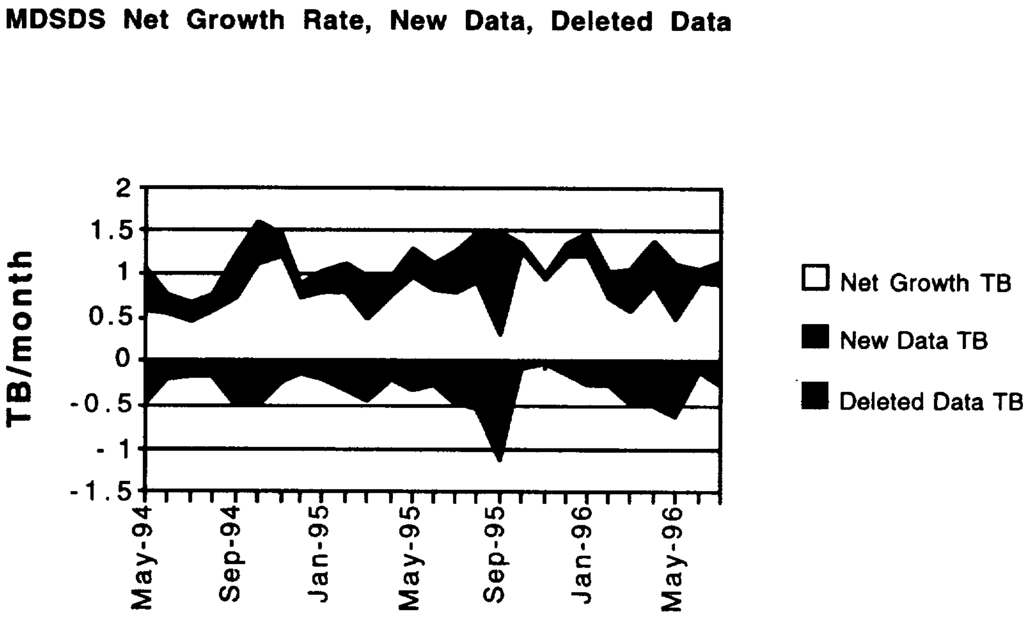


Figure 10. MDSDS Net Growth Rate, New Data, and Deleted Data.

Requirements Input: Some Projections for the Future

Several different sources have provided projected requirements information:

Near- and Medium- Term Research Program Changes:

The following changes are expected to have significant impact on MDSDS growth and volume of data transferred:

- Climate Data Assimilation researchers have announced plans to increase the climate model's vertical resolution and to save more diagnostic output starting in Fall 1996. These changes are expected to quadruple the amount of data produced by a climate model integration run.

- In FY98, the Climate Data Assimilation production work is currently slated to move off the NCCS supercomputers and onto EOSDIS-sponsored platform(s).

- A new Ocean Data Assimilation effort with projected requirements similar to Climate Data Assimilation production (CERRC [6]) is expected to begin (e.g., FY99 requiring 125 GFLOPs sustained and generating ~200 TB/year).

The NCCS MDSDS also anticipates providing long-term storage for the High Performance Computing and Communications (HPCC) program's Earth and Space Science (ESS) Cooperative Agreement effort, which will run through 1998-1999 (URL: http://nccsinfo.gsfc.nasa.gov/ESS/). Current plans show the Goddard Testbed machines and MDSDS connected via HiPPI, and perhaps later, via an ATM-to-HiPPI switch. ESS Grand Challenge Investigators who use those scalable parallel Testbed machines to be sited at Goddard are expected to store about 30 TB in the MDSDS system over the 3 years of the project.

Mission to Planet Earth Science-User Survey Results

In mid-1996, the Office of Mission to Planet Earth sent a survey to NASA Earth science researchers to help clarify the resources needed for computing and numeric modeling capabilities over the 5 years spanning 1997-2001. More than 200 researchers responded. Among the findings

- 35% wanted to double their model resolution (implies at least 4-fold increase in model output produced).

- 6% wanted to increase their model resolution by an order of magnitude (could easily increase model output by more than 2 orders of magnitude).

- 88% had requirement for access to high-speed networking and/or mass storage

NCCS Science Research Users: the CERRC Report

In January 1995 the Computer Environments and Research Requirements Committee (CERRC) reported on NCCS Earth and space science computing requirements for the years 1997-2004 after gathering information from researchers who use NCCS resources (CERRC [6]). The CERRC obtained input from both Goddard- and NASA-based researchers, and those at universities and non-NASA institutions (the latter use about

27% of NCCS resources at present). The CERRC report predates some of the recent Federal budget reduction exercises, and it describes computing requirements that presume the science investigations would be funded at reasonably favorable levels.

Briefly, the CERRC report relates the need for supercomputing CPU performance of 125 GFLOPs sustained by 1999 and 1 TFLOP sustained by 2004. The space and Earth Sciences research codes that drive these CPU performance requirements are expected to generate the need for 400 TB in robotic storage in 1997, and 2000 TB robotic storage by 1999.

NCCS Planned Supercomputer Upgrades

In response to the requirements detailed in the CERRC report and budget direction from management, the NCCS currently expects to make available to users increased supercomputing CPU power along the following lines:

- FY97: complete the 3-fold increase (to 19.2 GFLOPs) compared to FY95's 6 GFLOP Cray C90.
- FY99: an additional increase of nearly 5-fold to ~90 GFLOPs.
- FY00 and beyond: continuing incremental augmentations, as budgets and supercomputing technology costs allow.

Storage and Network Requirements

The CERRC report and ongoing observations of MDSDS usage contained the most concrete information on trends and projections, and so are the primary sources for the requirements presented here. As with all projections for the future, the results are only as good as the assumptions and initial data, so the NCCS monitors usage, program direction, and industry to revise and update the picture. Caveats aside, the results from requirements-projecting exercises have proven useful in resource and budget planning.

Sustained network rate requirements were derived directly from the CERRC report's [6] data traffic figures cited for the milestone years (500 GB to 1 TB daily data traffic in FY97 and 1-2 TB daily traffic by FY99). In FY97, this leads to the need for sustained bandwidth of 6-12 MB/s; FY 99 would require 12-24 MB/s sustained. Peak network bandwidth requirements incorporated the empirically observed need to accommodate burst rates an order of magnitude higher than sustained loads.

Growth rates for interim years were calculated 2 ways: in the conservative method, the growth rate is tightly coupled to the supercomputing power, and so remains constant in a year (such as FY98) in which there are no supercomputer augmentations. The "heavier traffic" estimate assumes that growth rates will continue to increase in the interim years (due to increases in resolution, complexity, and optimization for throughput, as noted above).

Total data stored (based on conservative and heavy growth rates) and the robotic capacity recommendations from the CERRC report are presented in *Figure 11.* Projections for growth rates and peak bandwidth are presented alongside expected increases in supercomputer power in *Figure 12.*

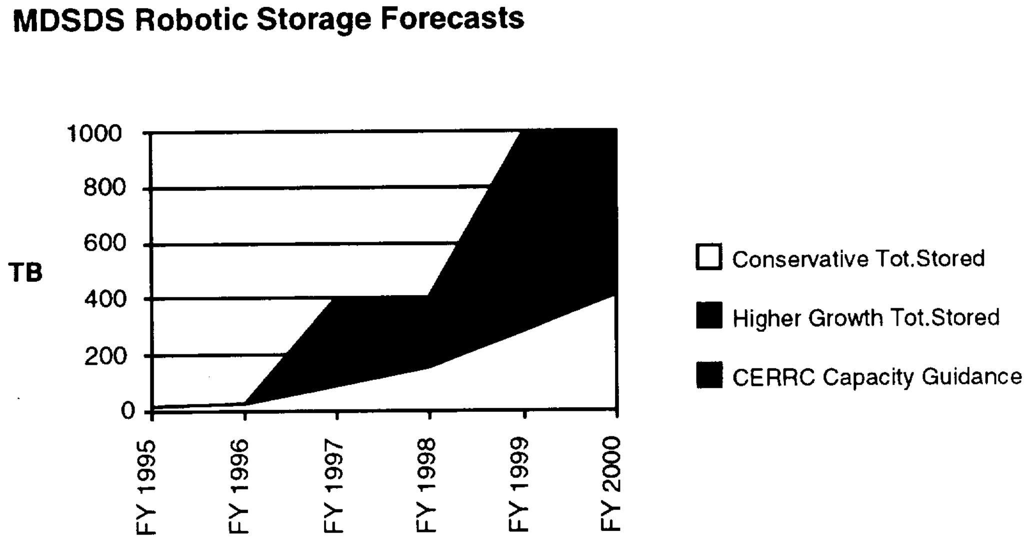


Figure 11. MDSDS Robotic Storage Forecasts.

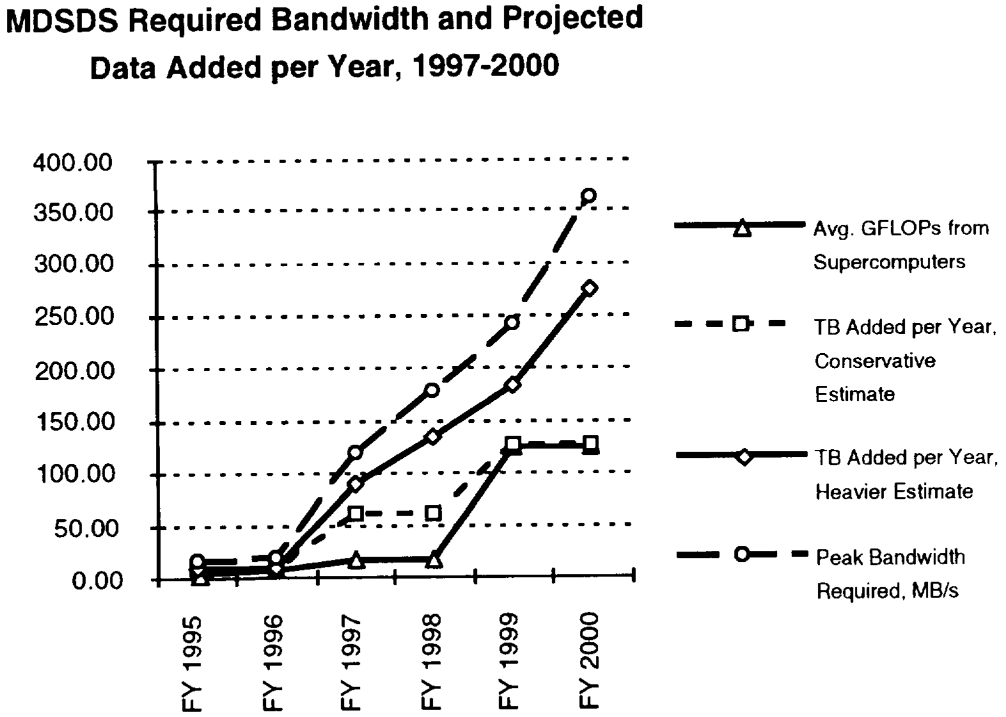


Fig. 12 MDSDS Required Bandwidth and Projected Data Added per Year, 1997-2000.

Acknowledgments

Many thanks to Adina Tarshish for her hard work and tenacity in gathering MDSDS statistics and creating charts. Thanks also to Nancy Palm, Michele Rienecker, Tom Schardt, Bob Theis, Pat Gary, and the authors of the CERRC report for their efforts at synthesizing abstract plans into concrete requirements.

References

[1] O. Pentakalos and Y. Yesha, "Evaluating the Effect of Online Data Compression on the Disk Cache of a Mass Storage System," *Proc. of the Fourth Goddard Conference on Mass Storage Systems and Technologies,* pp. 383-391, 1995.

[2] O. Pentakalos, D. Menasce, M. Halem, and Y. Yesha, "An Approximate Performance Model of a UniTree Mass Storage System," *Proc. of Fourteenth IEEE Symposium on Mass Storage Systems,* pp. 210-224, 1995.

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

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

[5] J. Behnke and J. King, "NSSDC Provides Network Access to Key Data via NDADS," *Proc. of Fourth Goddard Conference on Mass Storage Systems and Technologies,* pp. 359-381,1995.

[6] The Computer Environments and Research Requirements Committee, "NASA Earth and Space Science Computing Requirements 1997-2004," 56 pp., 1995.