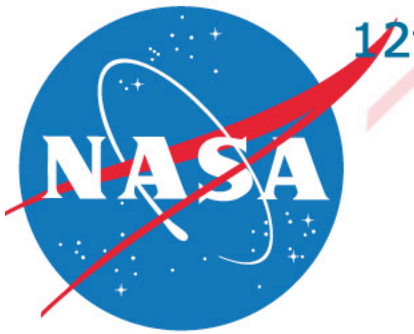
Hierarchical Storage Management at the NASA Center for Computational Sciences: From UniTree to SAM-QFS

Ellen.M.Salmon@nasa.gov
Science Computing Branch
Earth and Space Science Computing Division
NASA Goddard Space Flight Center Code 931



NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

NCCS's Mission and Customers

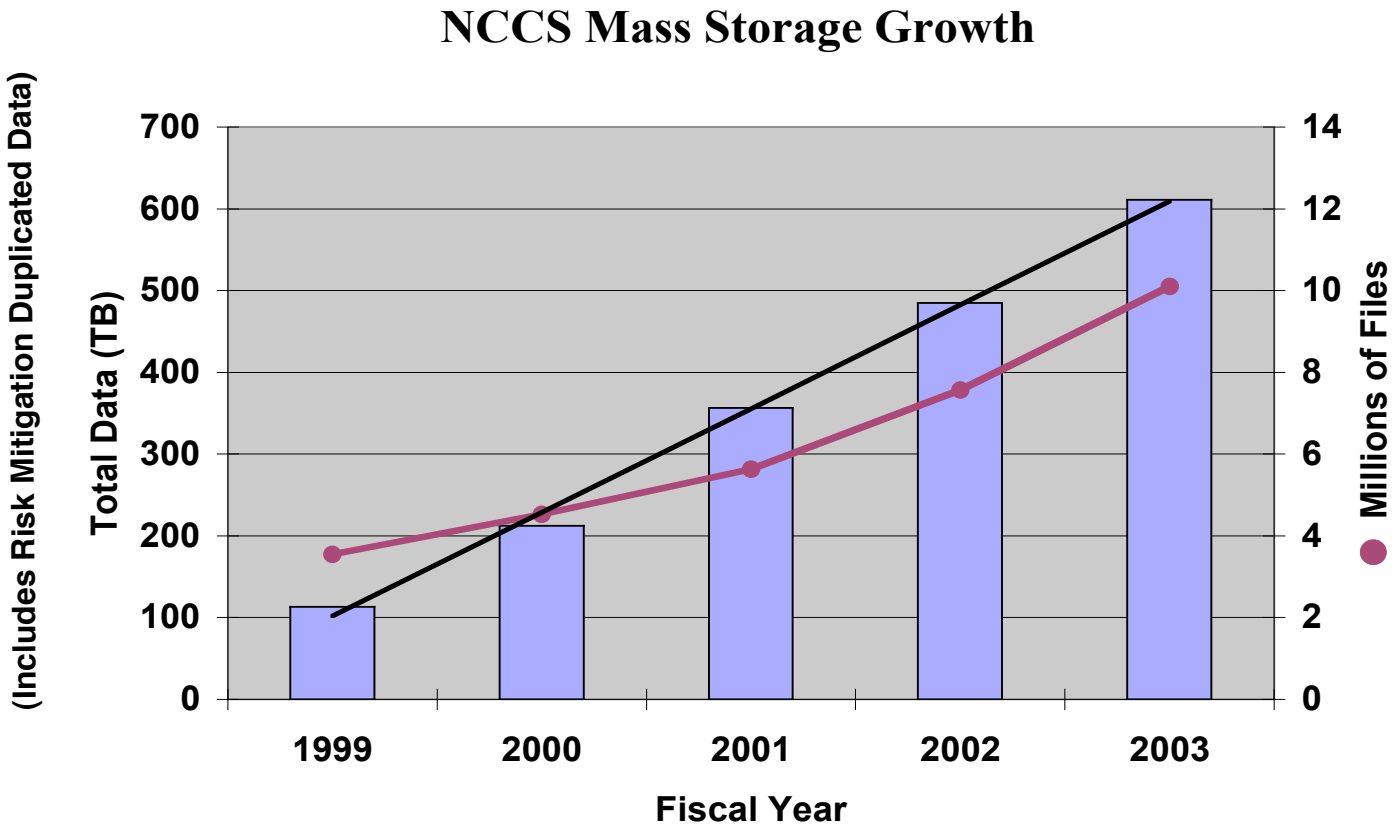
- NASA Center for Computational Sciences (NCCS) at NASA Goddard Space Flight Center
- Mission: **Enable Earth and space sciences research** (via data assimilation and computational modeling) by providing state-of-the-art facilities in
  - **High Performance Computing (HPC)**,
  - **Mass Storage**
  - **High-speed Networking**
  - **HPC Computational Science Expertise**
- Earth and space science customers:
  - Seasonal-to-interannual climate and ocean prediction
  - Global weather and climate data sets incorporating data assimilated from numerous land-based and satellite-borne instruments

NCCS Resources

- High Performance Compute Engines
  - HP Compaq ES 45 Alphaserver SC (1392p)
  - SGI Origin 3800s (608p total)
  - ~4.5 TFLOPs peak total
- Mass Storage Systems and Servers
  - Mass Data Storage and Delivery System (MDSDS), was UniTree now SAM-QFS, on Sun Fire 15K, 2 domains, Shared QFS "HA"
    - ~355 TiB*, ~12M files, DDN S2A 8000 disk arrays
  - SGI DMF, Origin 3800 server, to be converted to SAM-QFS via "DMS" (Data Management System, based on SRB)
    - ~350 TiB*, ~14M files, HDS 9960 and SGI TP 9x00 disk arrays
- Tape Libraries, Intra-Machine/Device Networks, Switches
  - Nine STK Powderhorn tape libraries (~51,000 slots)
    - STK 9840C, STK 9940B, STK 9840A tape drives
  - Gigabit Ethernet, Foundry BigIron 15K
  - 2-Gb Fibre Channel, Brocade Silkworm 12K, 3900s

*** Unique** (does not include risk mitigation duplicates)

NCCS Mass Data Storage and Delivery System (MDSDS) Growth



NCCS Projected Requirements

- Earth and Space Science drivers: increasing
  - Model resolution
  - Number of assimilated observations
  - Numbers of concurrent model ensembles
- Total data held (including risk-mitigation duplicates):
  - Current: ~1.5 PiB
  - End of FY 2005: ~6 PiB
  - End of FY 2007: ~19 PiB
- Files (unique):
  - Current: ~25 million, grows ~33% per year
  - FY 2005: ~44 million
  - FY 2007: ~78 million

HSM Evaluation (Spring 2002)

- High-end HSM vendors' responses to 60-some technical questions
  - SGI's DMF
  - Sun's SAM-QFS
  - Legato's DiskXtender (UniTree Central File Manager)
  - IBM's HPSS
- NCCS and CSC team evaluated:
  - Performance
  - Integrity, High Availability
  - Scalability, Modularity, Flexibility
  - Balance (avoiding bottlenecks)
  - Manageability

UniTree to SAM-QFS Migration

- Employed Sun's SAM migration toolkit and migration libraries written by Instrumental, Inc.
  - Legacy UniTree directory and file "inode" info harvested, then inserted into SAM-QFS filesystems
    - Legacy UniTree: ~300 TB, ~11M files
    - Only 5 days downtime, including QC checks and server recabling (~11M files, ~300K directories)
  - Transparent user retrieval of legacy files: SAM sees UniTree files as "stranger" media, so reads files via migration library, then archives to SAM tapes
    - Approach requires UniTree system to read legacy files/media
  - Background migration: via DQDuffy et al. Perl script, UniTree files pre-staged tape by tape for efficiency

Strong Benefits in NCCS's Current SAM-QFS Configuration

- Performance observed in daily use: over 10 TB/day archived while handling 2+TB/day user traffic
- Shared QFS works well to make the underlying cluster appear as a single entity
- Using "HA flip-flop" for significant software upgrades has greatly reduced downtime for significant software upgrades
- A test cluster system has been invaluable
- Restoring files after accidental deletions much simpler/faster than previous solution

Lessons Learned

- Complexities of clustered HSM systems make configuration of automated high-availability software challenging
- The “Release Currency Conundrum”:
  - Software release’s newest features will be the most immature
  - Keeping current on OS and HSM patches can help to avoid significant pitfalls
- Make “risk mitigation” duplicate tape copies
- Keep your expectations of vendors high
  - Great support/cooperation from Sun in getting “Traffic Manager” (a.k.a mpxio) to work with 3rd party Fibre Channel RAID array (DataDirect Networks S2A 8000)

Background Detail

The Large Team

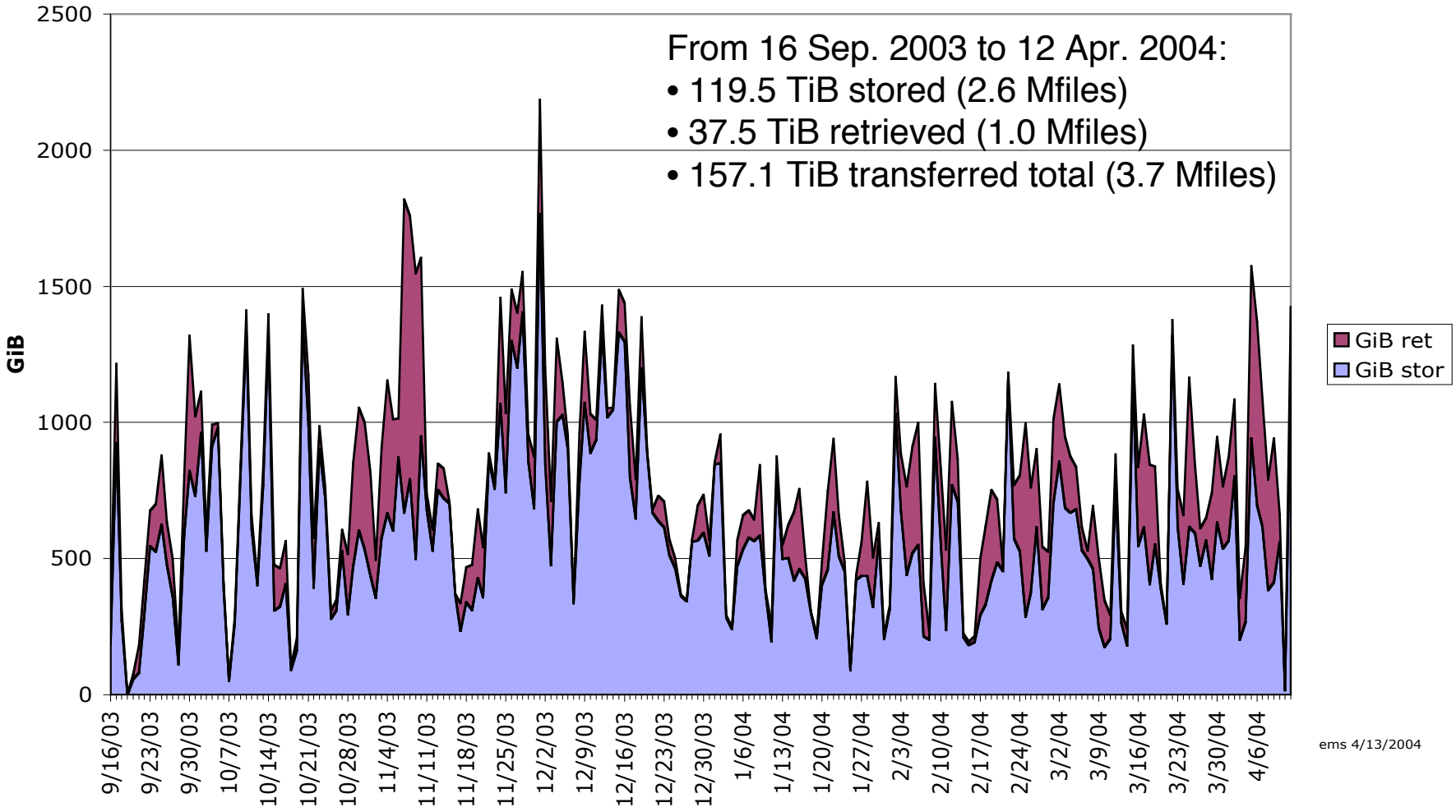
Ellen Salmon, Adina Tarshish, Nancy Palm, Tom Schardt
NASA Center for Computational Sciences (NCCS)
NASA Goddard Space Flight Center Code 931
Greenbelt, Maryland 20071
Ellen.M.Salmon@nasa.gov
Adina.R.Tarshish@nasa.gov
Nancy.L.Palm@nasa.gov
Thomas.D.Schardt@nasa.gov

Sanjay Patel, Marty Saletta, Ed Vanderlan,
Mike Rouch, Lisa Burns, Dr. Daniel Duffy, Roman Turkevich
Computer Sciences Corporation,
c/o NASA GSFC Code 931
Greenbelt, Maryland 20071
sjpatel@calvin.gsfc.nasa.gov
marty@nccs.gsfc.nasa.gov
evanderl@csc.com
mrouch@sanz.com
lburns3@csc.com
dqduffy@pop900.gsfc.nasa.gov
rturkevi@csc.com

Robert Caine, Randall Golay,
Craig Flaskerud, Linda Radford,
Matt Hatley
Sun Microsystems, Inc.
7900 Westpark Drive
McLean, VA, 22102
Email:
Robert.Caine@sun.com
Randall.Golay@sun.com
Craig.Flaskerud@sun.com
Linda.Radford@sun.com
Matt.Hatley@sun.com

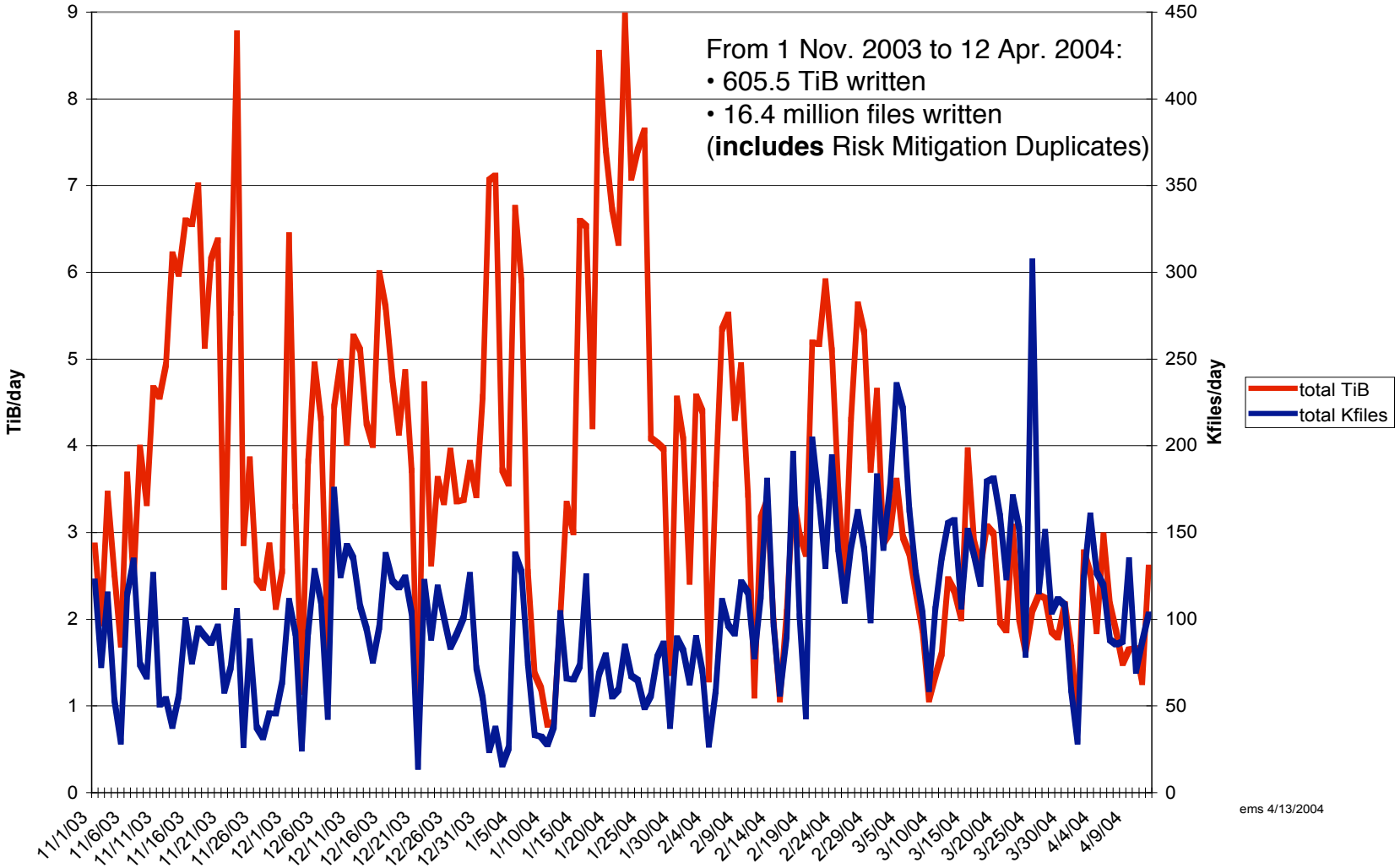
Jeff Paffel, Nathan Schumann
Instrumental, Inc.
2748 East 82nd Street
Bloomington, MN 55425
jpaffel@instrumental.com
nds@instrumental.com

NCCS MDSDS SAM-QFS User Transfer Traffic



ems 4/13/2004

NCCS MDSDS SAM-QFS Daily Tape Write Activity



ems 4/13/2004

The Future (1)

- Further optimize data placement on tape to favor data retrieval
  - Issue: adequately characterizing retrievals?
- Explore SATA disk as the most nearline part of the HSM hierarchy
  - NCCS data retrieval profile make this somewhat problematic
  - But becomes more attractive as time-to-first-data rises on growing-capacity tape
  - Not expected to *replace* tape any time soon
- National Lambda Rail participation: enable large scale, long distance science team collaboration

The Future (2): Data Management System

- Goal: help users ***manage*** their data
- Based on San Diego Supercomputer Center's Storage Resource Broker (SRB) middleware, system developed by Halcyon Systems, Inc.
- Replaces file system access
- Allows for extremely useful metadata and queries,for monitoring and management,e.g.
  - File content and provenance
  - File expiration
- Allows for transparent (to user) migration between underlying HSM

References

[1] http://nccs.nasa.gov.

[2] *Performance Management at an Earth Science Supercomputer Center*, Jim McGalliard and Dick Glassbrook.

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

[4] *Mass Storage System Upgrades at the NASA Center for Computational Sciences,* A. Tarshish, E. Salmon, M. Macie, and M. Saletta, Proceedings of the Eight NASA Goddard Conference on Mass Storage Systems and Technologies, Seventh IEEE Symposium on Mass Storage Systems, (2000) pp. 325-334.

[4] *UniTree to SAM-QFS Project Plan*, Jeff Paffel, Instrumental, Inc., NCCS internal report.

[5] *UniTree to SAM-QFS Migration Procedure*, Daniel Duffy, Computer Sciences Corporation, NCCS internal report.

[6] *Sun SAM-FS and Sun SAM-QFS Storage and Archive Management Guide*, August 2002; *Sun QFS, Sun SAM-FS, and Sun SAM-QFS File System Administrator's Guide.*

[7] http://www.npaci.edu/DICE/SRB/.

Standard Disclaimers and Legalese Eye Chart

- All Trademarks, logos, or otherwise registered identification markers are owned by their respective parties.
- Disclaimer of Liability: With respect to this presentation, neither the United States Government nor any of its employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.
- Disclaimer of Endorsement: Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. In addition, NASA does not endorse or sponsor any commercial product, service, or activity.
- The views and opinions of author(s) expressed herein do not necessarily state or reflect those of the United States Government and shall not be used for advertising or product endorsement purposes.