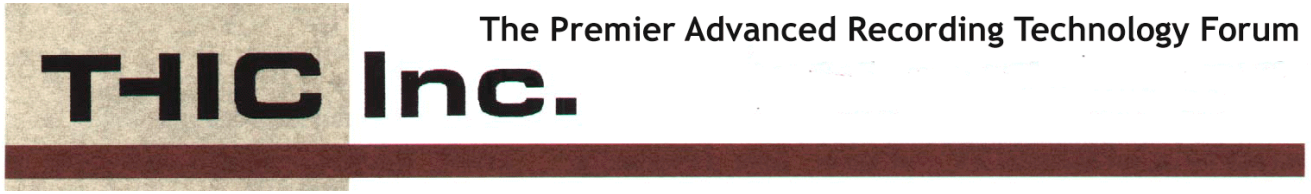
Large Scale Data Migration
at the
NASA Center for Computational Sciences

Ellen Salmon, Science Computing Branch
NASA Goddard Space Flight Center Code 931
Greenbelt, MD 20771
Phone:+1-301-286-7705 FAX: +1-301-286-1634
E-mail: Ellen.M.Salmon@nasa.gov

Presented at the THIC Meeting at the Raytheon ITS Auditorium
1616 McCormick Drive, Upper Marlboro MD 20774-5301
October 26-27, 2004



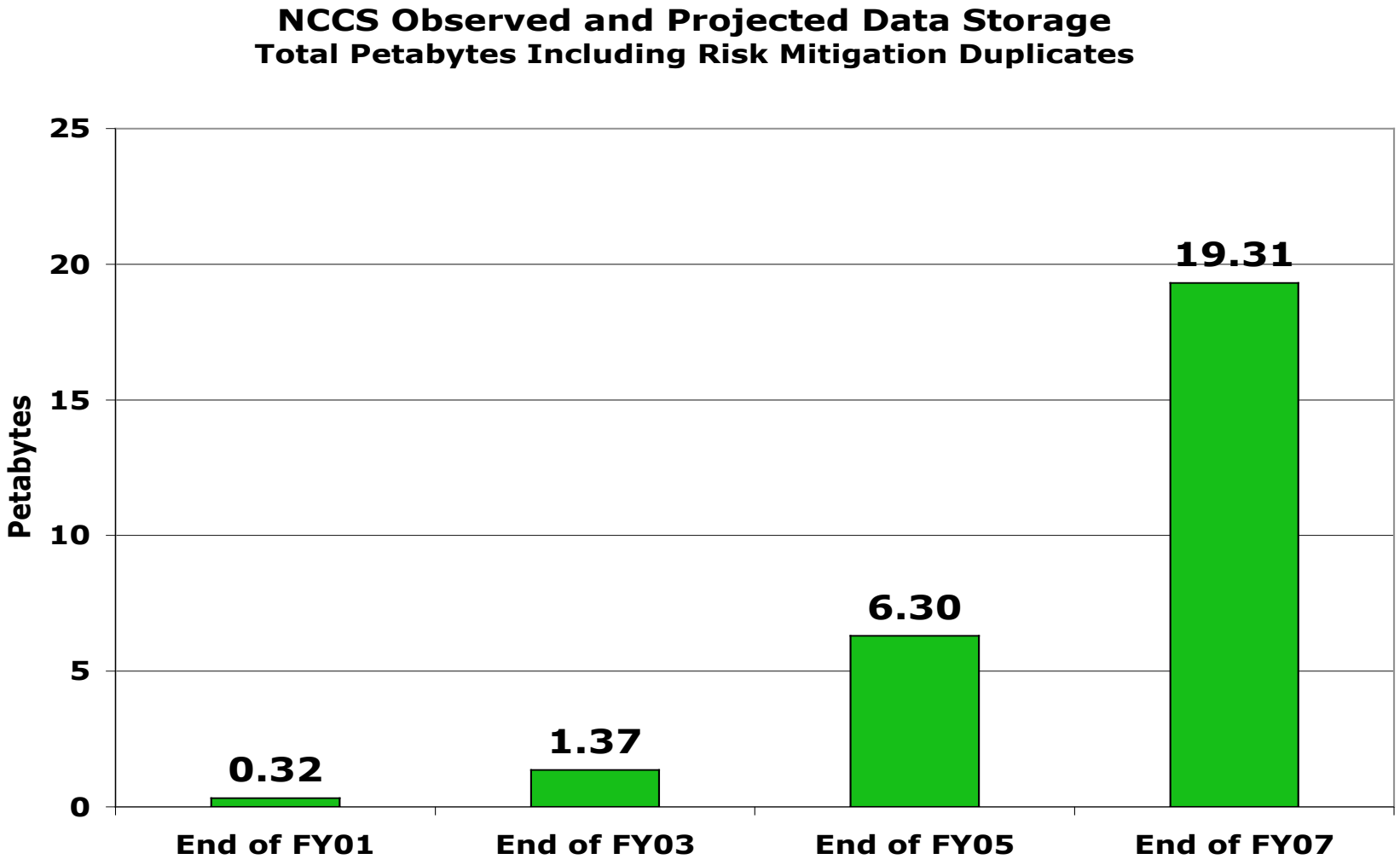
Standard Disclaimers and Legalese Eye Chart

- 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.

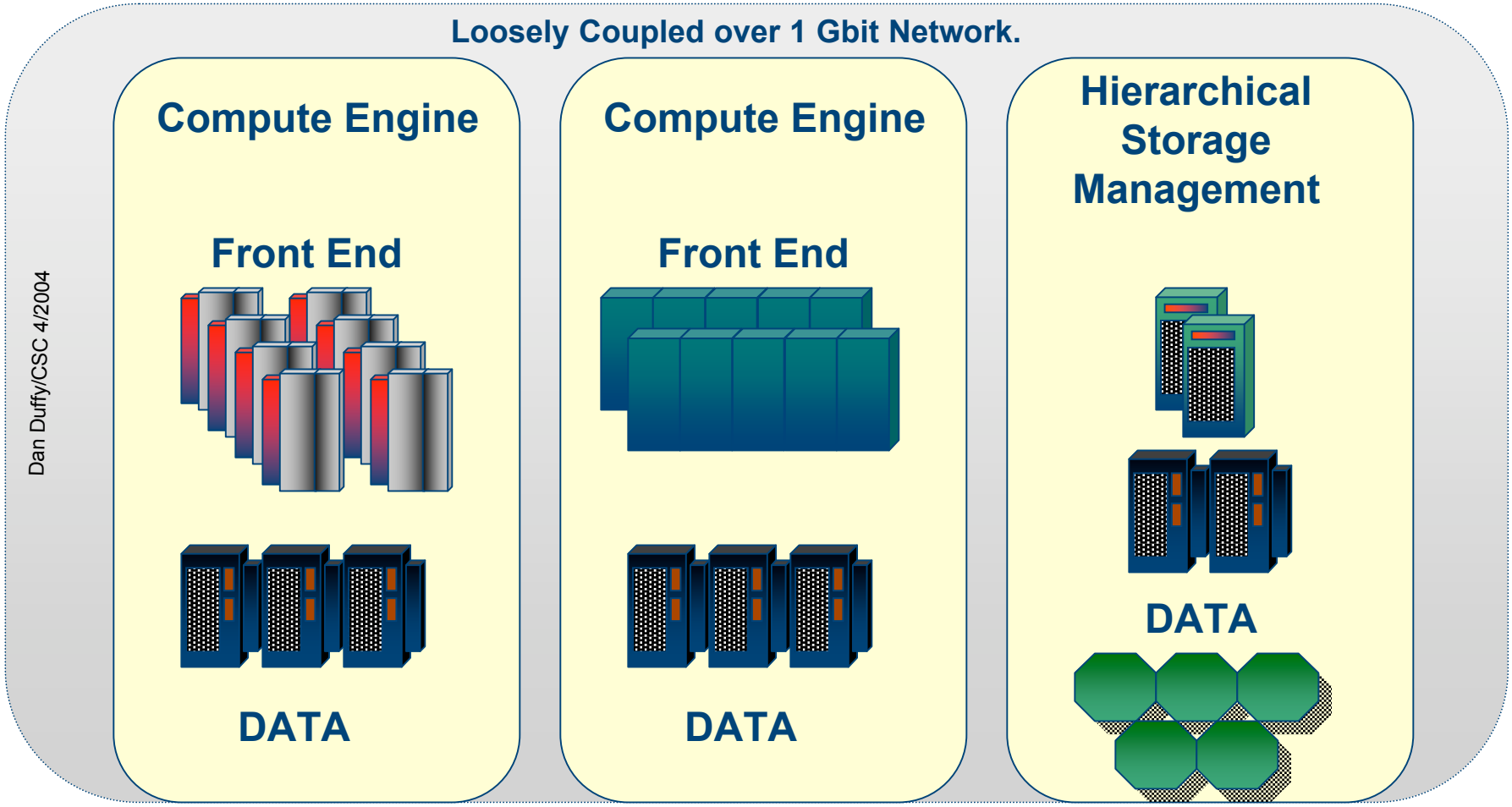
NCCS's Mission and Customers

- NASA Center for Computational Sciences (NCCS)
- 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 technologies, high-speed networking, and 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 User Requirements



Current NCCS Architecture (1)



Multiple interfaces and copies of data

Current NCCS Architecture (2)

Pros:

- Highly optimized storage resources for High-End Computing (HEC)
- Hierarchical Storage Management (HSM) system provides consolidated long term storage management

Cons:

- Local attached storage leads to multiple interfaces and copies of data

Current NCCS Architecture (3)

- Note that NCCS currently has *two* large HSM systems
- The MDSDS (Mass Data Storage and Delivery System) will be the focus here
  - The recent UniTree to SAM-QFS migration occurred on this system

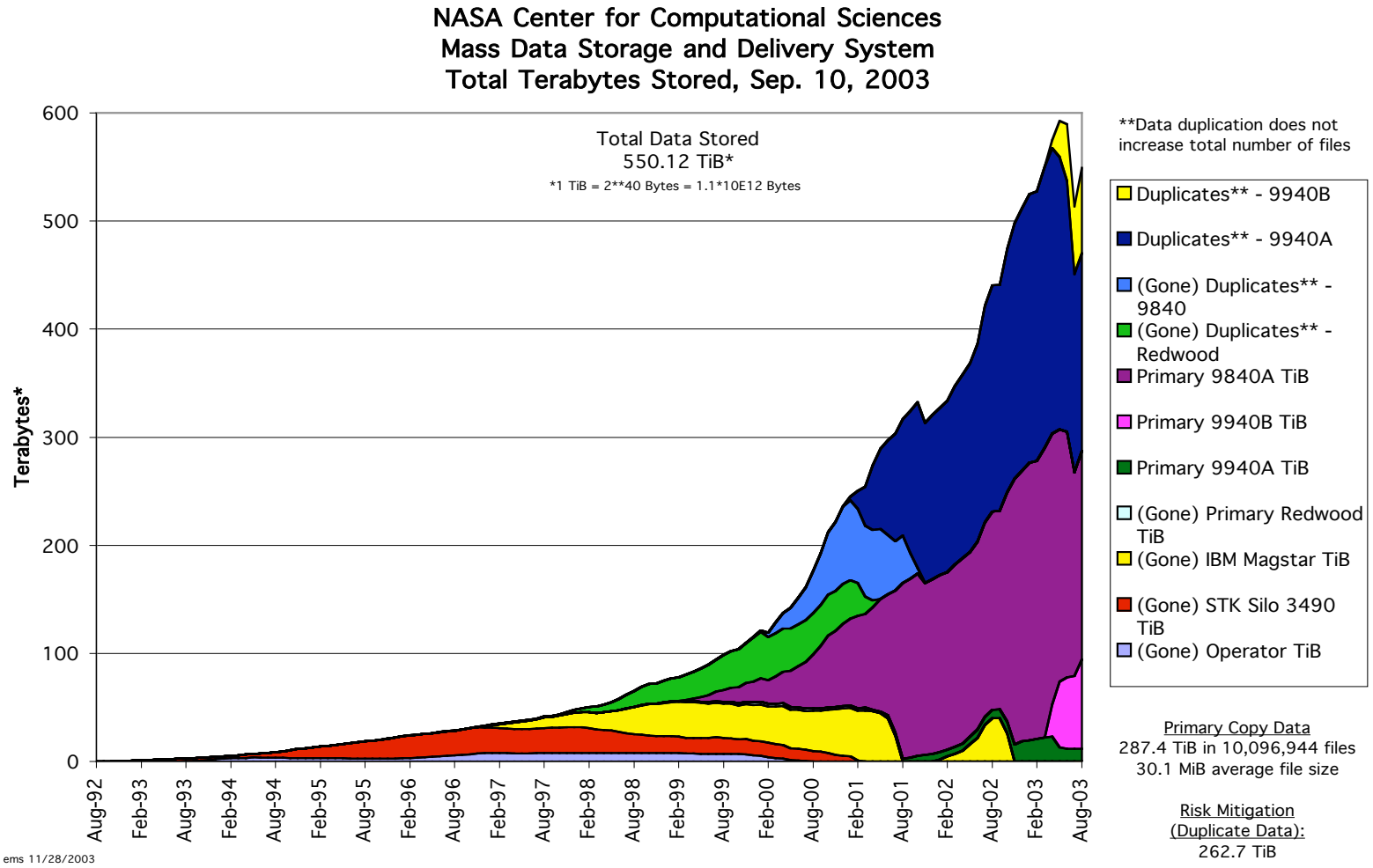
NCCS HEC, MDSDS, Network Evolution Snapshots

	HEC Engine(s)	HSM, Storage Software	Network "Media"	Network Software
1985	CDC Cyber 205	IBM's DFHSM	Block Mux	CDC's MFLINK
~1990	Cray YMP	Convex UniTree	UltraNet	ftp
~1996	Cray J90, Cray T3E	HP/Convex UniTree+	HiPPI	ftp
2004	HP ES 45 SC, SGI O3800, SGI Altix (soon)	Sun's SAM-QFS, SGI's DMF, DMS/SRB	Gigabit Ethernet	sftp,scp, SRB's Sput, Sget

(Dis)Continuum of Migrations

- Evolutionary migrations
  - Hardware: tape, network
  - Software: new releases
- Disruptive, “fork lift” migrations
  - Hardware: servers, disk arrays
  - Software: replacement

Evolutionary Migration: NCCS Mass Data Storage and Delivery System Tape Media



Alternatives for Large Scale Heterogeneous HSM Migrations (1)

- New software reads old system's tapes directly; all data migrated
  - Rarely done, risky
    - Can be intellectual property issues
    - Could be less expensive because can be done without retaining old system's hardware/software
      - But what recourse if something goes wrong?

Alternatives for Large Scale Heterogeneous HSM Migrations (2)

- Attempt at “data interchange definition”: MS66 standard
  - Vendors participated in large part because they wanted to be able to read /convert from other HSMs to their own (not surprising)
  - A few vendors used its principles to enable movement or pruning/grafting of directories from one site to another when both sites were running that vendor’s HSM product
  - Not clear that any vendors have used the standard otherwise
  - Not clear whether any user sites specified complying with this standard in any Requests for Procurements

Alternatives for Large Scale Heterogeneous HSM Migrations (3)

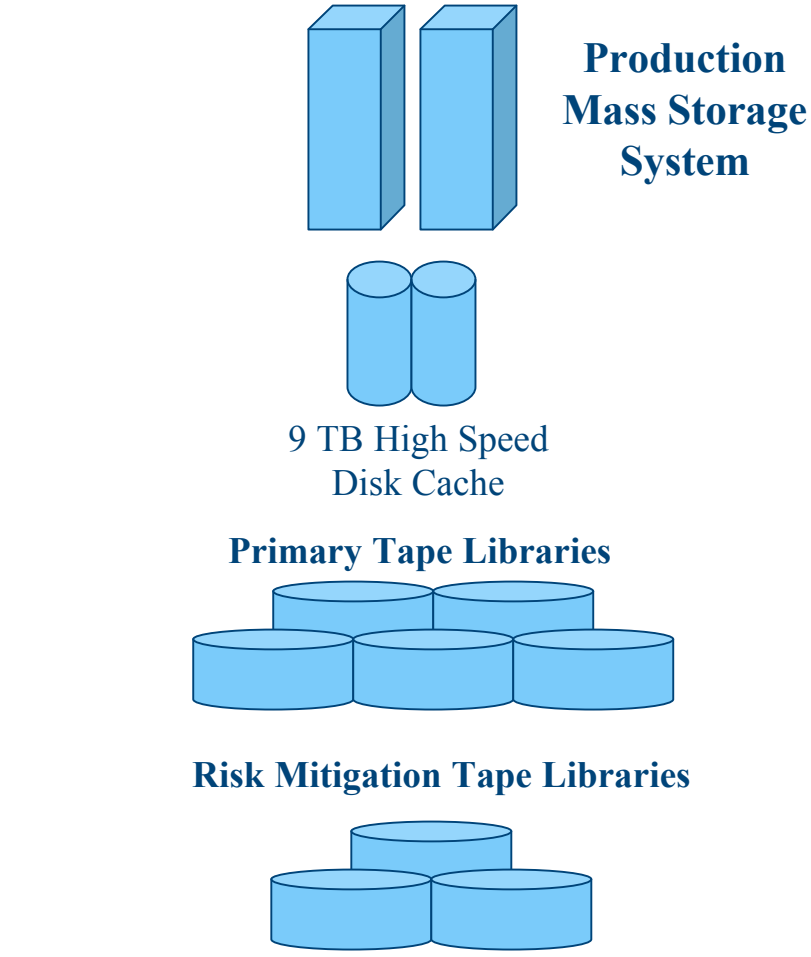
- Have users identify/move only data of value
  - Few tools to help identify valuable data
  - Difficult to optimize moves
- Transparent migration
  - Populate new system with old system's filename/directory structure
  - Automatic on-demand "hook" in new system to read old files(using old system)
  - Background migration from old to new

Large Scale Data Migrations at the NCCS

- 1992-1994: MVS to Open Systems:
  - IBM's DFHSM to Discos/Convex UniTree (2 TiB, 0.5M files), created the MDSDS
- 2003: Homogeneous, West Coast to East Coast:
  - "Grafting" DMF onto DMF (190 TiB, 10M files)
- 2003-2004: ~SPOF* to ~HA** for "MDSDS" (Completed August 30, 2004):
  - UniTree/DiskXtender to SAM-QFS (290 TiB, 11M files)
- Future?: Native Filesystem to Data Management System(DMS)
  - Combined DMF and SAM-QFS, October 2004: 835 TiB, 33M files

*** ~SPOF == Single Points of Failure, but failures rare**
**** ~HA == Quasi-High Availability**

NCCS SAM-QFS HSM Configuration



- Sun SAM-QFS
- Sun Fire 15K Server
  - Two production domains
  - Shared QFS filesystems
  - Veritas Cluster Server
- Test Cluster
  - Small domain on SF 15K
  - Sun V880

Strong Benefits in NCCS's Current SAM-QFS Configuration (1)

- 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
- Restoring files after accidental deletions much simpler/faster than previous solution
- A test cluster system has been invaluable
  - Changes our site considers making are not necessarily widely done in the SAM "mainstream"; test cluster allows thorough testing before we commit to them

Strong Benefits in NCCS's Current SAM-QFS Configuration (2)

- Using "HA flip-flop" for significant software upgrades has greatly reduced downtime for significant software upgrades
  - Requires "over engineered" high performance and capacity on each of two domains
  - Filp-flop:
    - Intentionally "fail over" the filesystems and functions of the first domain to be upgraded, with remaining domain handling all user traffic.
    - Apply software upgrades to the "failed" domain
    - Intentionally "fail over" un-upgraded domain to newly upgraded one, so that newly upgraded one handles all user traffic
    - Apply software upgrades to the remaining "failed" domain
    - "Fail back" the respective filesystems and functions to the most recently upgraded domain.

NCCS's UniTree to SAM/QFS HSM Migration Lessons Learned (1)

- Automating high-availability software is challenging for clustered HSM systems
  - ***Tape drive sharing*** between SAM-FS cluster members
    - Heavy use exposed problems; NCCS disabled sharing
    - Result: increased costs because more tape drives needed than if the drives could be shared successfully
  - ***Veritas Cluster Server*** (VCS)
    - Configured to monitor and fail SAM-QFS cluster services by Sun staff, before local staff was familiar with SAM-QFS and with VCS.
    - Heavy activity on our system resulted in problems that caused us to disable VCS (and instead fail over cluster members by hand, when needed) until such time as we can more fully understand and appropriately implement VCS.

NCCS's UniTree to SAM/QFS HSM Migration Lessons Learned (2)

- 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)

Near Term Explorations, Longer Term "Twinkles in Our Eyes"

- 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 problematic
    - Strong pattern of file retrieval within first month of creation
    - Also strong pattern of retrieval for older files (months to years), but often with little "locality of reference"
  - 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
  - Exploring long-range SANs,data grids in support of geographically distant teams, etc.

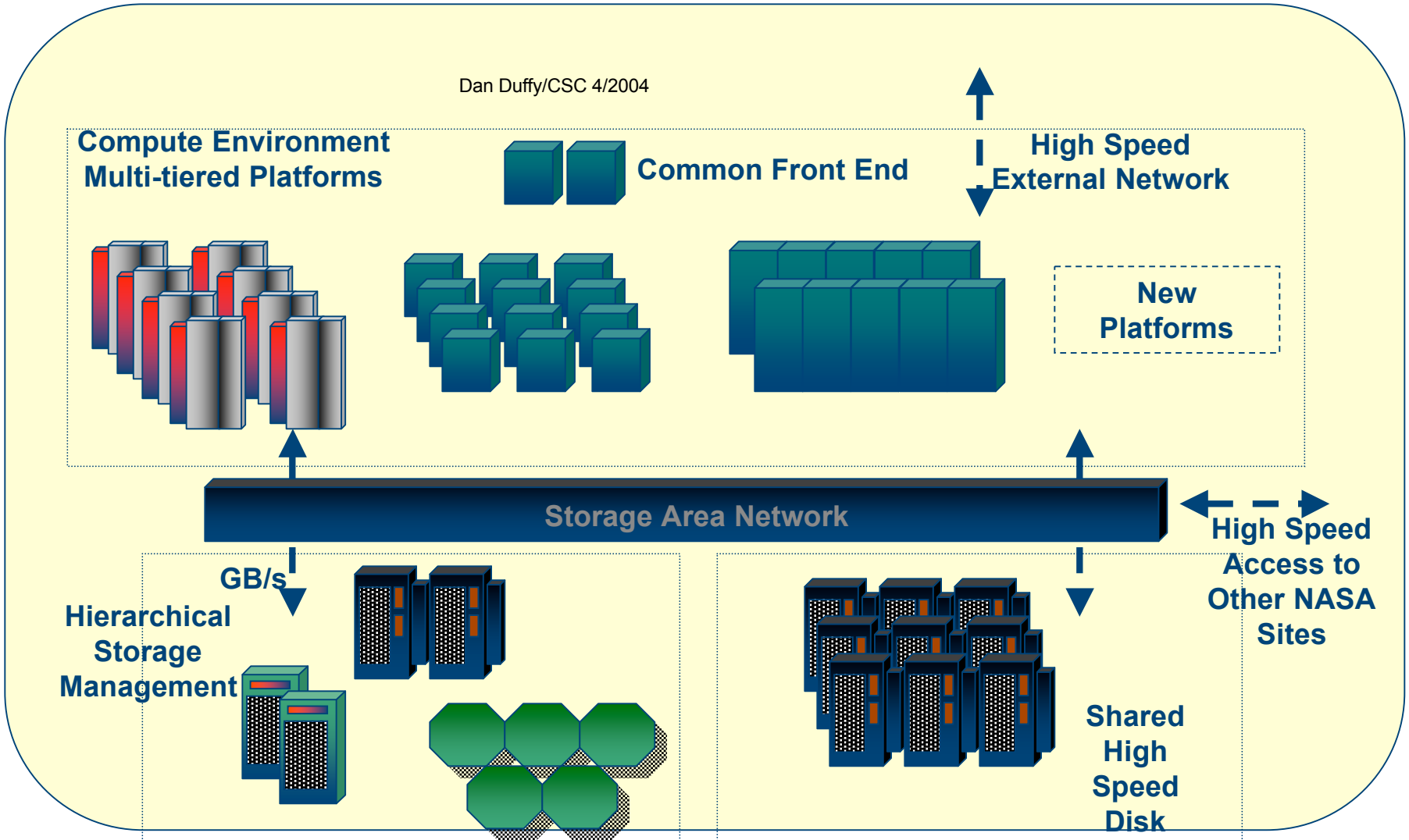
Continuing in-the Trenches Challenges

- High-performance sites: it's not just big data, it's also lots of files
  - Difficulty migrating from UniTree a user directory with 92K+ small files
  - Tape and disk intentionally optimized for larger files and sequential I/O; addition of millions of tiny files adds challenges
- Requests to move tens of thousands of files (already written on tape) to new filesystem:
  - Currently requires copying from tape-(to-disk-)to-tape
  - By contrast, "virtualization" of the file location could allow a simple rename instead

NCCS's Incipient Data Management System (DMS)

- Requested by largest customer's management to help them 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

NCCS Evolving Architecture: Data Centric, Multi-Tiered (1)



NCCS Evolving Architecture: Data Centric, Multi-Tiered (2)

Pros

- Ease of use leads to higher user productivity and better support
- Common front ends allow for greater utilization of resources
- Storage area network provides large, fast storage from all compute platforms
- Multi-tiered HSM environment integrated into the SAN
- Multi-tiered computational engines provides the appropriate platform for the application, rather than the other way around
- Extensible architecture makes it more adaptable to new architectures and changing requirements

Cons

- Local attached storage may still be needed for certain applications
- Vendor specific SAN software could limit future integration efforts
- HSM is typically tightly coupled to the SAN software

Dan Duffy/CSC 4/2004

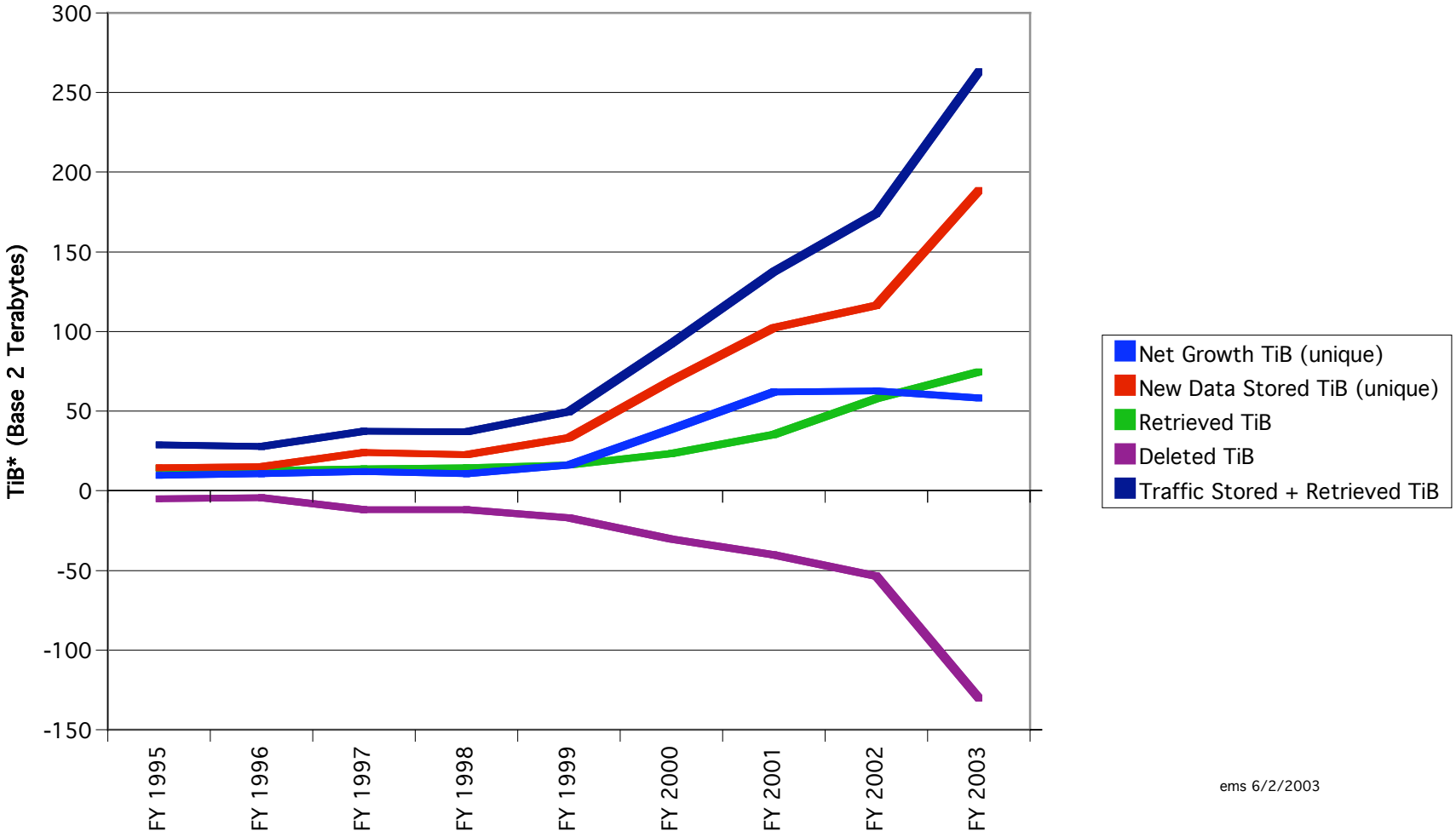
References

- References:
  - SRB URL: http://www.npaci.edu/DICE/SRB/
  - Goddard IEEE Conference on Mass Storage Systems and Technologies http://storageconference.org
- Acknowledgements:
  - CSC: Dan Duffy, Sanjay Patel, Marty Saletta, Lisa Burns, Ed Vanderlan
  - NCCS: Nancy Palm, Adina Tarshish, Tom Schardt
  - Sun: Mike Rouch (SANZ),Bob Caine, Randy Golay, Linda Radford
  - Instrumental, Inc.: Jeff Paffel, Nathan Schumann
  - Halcyon Systems, Inc.: Jose Zero, David McNab, Ignazio Capano

Backup

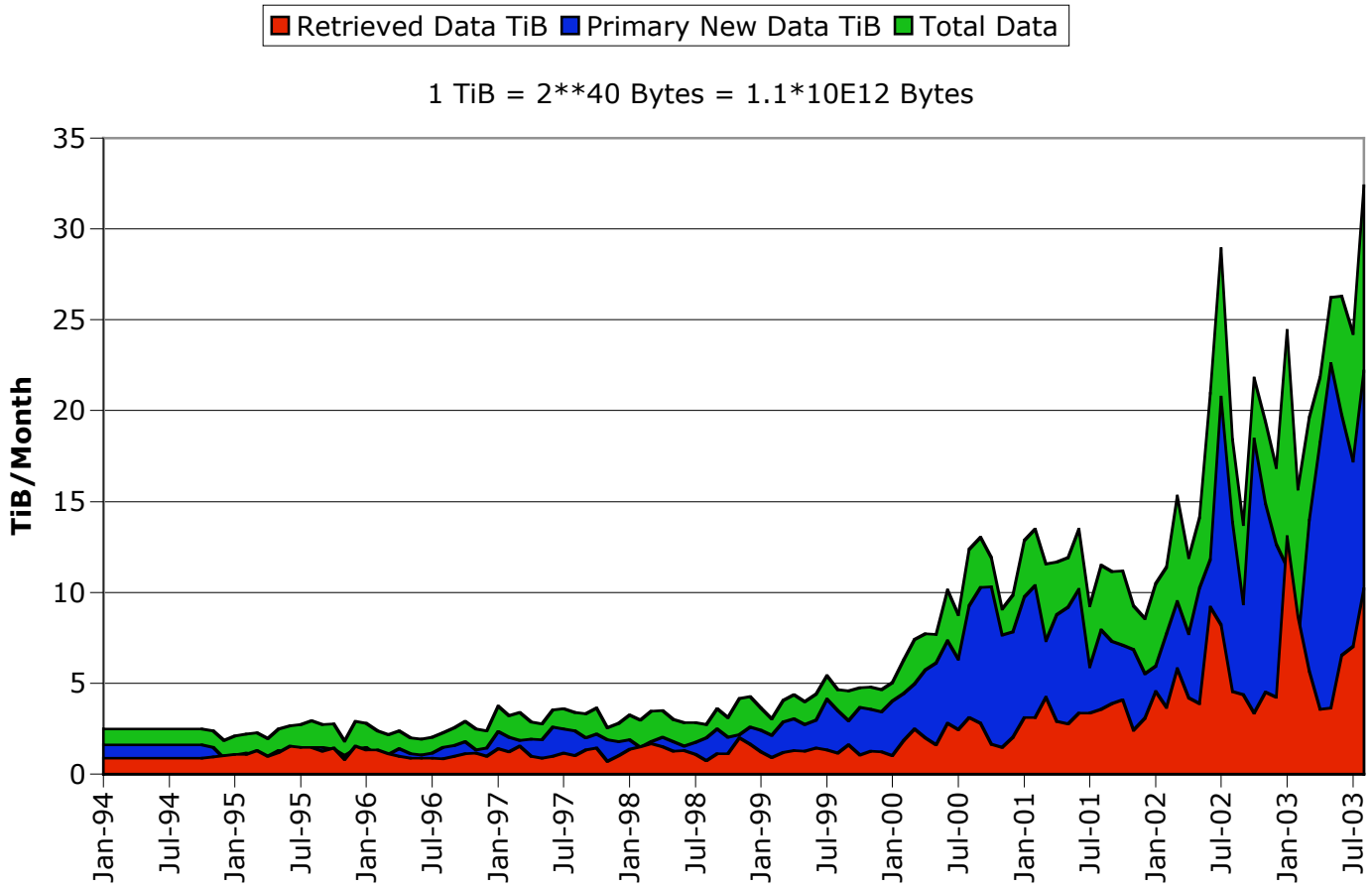
NASA Center for Computational Sciences
Mass Data Storage and Delivery System
Data Stored, Retrieved, and Deleted FY1995 - FY2003*

1 TiB = 2**40 Bytes = 1.1*10E12 Bytes

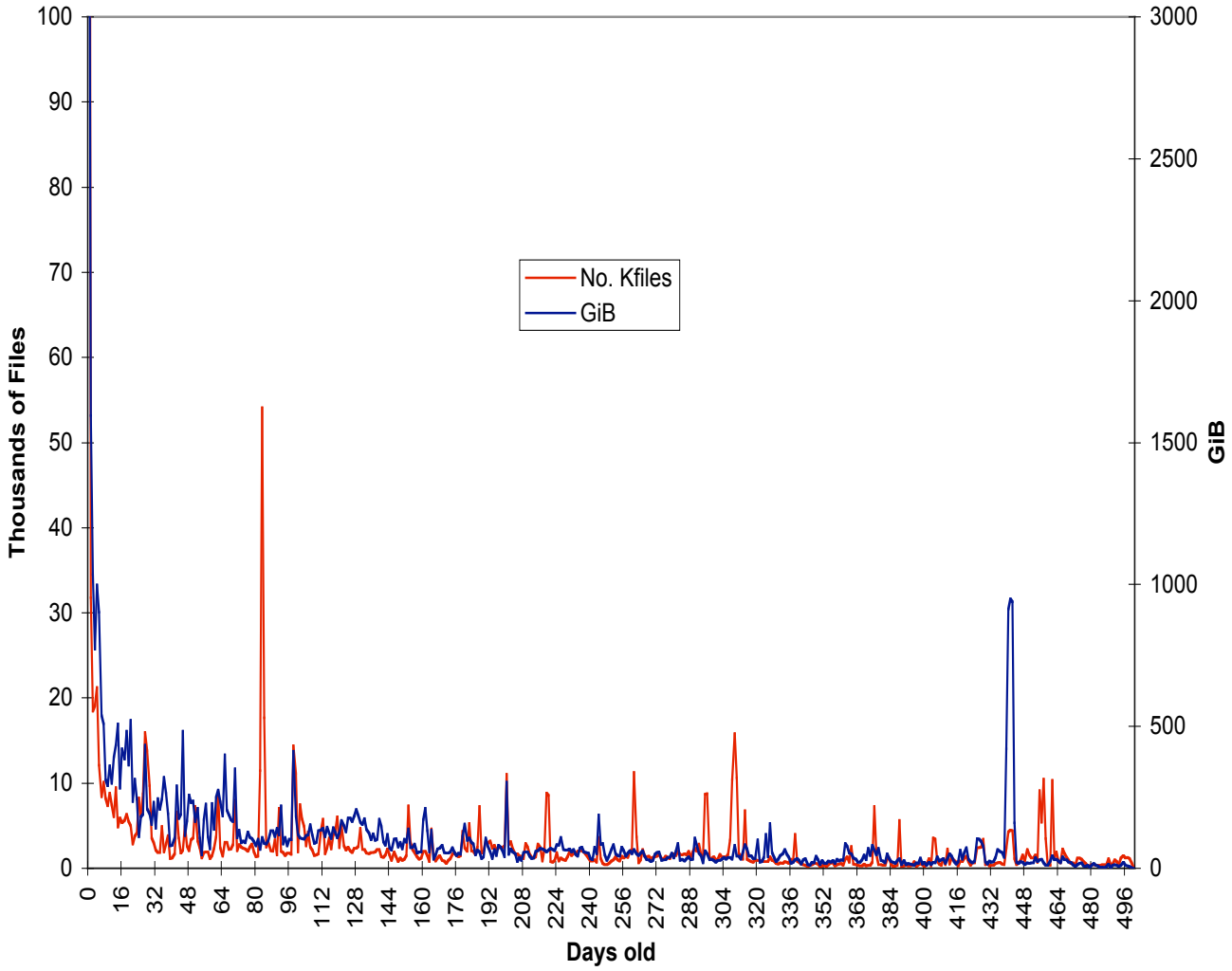


ems 6/2/2003

NASA Center for Computational Sciences
Mass Data Storage and Delivery System
Monthly Store and Retrieve Traffic



ems 11/28/2003


MDSDS Age of Files Retrieved Jan. 1, 2003 - Feb. 10, 2004
for files 500 days old or younger
Thousands of Files
GiB
Days old
No. Kfiles
GiB
0
10
20
30
40
50
60
70
80
90
100
0
500
1000
1500
2000
2500
3000
0
16
32
48
64
80
96
112
128
144
160
176
192
208
224
240
256
272
288
304
320
336
352
368
384
400
416
432
448
464
480
496