ARSC Storage Solution

Gene McGill

Storage Specialist

mcgill@arsc.edu

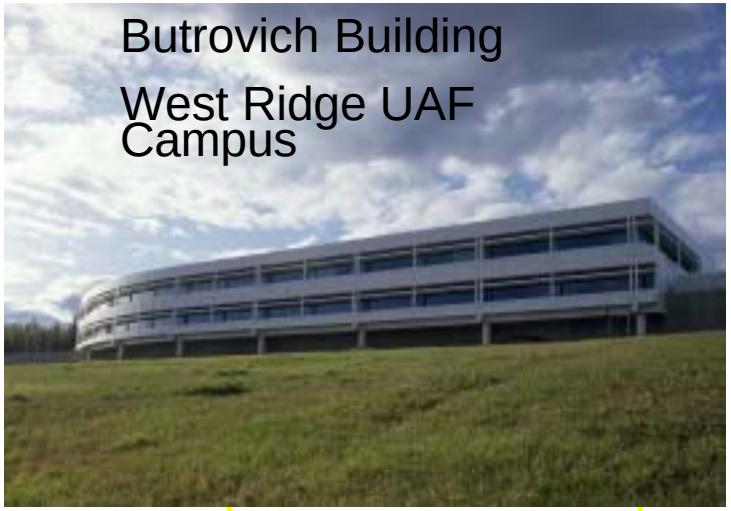
www.arsc.edu

Overview

- About ARSC
- Where we've been in storage
- Why change?
- Goals for the new storage systems
- Migration challenges
- Where are we now?
- Some performance data
- Questions

Who we are

- High performance computing center
- University owned and operated
- DoD funded through HPCMP



What we do

- Support computational research in science and engineering with emphasis on high latitudes and the Arctic
- Provide HPC resources and support
  - Computing, Data Storage, Visualization, Networking
- Conduct research locally and through collaborations

Hardware

IBM P690 Regatta
Symetric Multiprocessing (SMP)
- 32 processors
- 64 gigabytes RAM
- 166.4 gigaflops
- 1.4 terabytes disk

Cray SX-6
Parallel Vector
Computer
- 8 processors
- 64 gigabytes RAM
- 64 gigaflops
- 1 terabyte disk

Cray SV1 EX
Parallel Vector
Processor
- 32 processors
- 32 gigabytes RAM
- 64 gigaflops
- 2 terabytes disk

IBM SP
Massively Parallel
Processor
- 200 processors
- 100 gigabytes RAM
- 276 gigaflops
- 1.2 terabytes disk

Cray T3E 900
Massively Parallel
Processor
- 272 processors
- 26 gigabytes RAM
- 230 gigaflops
- 522 gigabytes disk

ARCTIC REGION SUPERCOMPUTING CENTER

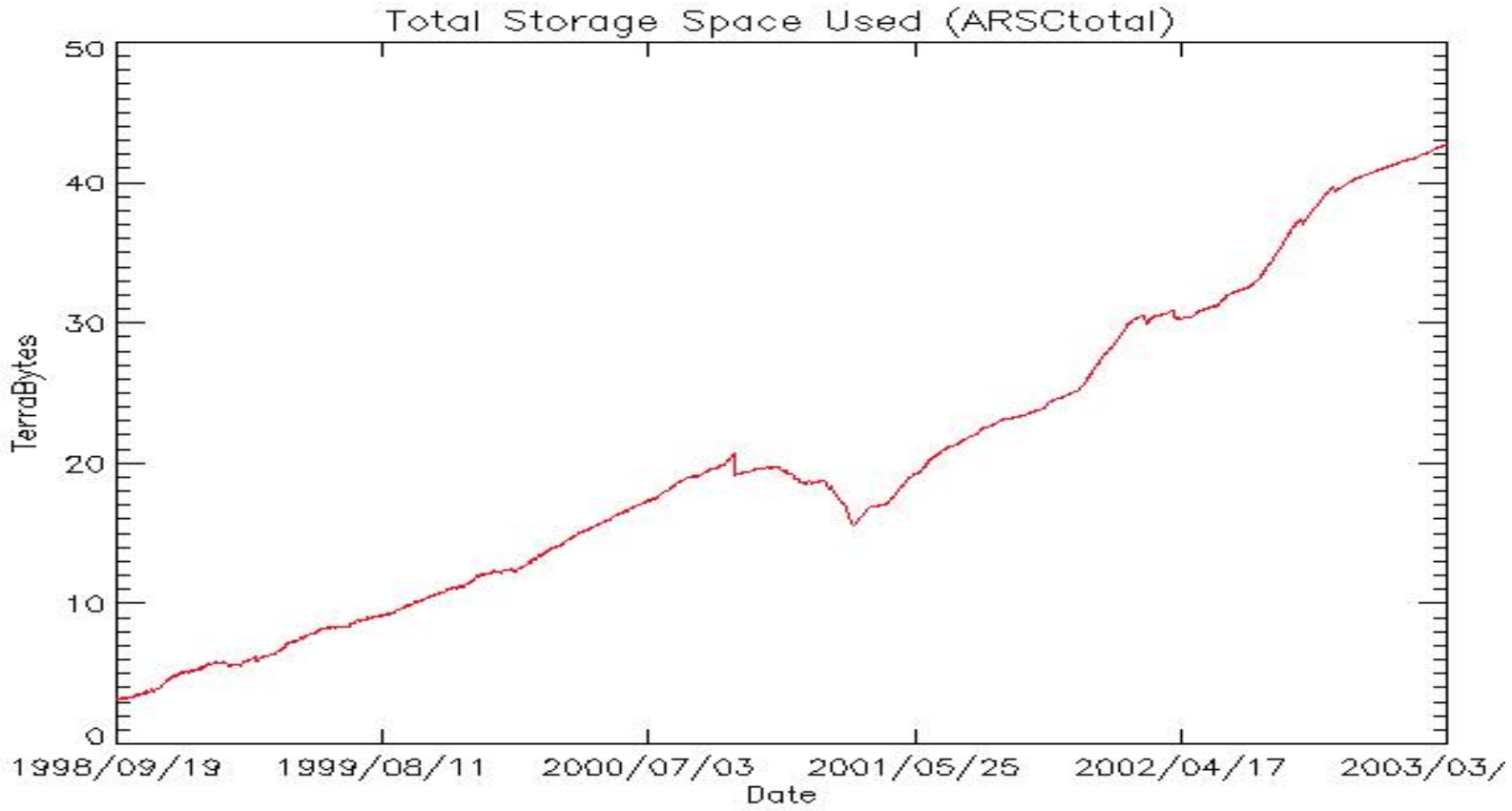
Center Evolution

Upgrades in power, memory and storage over the years

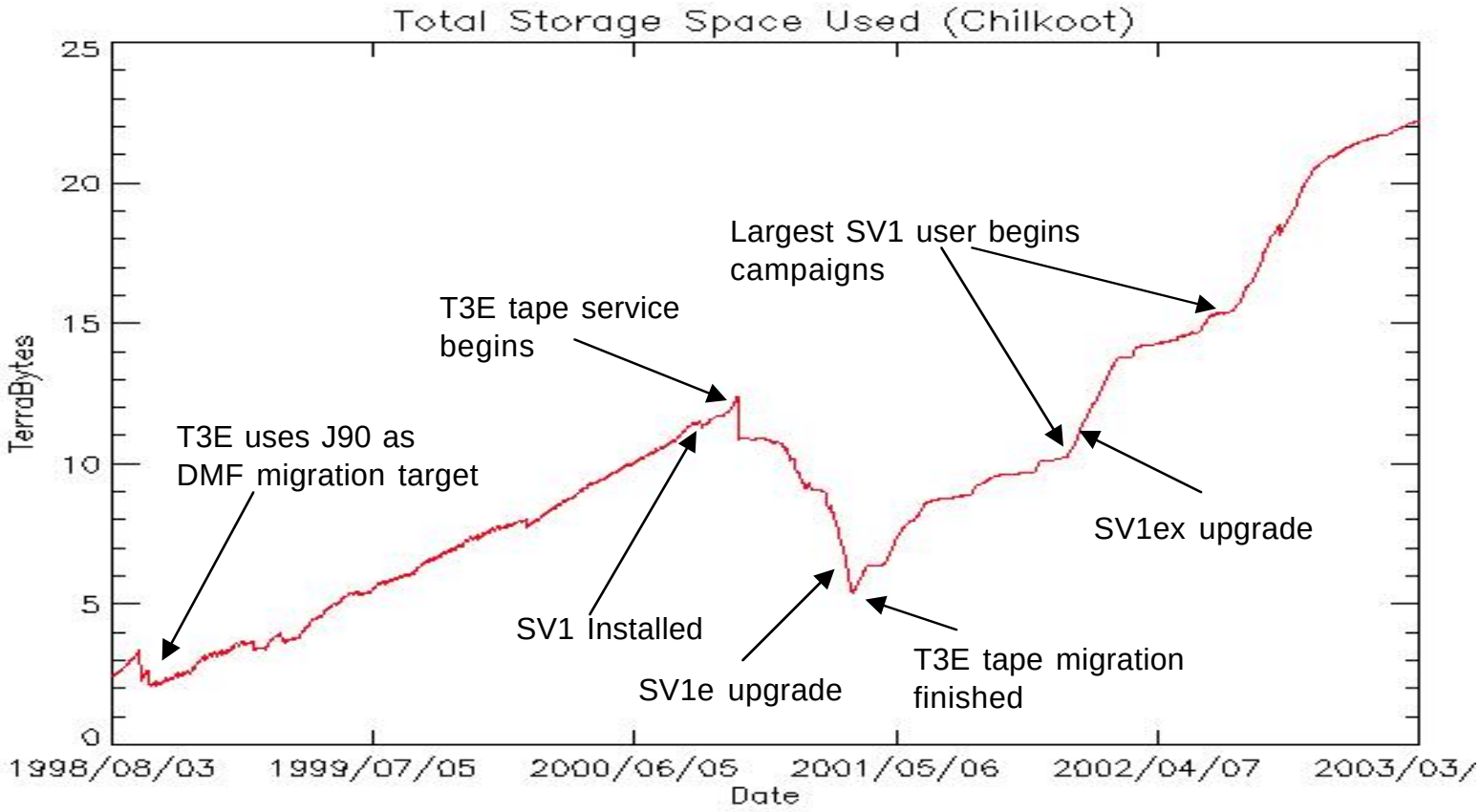
Where we've been in storage

- ARSC storage has always been Cray/DMF
  - DMF hosted on each Cray
- Vector Cray NFS served filesystems to other machine

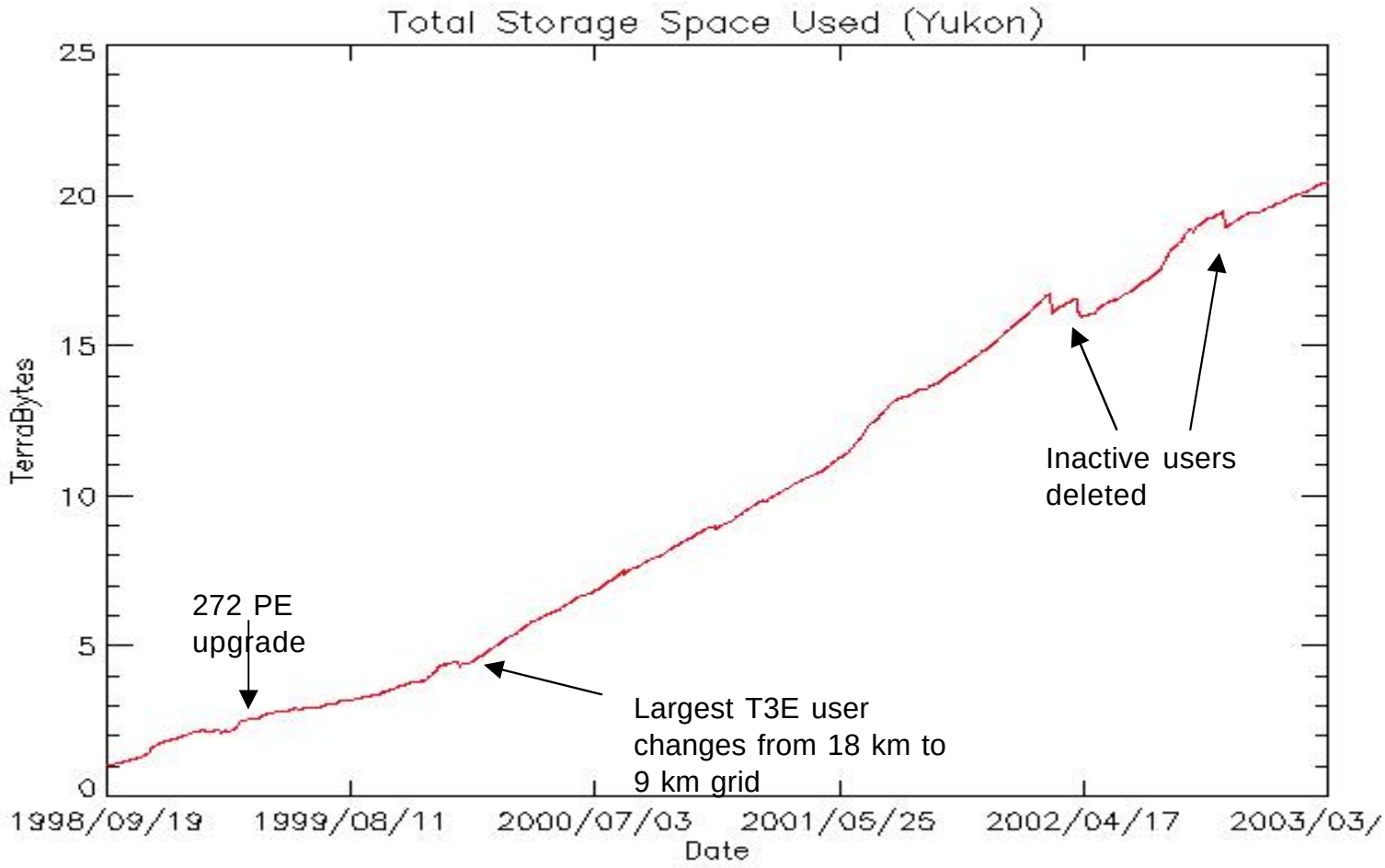
Where we've been (cont.)



Where we've been (cont.)



Where we've been (cont.)



Why change?

- Cray limitations
- Challenges of maintaining storage in multi-system, multi-vendor environment
- SAN technologies maturing
- DoD-mandated separation

Goals for new storage systems

- Reliable, maintainable systems
- High performance but not bleeding edge
- Systems must meet DoD needs
- Scalable in capacity, bandwidth
- Productive for users

Goals: Reliable, maintainable systems

- 2 Sun Fire 6800s each with:
  - Resilient configuration (though not true HA)
  - 8 900MHz processors, 16 GB memory on 2 system boards
  - 10.5 TB raw, dual-pathed, fibre channel T3+ disk
  - 6 STK T9840B, 4 T9940B tape drives
  - Sun-branded Qlogic switches, HBAs
  - Trunked & multi-pathed networks
  - SAM-QFS HSM software
- 1 Sun Fire 4800
  - Subset of above
  - For testing only

Goals: High performance but not bleeding edge

- Aggregate BW to disk = 800MB/s in redundant configuration
- Aggregate native BW to tape = 234MB/s
- Separate NFS, login networks
- Already in use at MSRCs, elsewhere

Goals: Meeting DoD needs

- Configuration is compatible with MSRCs
- Two systems enables separation of DoD-sensitive data from systems serving UA needs

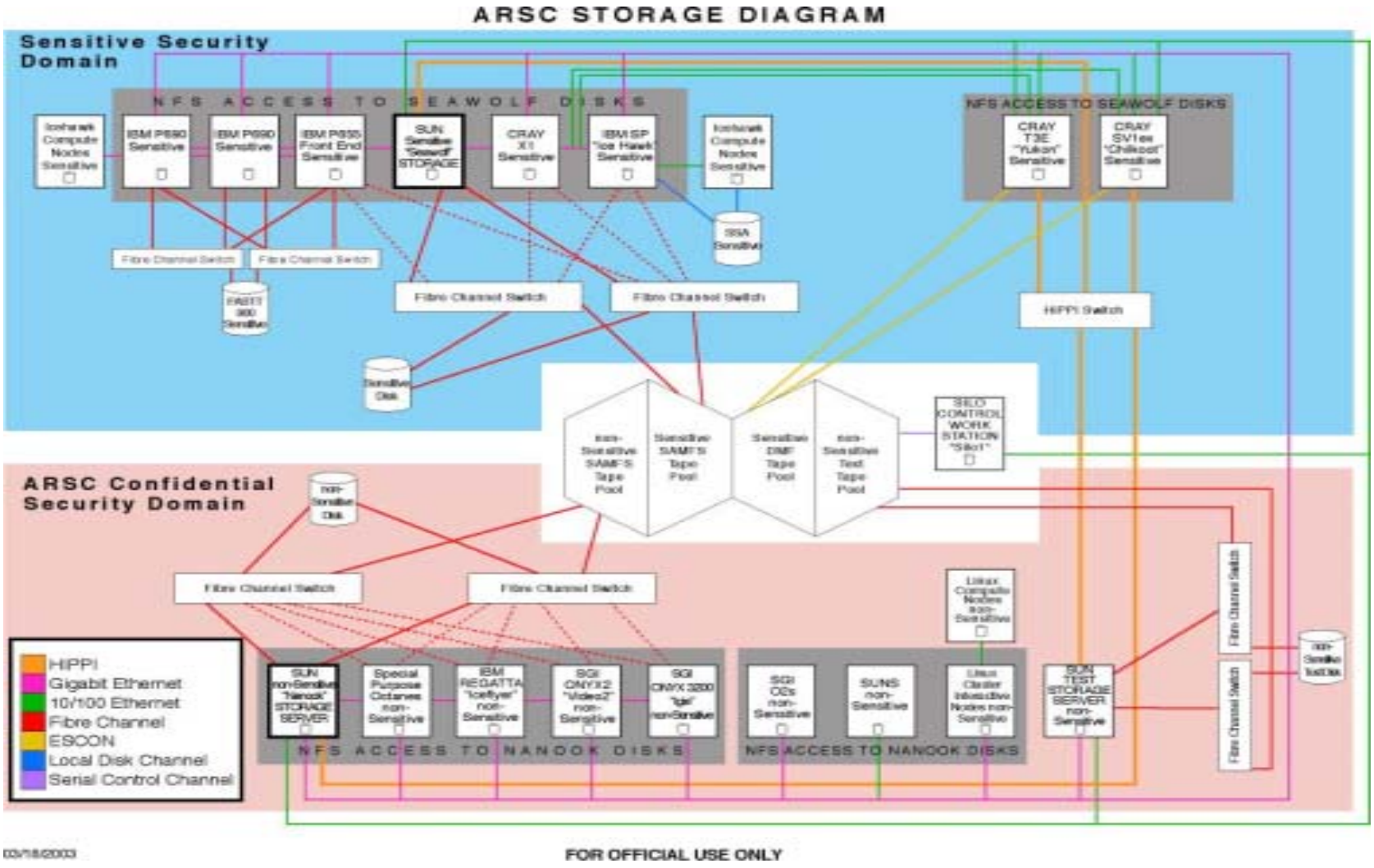
Goals: Scalable in capacity, bandwidth

- Initial configuration targeted for Cray X1, IBM P690+/P655+ demands
- Can add CPUs, memory
- Some I/O expansion room
- Future SAN software may help I/O bandwidth

Goals: Productive for users

- Standardized environment variables on all ARSC machines
- Small, quota-controlled $HOME that is platform-specific (local to all but SGIs)
- Large $WRKDIR (purged, temp space) on compute platforms (local on all systems)
- Large $ARCHIVE for storing bulk of data
  - Larger disk quotas
  - NFS exported from SF 6800s within security domain
- Other special-purpose filesystems where beneficial
- New features not available in old configuration

Goals: The big picture



Migration challenges

- DMF -> SAM-QFS
  - Limited functionality vs. more flexibility (and complexity)
  - DMF is mature, SAM-QFS is less mature (as is the support organization)
- Fibre Channel SAN
  - New skill set to learn
  - Multi-vendor SAN makes support a challenge

Migration challenges (cont.)

- Combined storage, compute -> separate storage, compute
  - More complicated for users and support staff
- Data on multiple hosts going to other multiple hosts
  - Most Cray-based data going to sensitive 6800 via SAM migration toolkit
  - Remaining data moving to the other 6800 in a more manual manner
  - Migration estimated to take 6-12 months

Where are we now?

- **Sun systems**
  - **Solaris mostly hardened**
    - Most ARSC packages installed, working
  - **Networks**
    - Copper GigE trunking not quite here
    - Single GigE performance is good
    - HIPPI support best under Solaris 8
  - **SAM-QFS is still in test**
    - Initial tests had unacceptable error rates
    - Discovered obscure error with st, SAM,we're testing a workaround
    - Migration toolkit ready for test
- **STK tape drives working as advertised**

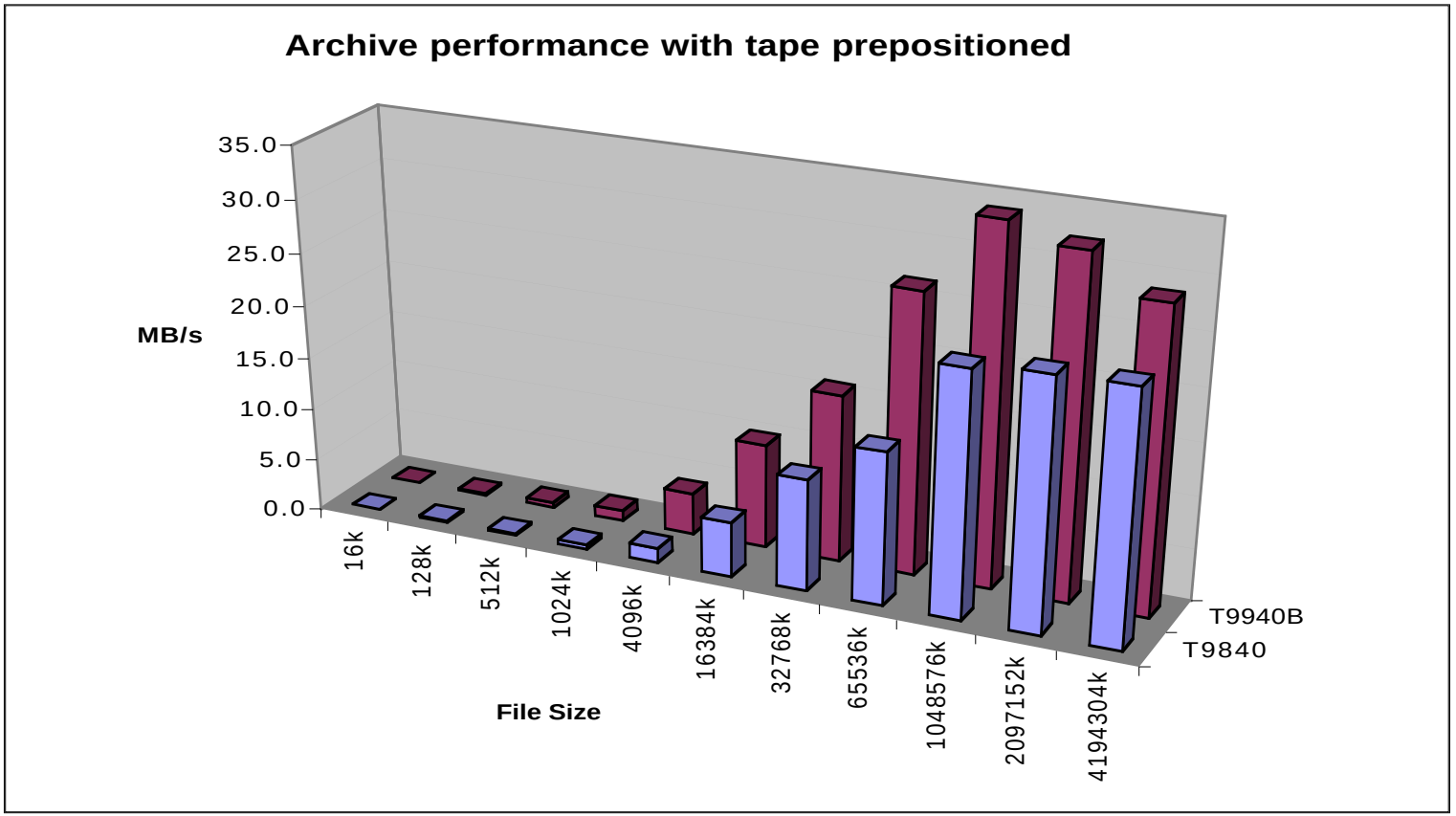
Some performance data

- Tape
- Disk
- Network
- SAM vs. DMF

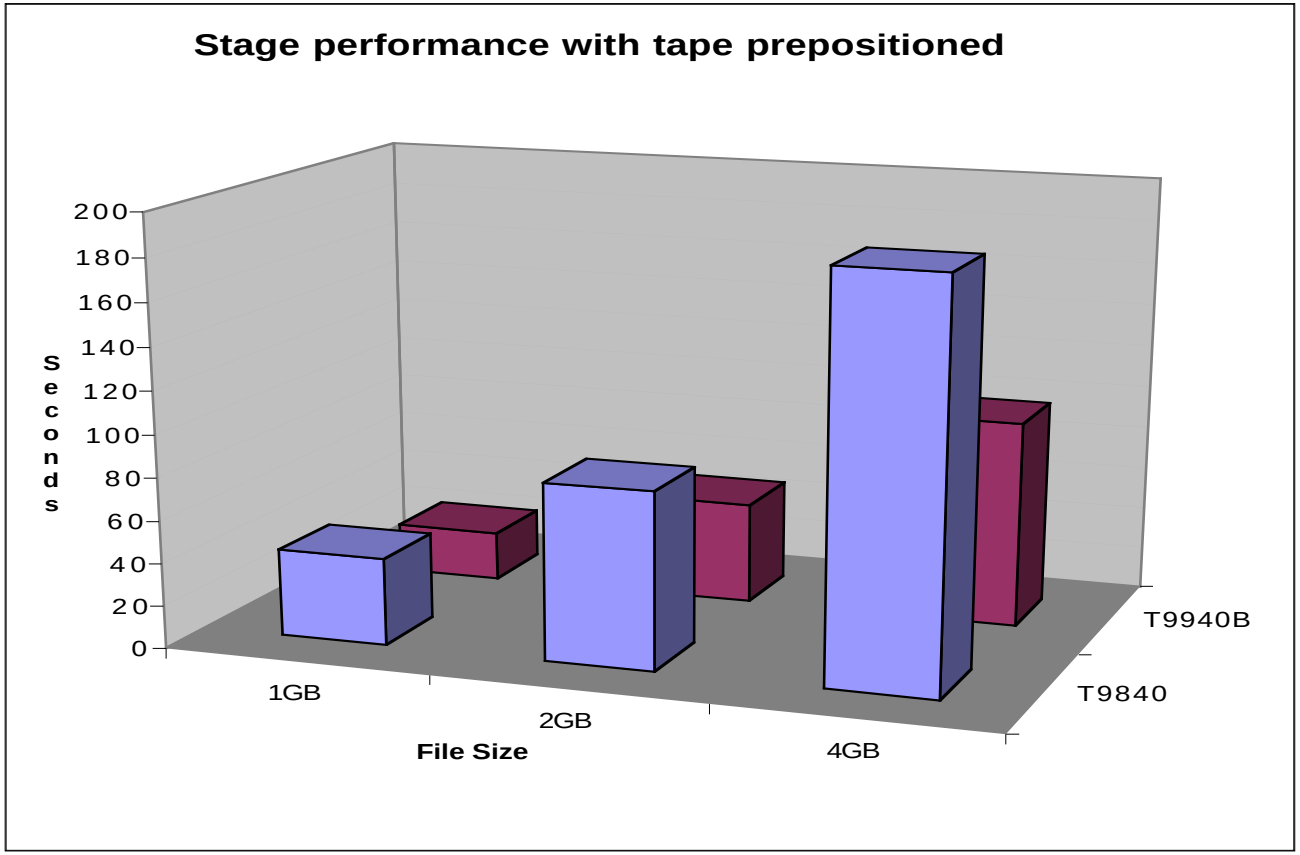
Performance: Tape

- **Capacity of tapes:**
  - **128MB files**
  - **Random data**
  - **Compression on**
  - **256K blocks**
- **T9840: 27.0GB**
- **T9940B: 280.5GB**

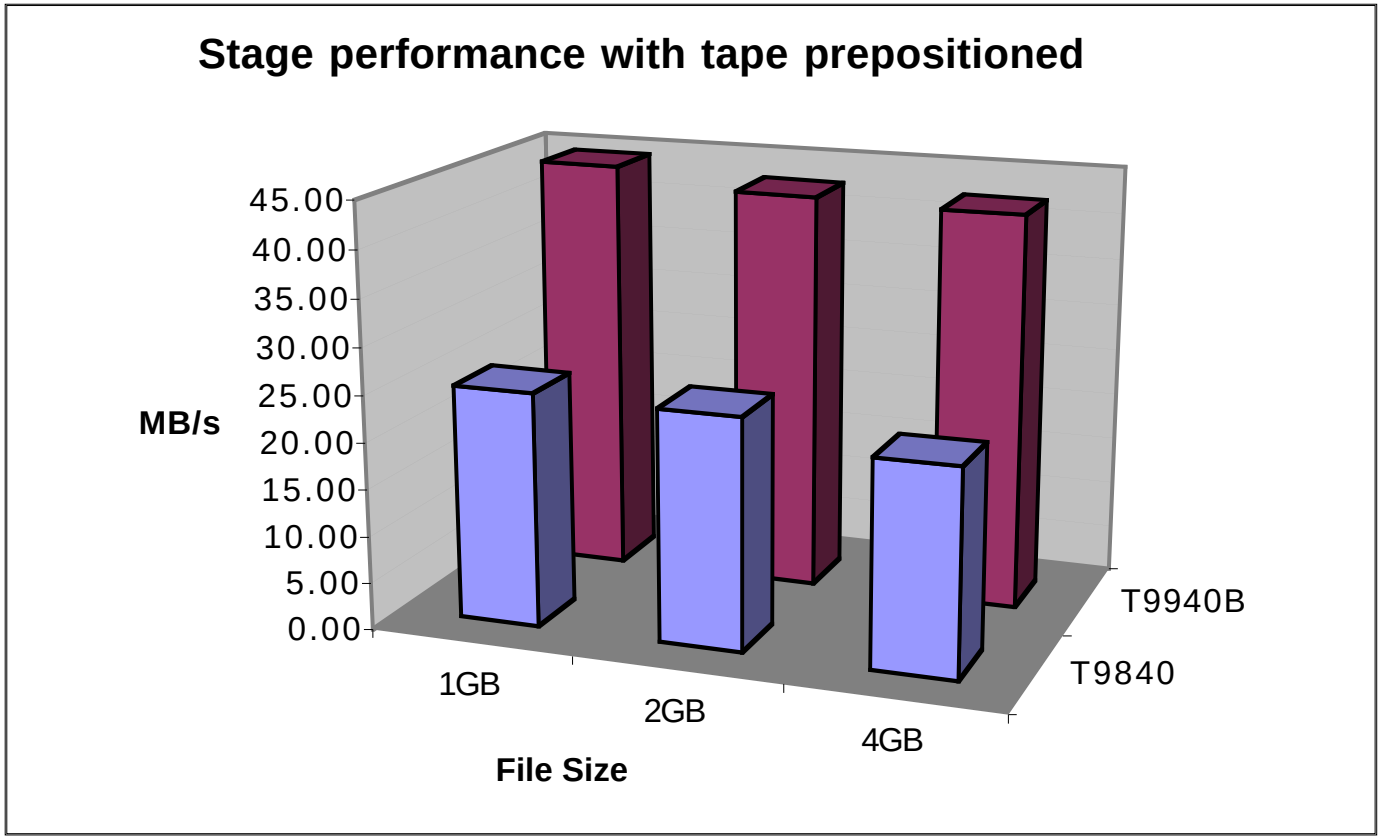
Performance: Tape



Performance: Tape

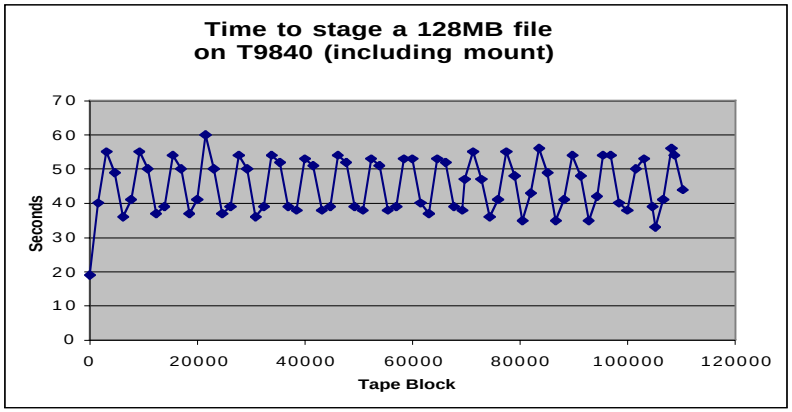


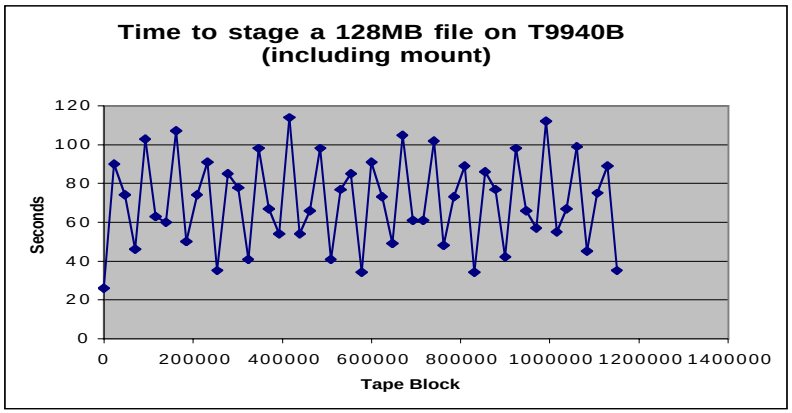
Performance: Tape



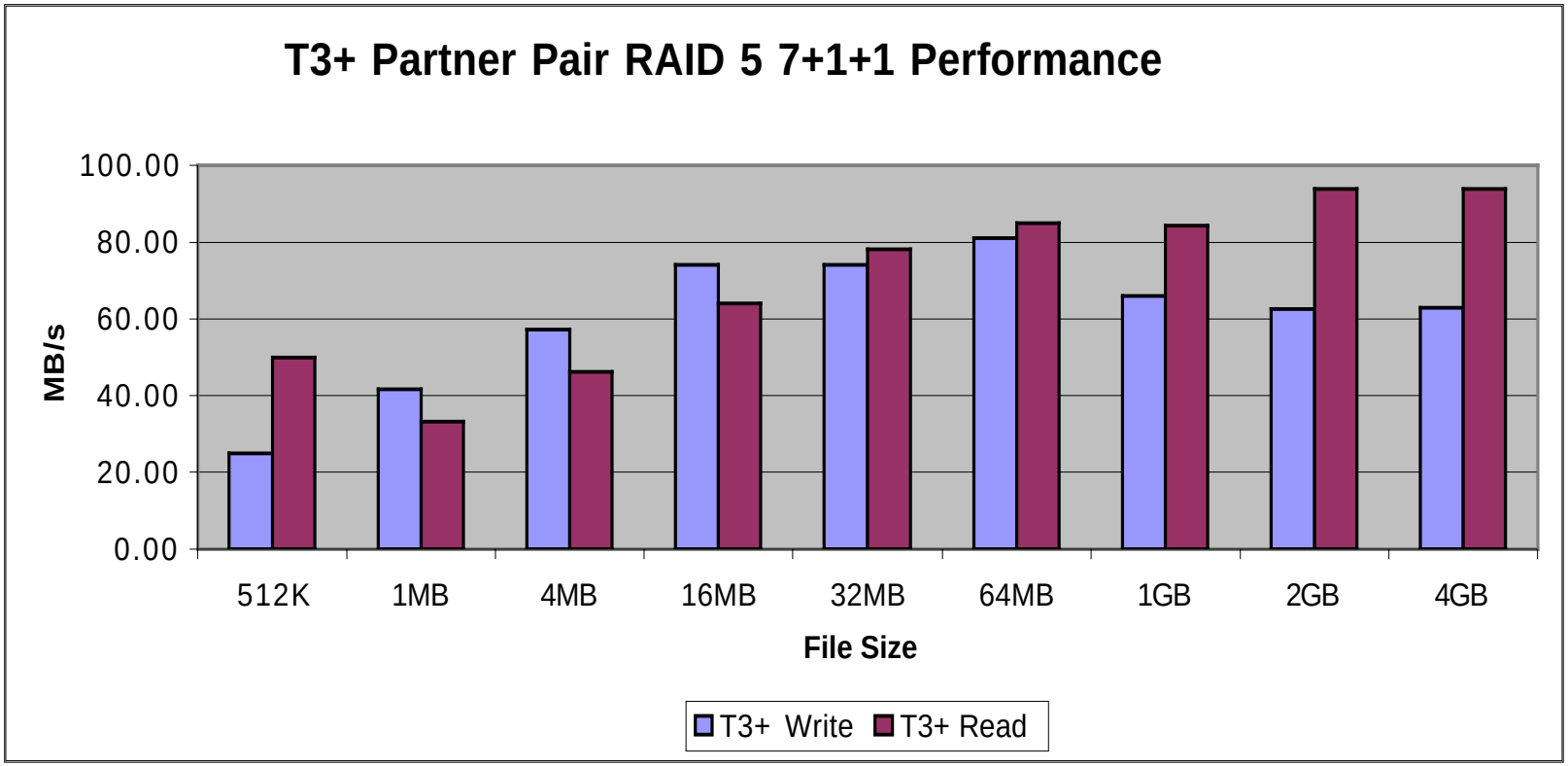
Performance: Tape

- **Positioning**
  - **T9840 takes between 20 & 60 seconds**
  - **T9940B takes between 25 & 115 seconds**





Performance: Disk



Performance: Net

- **Gigabit Ethernet**
  - **12 input streams from 3 SGIs = 940Mbit/s**
  - **1 input stream from Onyx 3200 = 579Mbit/s**
  - **18 input streams from Onyx 3200 = 706Mbit/s**
- **HIPPI**
  - **SV1 to Sun Fire**
    - 1 input stream = 268Mbit/s
    - 2 input streams = 521Mbit/s
    - 4 input streams = 575Mbit/s
    - 6 input streams = 576.3Mbit/s

Performance: SAM vs. DMF

- **SAM**
  - **Access to data during stage**
  - **Fast filesystem restores**
  - **Easy to map files on tapes**
  - **Tar-like tape format**
- **DMF**
  - **Must recall entire file before 1st data access**
  - **Slow filesystem restores**
  - **Difficult to map files on tapes**
  - **Proprietary tape format**

Performance: SAM vs. DMF

- **SAM**
  - **Error reporting not well-developed (SEF shouldn't be used yet!)**
  - **Copies > 1 and disk quotas don't play well yet**
  - **More flexible, more complex to set up**
  - **By default, SAM won't archive a file larger than the media!**
- **DMF**
  - **Error reporting more straightforward**
  - **Works better with disk quotas and copies > 1**
  - **Simple configuration (but less flexibility)**
  - **DMF copies all files by default.**

Questions