Storage Issues at NCSA: How to get file systems going wide and fast with in and out of large scale Linux cluster systems

Michelle L. Butler

Technical Program Manager

Storage Enabling Technologies Group

NCSA™

University of Illinois at Urbana-Champaign

Overview

- Mass storage
  - Convex to Convex Convex = HP to SGI
  - Growth
  - Lessons learned?
  - Where do we go from here?
- Supercomputer Storage
  - Different file systems used
  - File system “niches”
- TeraGrid and what that means to storage and mass storage.

History of Mass Storage System at NCSA

1985-1990

System: Amdahl dual processor 15 mips (if that ☺)

Software: CFS

Specifics: 36GB disk, 15MB memory, a single 1.5Mb hyperchannel connection, 10 3480 tape drives

Upgraded for new tape technologies and very limited by network performance

1991-1993

System:Convex C220I

Software: UniTree (From Convex)

Specifics: 100GB disk, 500MB memory, single Ethernet connection, 10 3480, and added 8 metrum drives

Upgraded for better network performance

Needed Hippi

1994-1997

System:Convex C3880 8 processors

Software: UniTree (From Convex)

Specifics: 200GB disk, 2GB memory, dual Ethernet connection, 1 Hippi, 8 metrum drives

Upgraded for new tape technologies, metrum being phased out, newer Hippi needed

1997-1998

System:Convex Exemplar 8 processors

Software: UniTree (HP… HP bought out Convex)

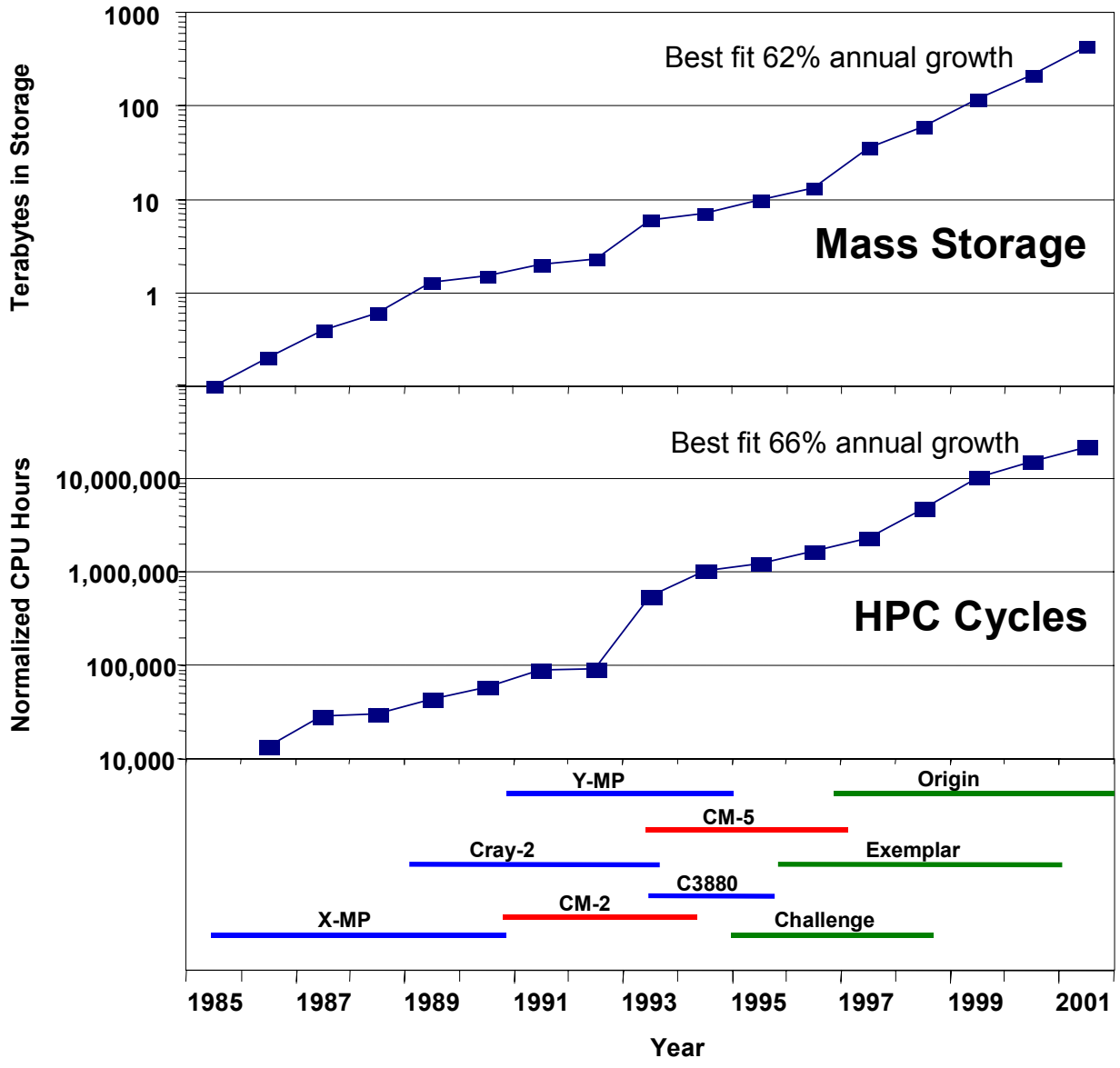
Specifics: 500GB disk, 4GB memory, three Ethernet interface, 2 Hippi, 8 metrum drives, and added 6 3590 drives with first robot

Upgraded more stable environment

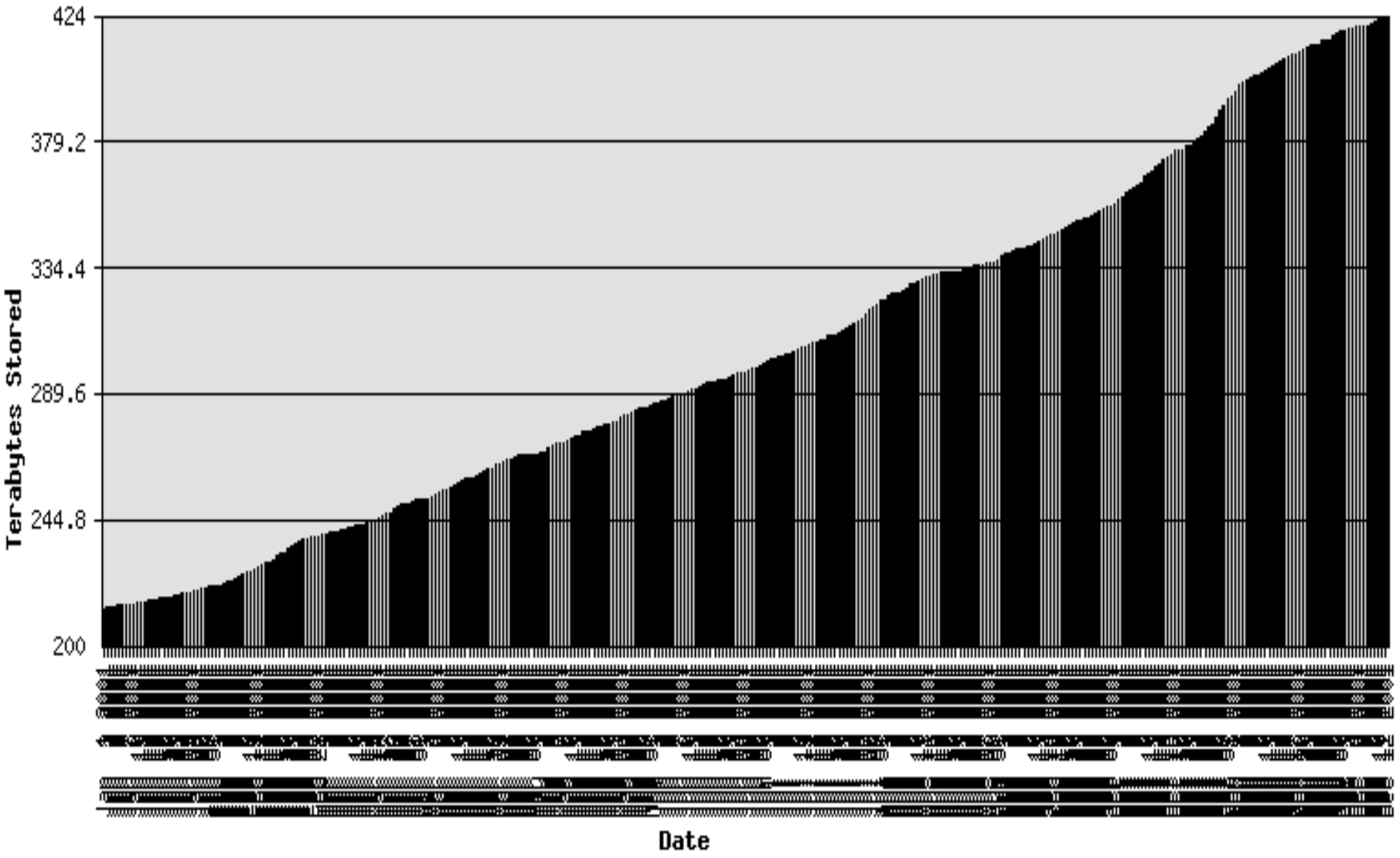
UniTree Inc and SGI

- 1999 - current
  - System:SGI
  - Software: UniTree from UniTree Inc
  - Specifics: 2TB disk, 4GB memory, dual 100baseT interfaces, 3 Hippi, 6 3590 drives with first robot
- **Never needed to move to new machine or software base, but did need to "add" capabilities**
- Current Machine:
  - 16 processors (295 MHz)
  - 12GB memory
  - 3 Hippi interfaces and 10 GigE interfaces
  - 12TB disks,10 3590 tape drives, 8 9840 tape drives with 1 Powderhorn library, 14 LTO tape drives with ADIC library of 720TB media capacity housing both LTO and 3590 drives

Growth over the years



Growth for 2001 alone



Lessons learned

- Separated enterprise wide backup system out from mass storage
  - Too many competing strategies
    - Small files, slow transfer, 8*5, different tape needs.
- Growth doubles every year;
  - We are now preparing for 440*2 for this year.
- Migrating to new tape technologies takes a LOT of time and processor resources. Not fun!
  - Found that each year if we are not upgrading technologies, we are changing tape technologies for faster performance and better foot print.

Lessons learned continued

- Dual writes to tapes has been a good but expensive insurance policy
- Users at NCSA have changed usage patterns from write only to 50% retrieval for first few months of file's life

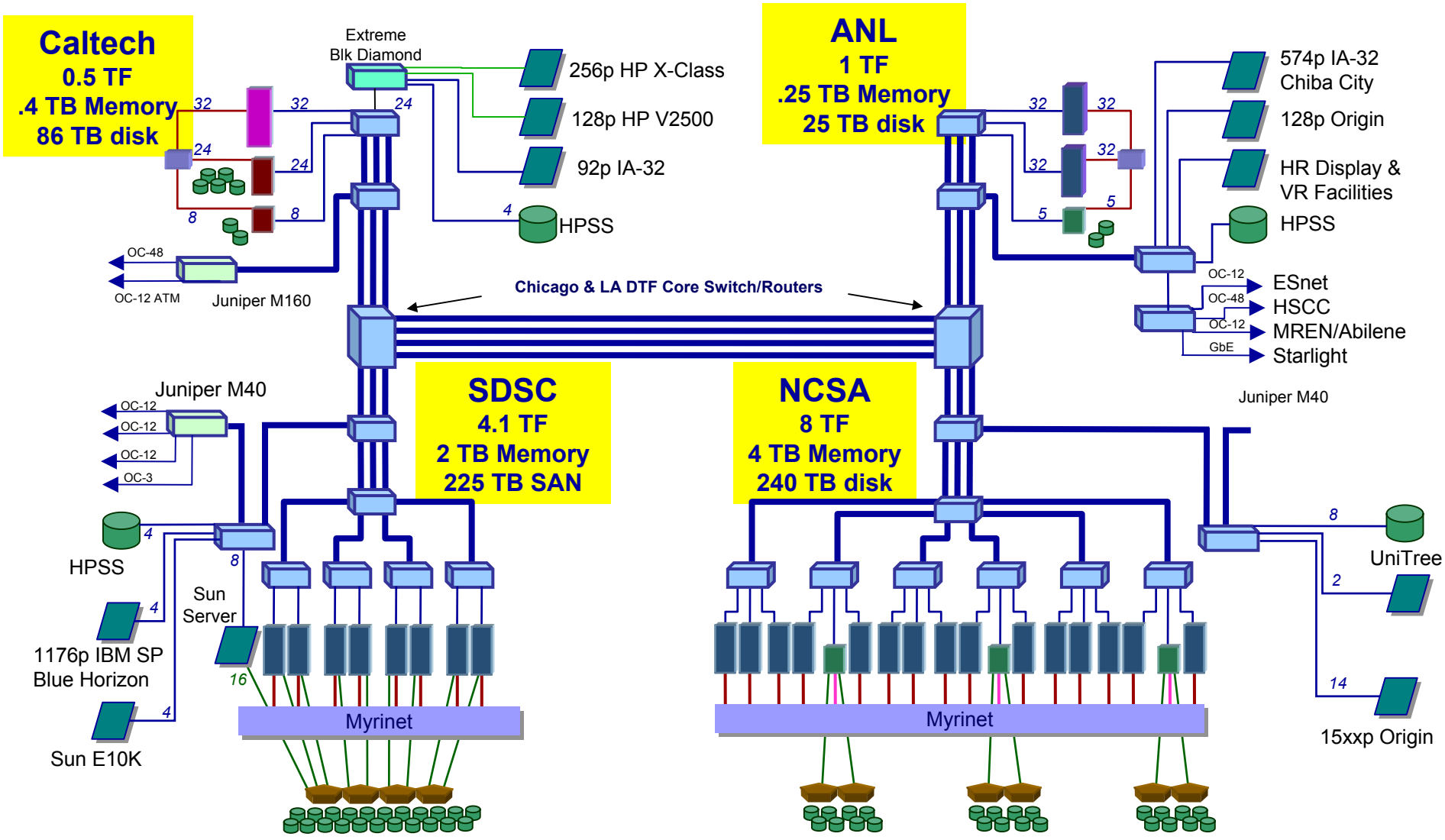
Where do we go from here?

- Improved aggregate throughput for more single streams of data
  - GridFTP
    - Striping data tranfers: Transfer file in parrallel through multiple network interfaces or just multiple ports on host
    - Using Extended Retrieve on top of striping of fetching data from multiple sites
  - Distributed UniTree (DiskXtender) cache
    - Building in another system to function as a disk cache for where data can land on based on original IP address, or GID/UID, or file size. One common namespace.

Supercomputer File systems

- NFS
  - Common name space, but performance problem
  - Used for small files for application binaries or license binaries/keys ..etc
- AFS
  - Common name space, and data movement to other machine environments (visualization) seamless, but performance is problem.
  - Used for centerwide installed packages such as perl, Matlab…
  - Users also use for small files for seamless view of the data from where ever.
- Local Scratch
  - Local filesystems for best I/O rates, but data space purged as batch jobs end. Users responsible for storing files that need to be kept.
  - Process started like this with the Crays and still running like that today

TeraGrid



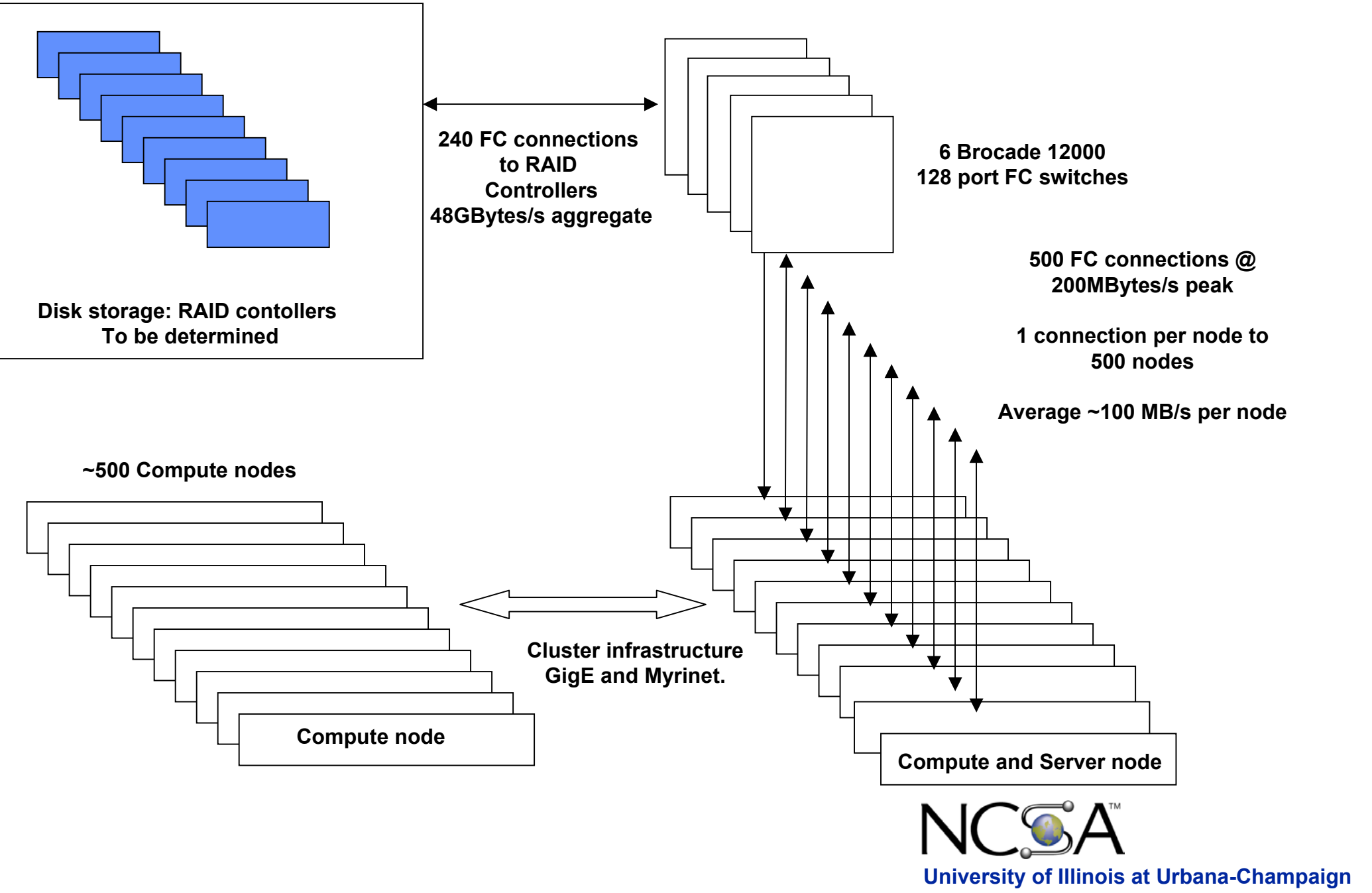
4 Clusters based on ~3300 Intel McKinley processors

NCSA™
University of Illinois at Urbana-Champaign

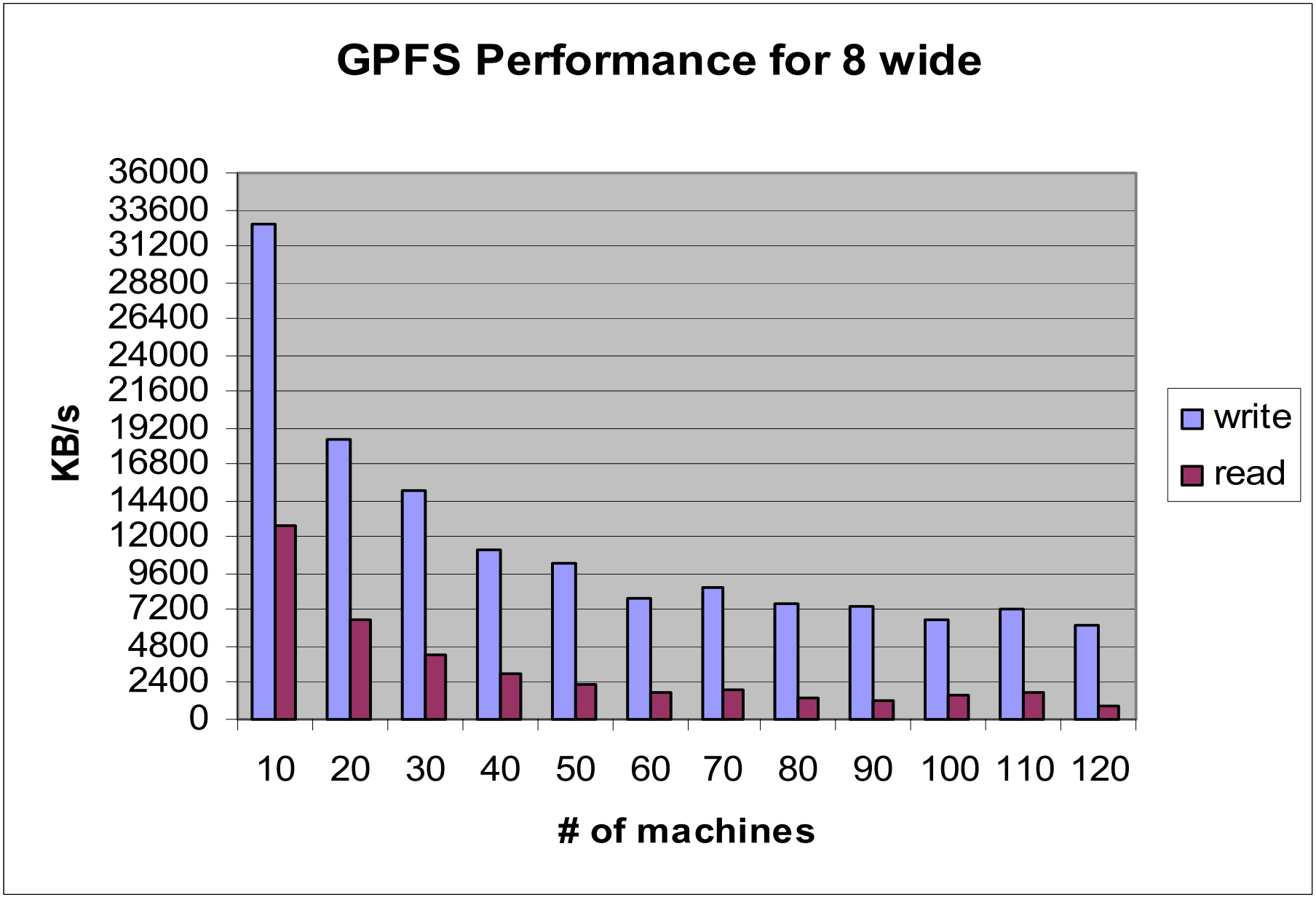
Changes in file systems for TeraGrid

- Resident data
  - Databases, Archives, Metadata services
  - Can't just be purged at end of job.
- Data needs to be seen by "all" machines. NFS does not scale to 1000 clients besides the performance being not that great.
- GPFS Global Parallel File system from IBM
  - Testing IA32
    - **128 clients with 8 servers tested**
    - **256 clients with 16 servers ongoing to be completed in May**
  - Testing IA64 Just starting that one
  - Working towards every client reading and writing to all data disks owned by "gpfs" environment within a SAN.

Strawman FC Connect for NCSA



GPFS to date


GPFS Performance for 8 wide
KB/s
36000
33600
31200
28800
26400
24000
21600
19200
16800
14400
12000
9600
7200
4800
2400
0
10
20
30
40
50
60
70
80
90
100
110
120
# of machines
write
read

Conclusions

- We have come a long way in 17 years
  - From nothing to 424TB and many iterations of system architectures and system componets (disk and tape)
- Storage scales with processor increase in the past, but will that continue?
  - Storage continues to double every year
  - Predicting what storage will be with huge jumps in processor (Teragrid) is difficult to judge
- TeraGrid changes not only mass storage components, but also file systems on supercomputers
  - Old tools don't scale (NFS; AFS; local scratch)
  - GPFS file system is being tested; working with IBM to evolve this to a large pure FC SAN file system.
  - Resident data and databases becoming very common place for supercomputer jobs.