Storage and Data Movement at Fermi Lab

Don Petravick
Fermi National Accelerator Laboratory
P.O. Box 500,Batavia IL 60510
Phone: +1-630-840-3935
E-mail: petravick@fnal.gov

Presented at the THIC Meeting at the National Center for Atmospheric Research

Boulder CO 80305-5602

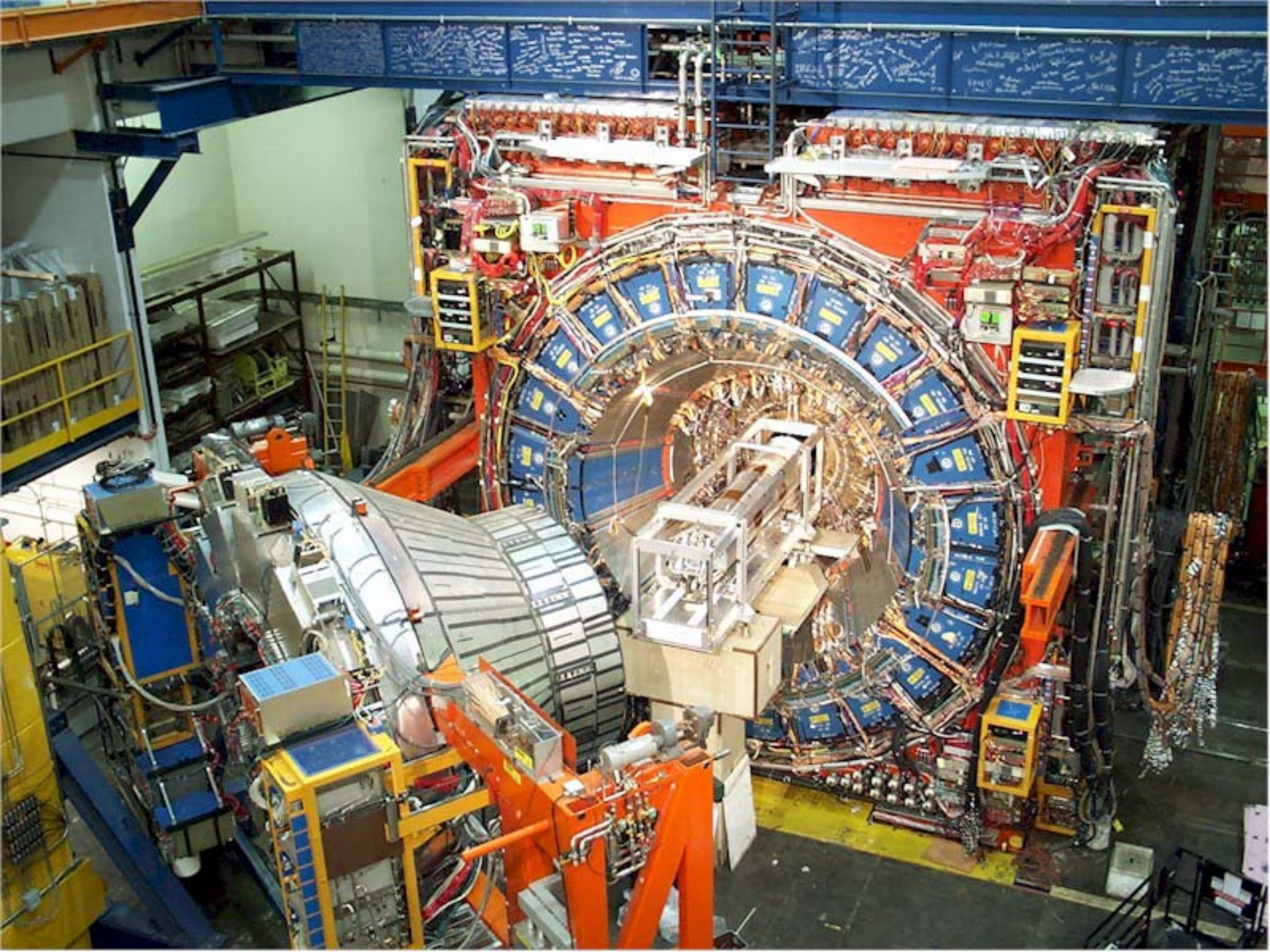
June 11-12, 2002

Storage and Data Movement at Fermilab

D. Petravick
Fermi National Accelerator Laboratory
P.O. Box 500
Batavia, IL 60510
630-840-3935
Presented at the THIC meeting NCAR, June 11, 2002

Fermilab

- Proton anti proton particle Accelerator.
- Bunches of protons, anti protons circle the ring in different directions.
- Beams are "crossed" at well defined, "instrumented" locations, (tens of) millions of times a seconds.
- Heavily filtered samples are recorded and analyzed. ~ 1PB/year for two experiments



System Model in a nutshell.

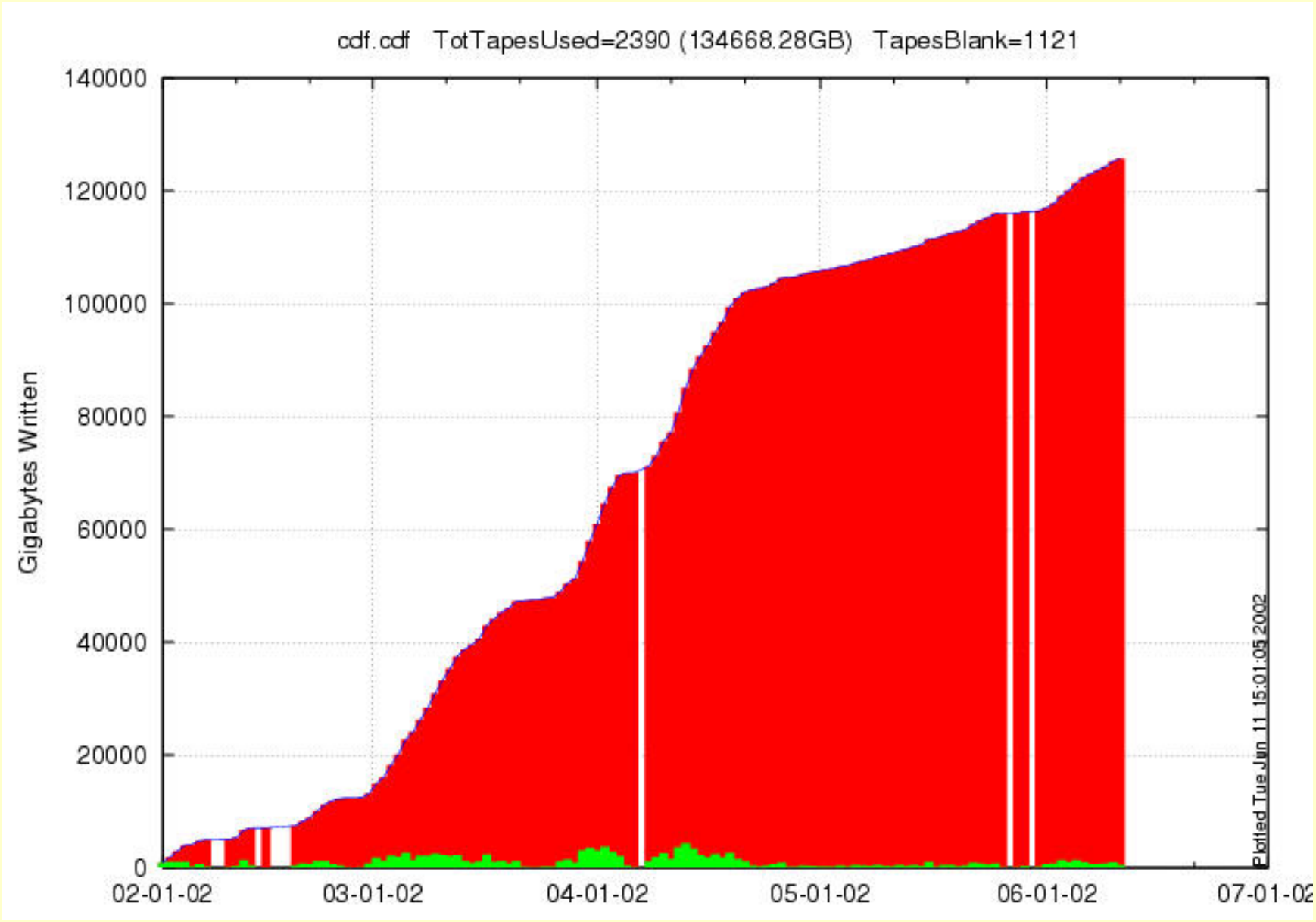
- Separate facilities for each experiment.
  - Big SGI
  - Much more CPU in "farms" of linux computers.
  - Storage, computers all on "Network" not direct attached.
- Data
  - re in flat file format.
  - RDBMS describes the files.
- Both are storage-system centric.
  - Data move to tape and then are read.
  - Moving away from this.

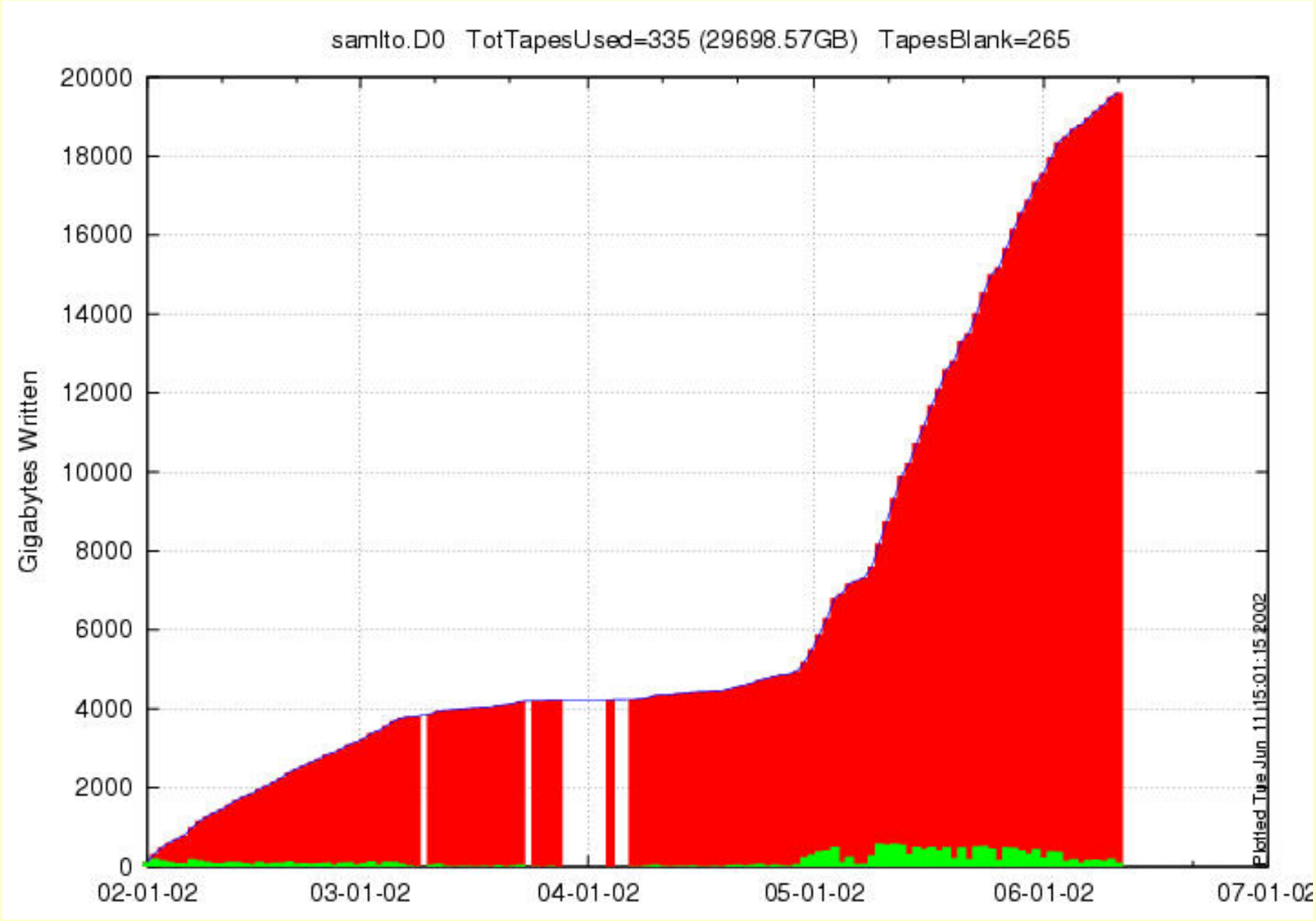
Networking

- Networks are nearly exclusively Ethernet.
- Primary infrastructure is Cisco 6509.
- Ethernet based networks are crucial for exploiting commodity technology.
  - Works well at the switch level
- Systems needs careful system design, performance analysis.
  - These are issues in the host stack and application design.
- FC as a networking technology is not expanding
  - Had been tried by CDF.
- WAN connectivity requirements increase as experimenters understand data grid computing.

Run II Permanent Storage

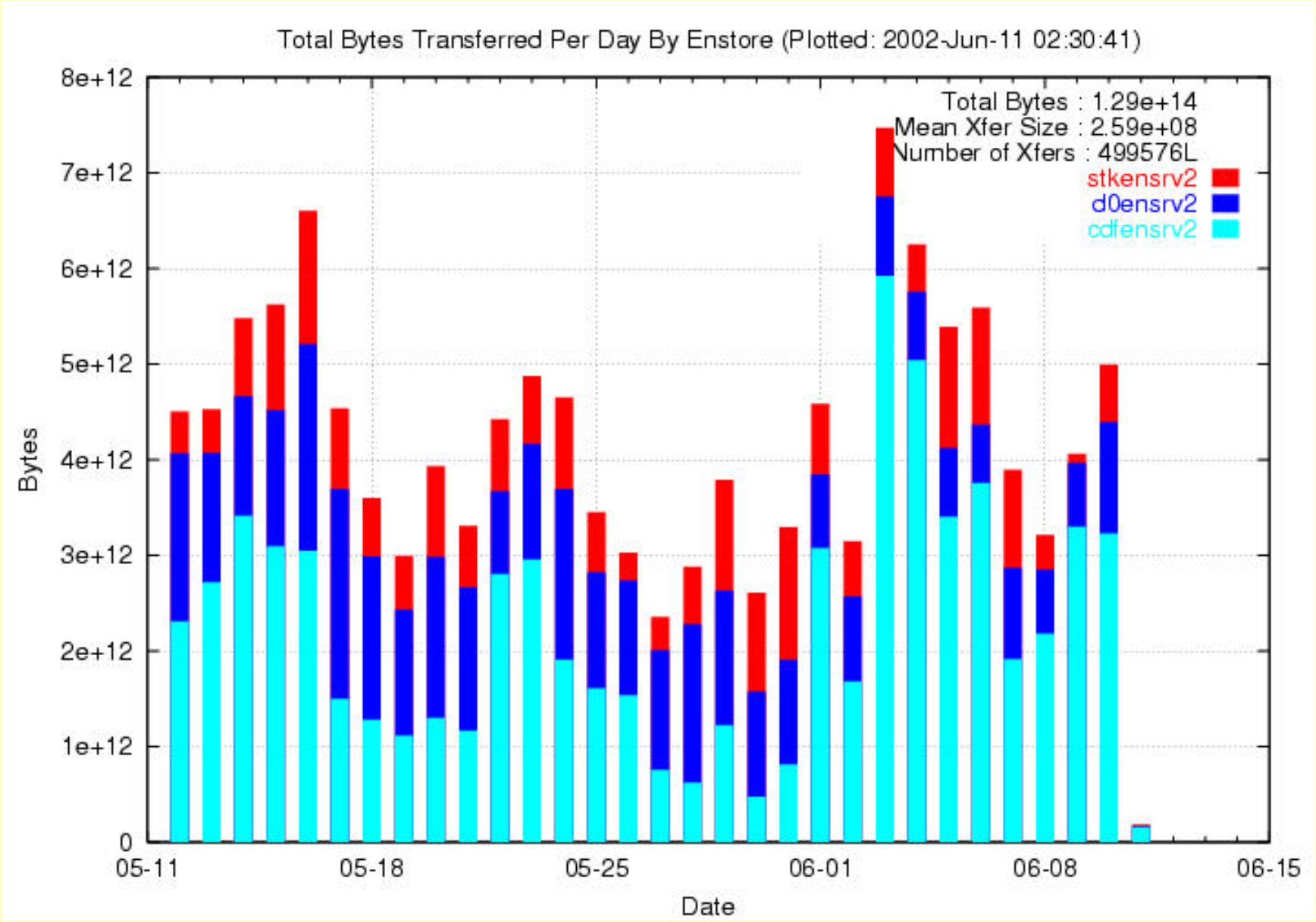
- Strategy: flexible tape media.
  - Tape is predominately STK 9940
  - LTO in AML/2 is being looked as an option.
  - FNAL is a Beta site for T9940B
  - 8mm tape (nearly) phased out.
- Main usability criteria is tolerable error rate @ ~25 TB of data movement/day
- Unendorsed for Run II – Disk replacing tape.
  - Tape is active store and permanent archive


cdf.cdf TotTapesUsed=2390 (134668.28GB) TapesBlank=1121
Gigabytes Written
140000
120000
100000
80000
60000
40000
20000
0
02-01-02
03-01-02
04-01-02
05-01-02
06-01-02
07-01-02
Plotted Tue Jun 11 15:01:05 2002


samlto.D0 TotTapesUsed=335 (29698.57GB) TapesBlank=265
Gigabytes Written
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
02-01-02
03-01-02
04-01-02
05-01-02
06-01-02
07-01-02
Plotted Tue Jun 11 15:01:15 2002

Data Movement -- Tape

- Is via TCP/IP direct to tape.
- 10 MB/sec STK 9940 is a good match for commodity networking, esp on the user's side.
- 30 MB/sec STKT9940B rates will be harder to exploit.
  - Speed up experiment end stations.
  - Forego some of the rate (but not the capacity)
  - Use path to disk. (dCache software w/ DESY)


Total Bytes Transferred Per Day By Enstore (Plotted: 2002-Jun-11 02:30:41)
Total Bytes : 1.29e+14
Mean Xfer Size : 2.59e+08
Number of Xfers : 499576L
stkensrv2
d0ensrv2
cdfensrv2
Bytes
8e+12
7e+12
6e+12
5e+12
4e+12
3e+12
2e+12
1e+12
0
05-11
05-18
05-25
06-01
06-08
06-15
Date

ENSTORE Encp History - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://www-cdfen.fnal.gov:/enstore/encp_enstore_system.html Go Links

Time	Node	User/Storage Group	Mover Interface	Bytes	Volume	Data Transfer Rate (MB/S)	User Rate (MB/S)
2002-06-11 15:30:34	fncdf76	cdfprod0/cdf	cdfenmvr10a	1053588495 (1)	from IA1725	10.1	10
2002-06-11 15:30:34	cdfensrv3	enstore/cdf	cdfenmvr5a	1007871408 (2)	from IA3421	9.98	9.89
2002-06-11 15:29:44	fncdf81	cdfprod0/cdf	cdfenmvr1a	727762455 (3)	from IA1688	10.1	10.1
2002-06-11 15:29:26	fncdf153	cdfprod0/cdf	cdfenmvr9a	1056134535 (4)	to IA2072	9.47	9.38
2002-06-11 15:28:53	fncdf76	cdfprod0/cdf	cdfenmvr10a	1069754140 (5)	from IA1725	10.1	10.1
2002-06-11 15:28:35	fncdf81	cdfprod0/cdf	cdfenmvr1a	85804077 (6)	from IA1688	10	9.56
2002-06-11 15:28:25	fncdf81	cdfprod0/cdf	cdfenmvr1a	1075127750 (7)	from IA1688	10.1	10

Internet

Evolutionary System Direction Disk:

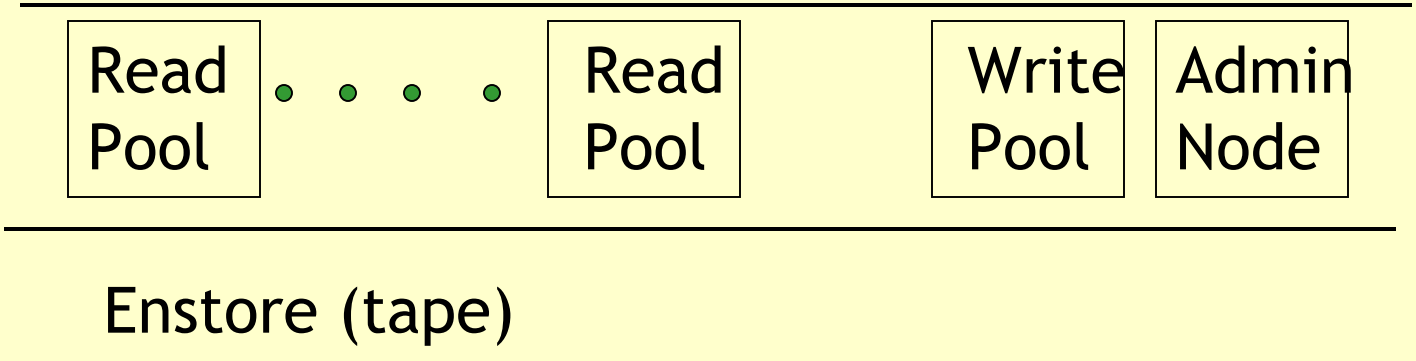
- Enabled by:
  - Super-moores law increase in disk capacity
  - Recognition of Ethernet as a competent SAN technology
  - IDE raid controllers, appropriate PC rack mount chassis
  - Linux, open source systems toolkit.
- Bypassed: Value added disk systems integrators

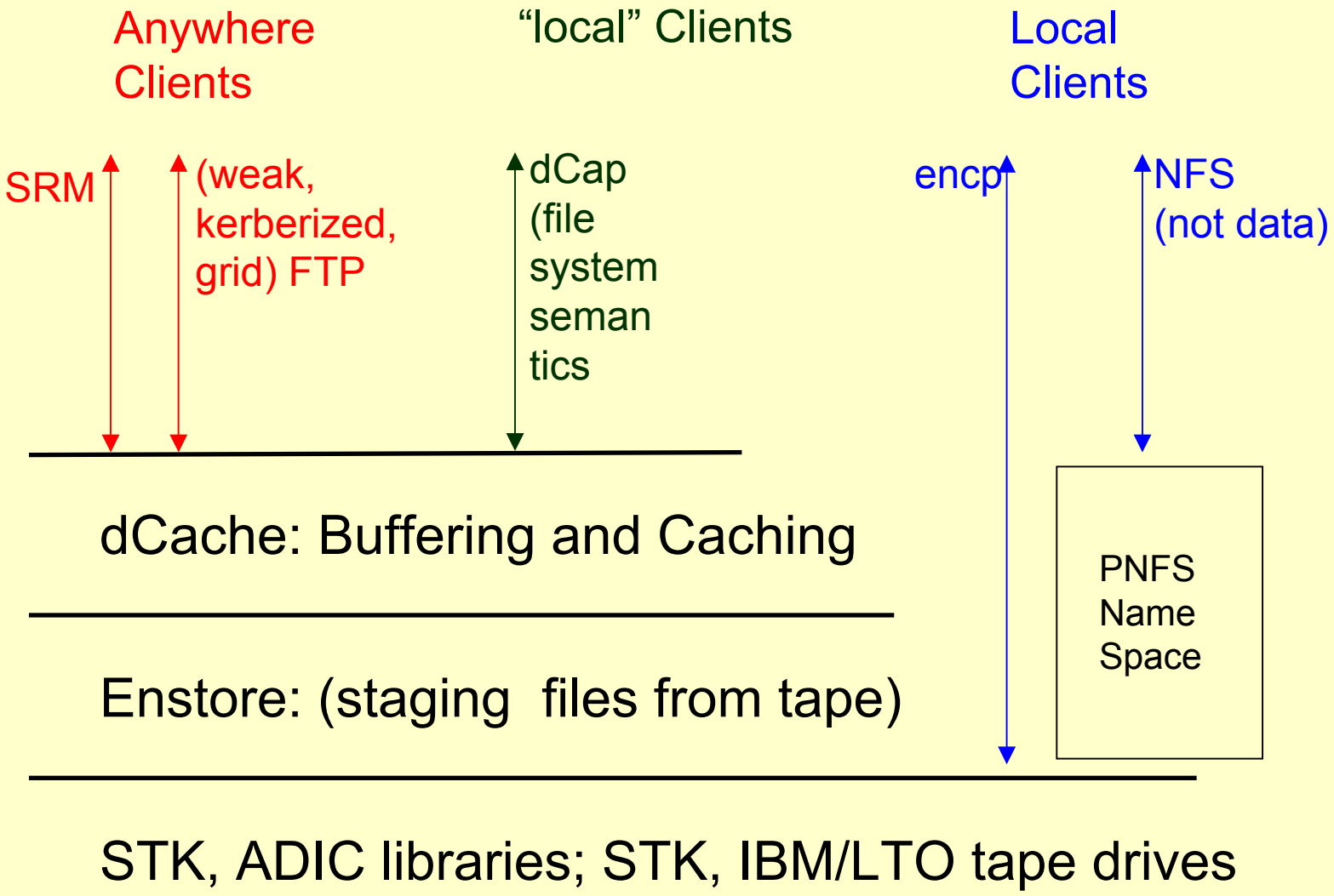
Large Disk ~ 100 TB for Scientific Data Sets

- Consensus Direction -- Linux
- Implementation – Scalable Unit
  - White Box Linux Box, ~4U chassis
  - >= 10 IDE Disk/box
  - PCI/IDE Raid controller
    - (Reason? Fear! Freaks, infant mortals, MTBF)
  - Gigabit Ethernet
- Recent scalable Unit Cost $4.65/gigabyte

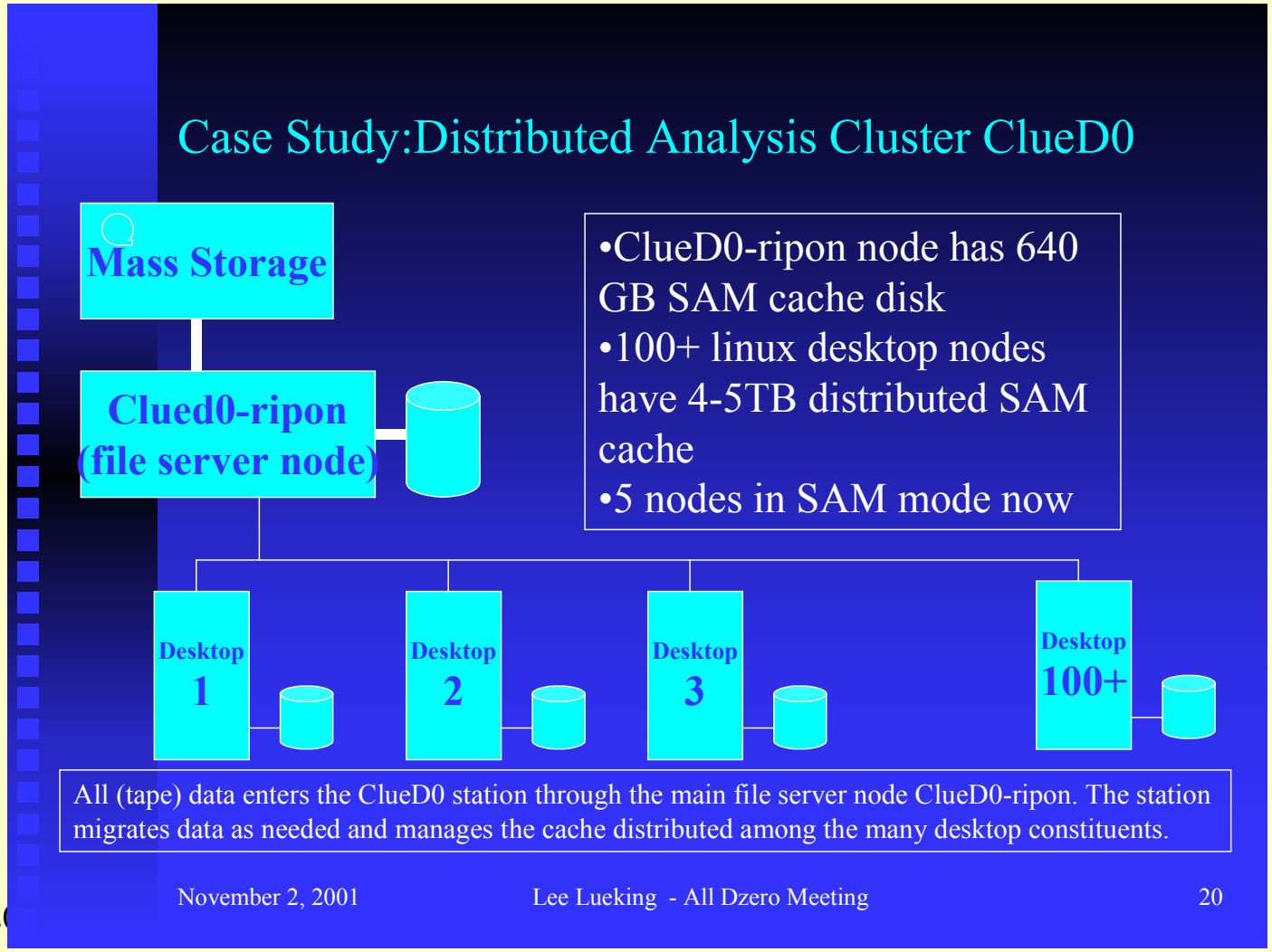
dCache

- Autonomous Pools (read and write)
- Permanence -- Backed by tape
- Protocol independent architecture
- Usability Model – Read and Write Pools
- Collaboration w/ DESY, written in JAVA



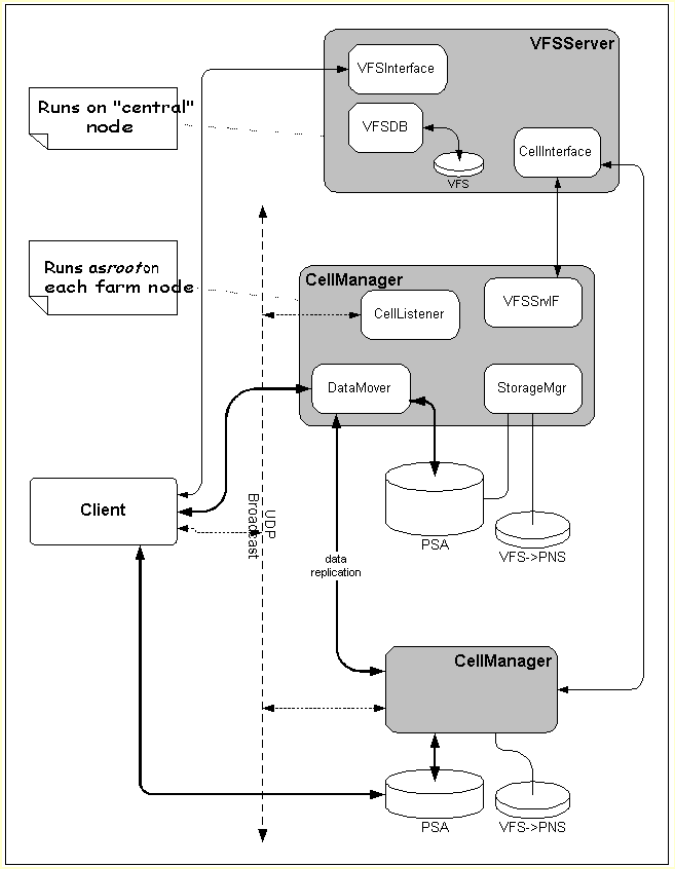

Anywhere Clients
"local" Clients
Local Clients
SRM
(weak, kerberized, grid) FTP
dCap (file system seman tics
encp
NFS (not data)
dCache: Buffering and Caching
PNFS Name Space
Enstore: (staging files from tape)
STK, ADIC libraries; STK, IBM/LTO tape drives

Linux, Box, Staging middleware

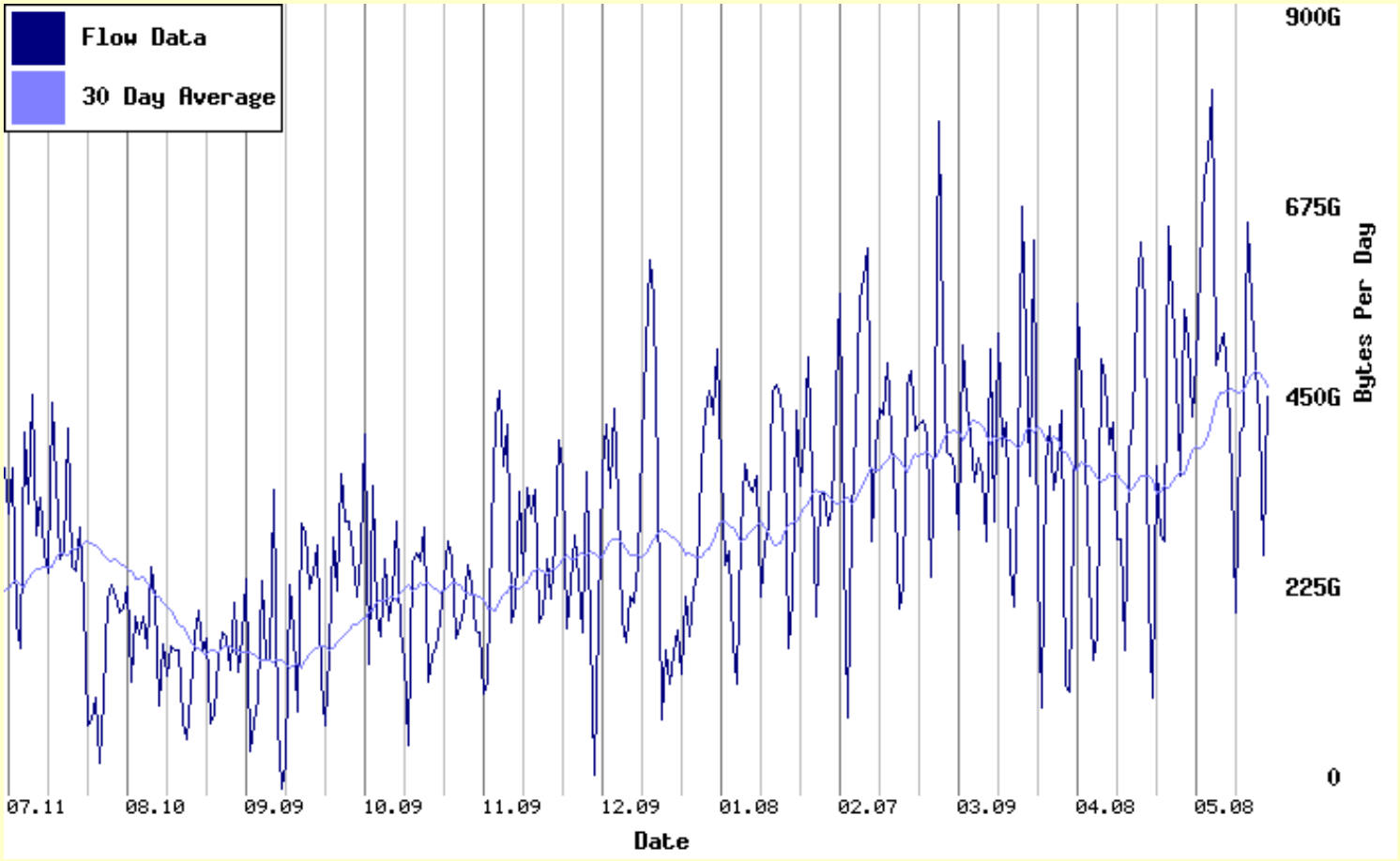


Bonus Disk -- dFarm

- Exploit the excess disk + competent networks we are currently blessed with.
- Provide redundant, temporary storage, robust against failure of any farm node



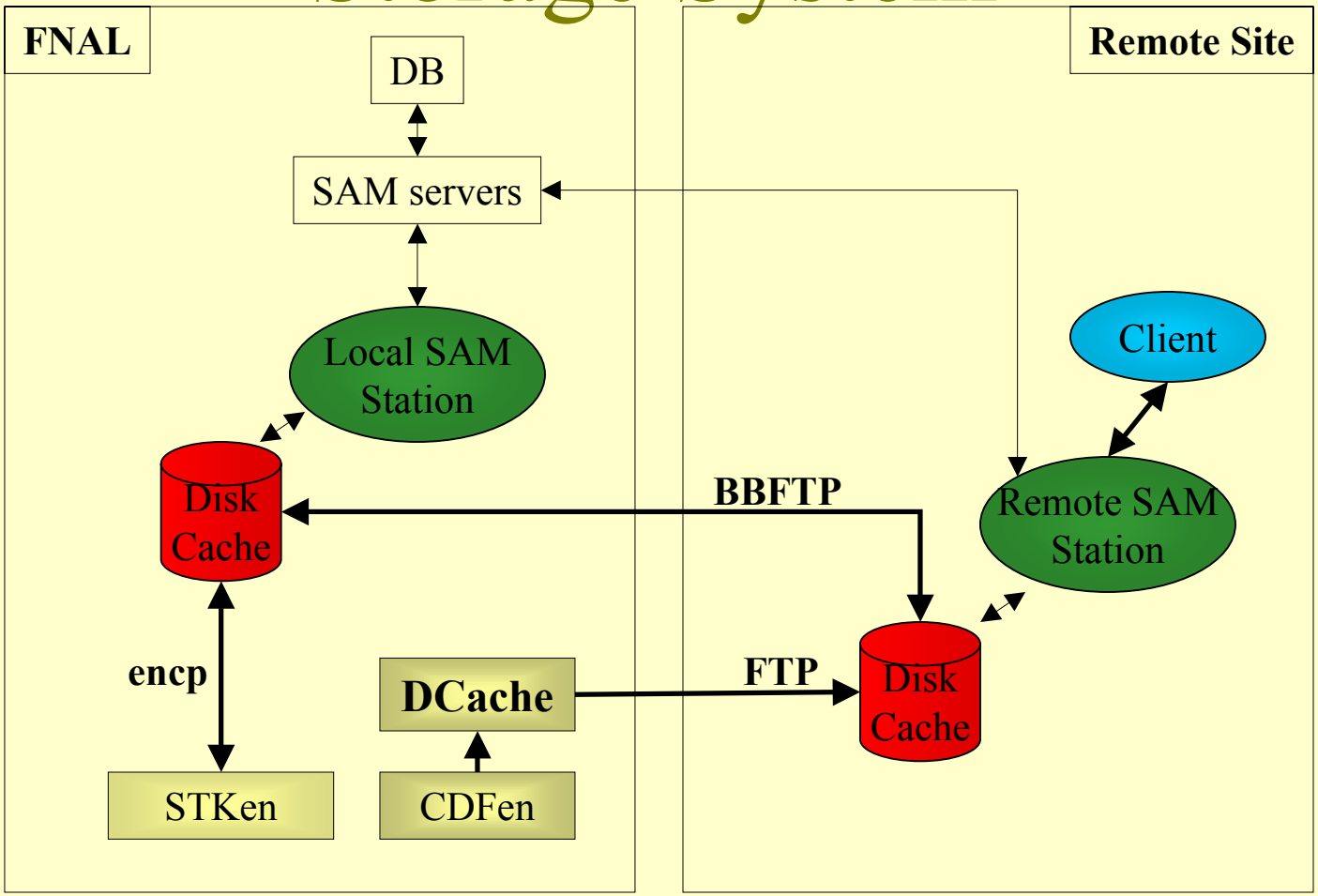
FNAL WAN throughput



Demise of Tape for inter-lab transfer

- Embracing Network for off-site data transfers.
- Aspiring to grid middleware.
  - Ftp -> Grid FTP
  - Storage Resource Manager Interfaces
- Comments
  - tape meta data , format was never std.
  - Demise of desktop tape drives
  - Enabled by network, big disks

Grid Techniques direct from Storage System



Next FNAL experiment BTeV

- 2000-6000 CPUs
- Runs in 2005.
- TB disks? Might get "storage for free"
- Desire to exploit grid technologies to enable University collaborators.
- Tape no tin their current TDR.
  - Permanency via redundancy.

Summary:

- Compared to plans, we used high end tape.
- Use of Linux has been very important.
- As a networking technology, only ethernet withstood the test of bringing large systems to bear.
- New experiments planning for the ~2005 era are not convinced of the value of tape.
- Data distribution over wide area networks are receiving a lot of attention