Run II Data Handling Infrastructure

D. Petravick

System Model in a nutshell.

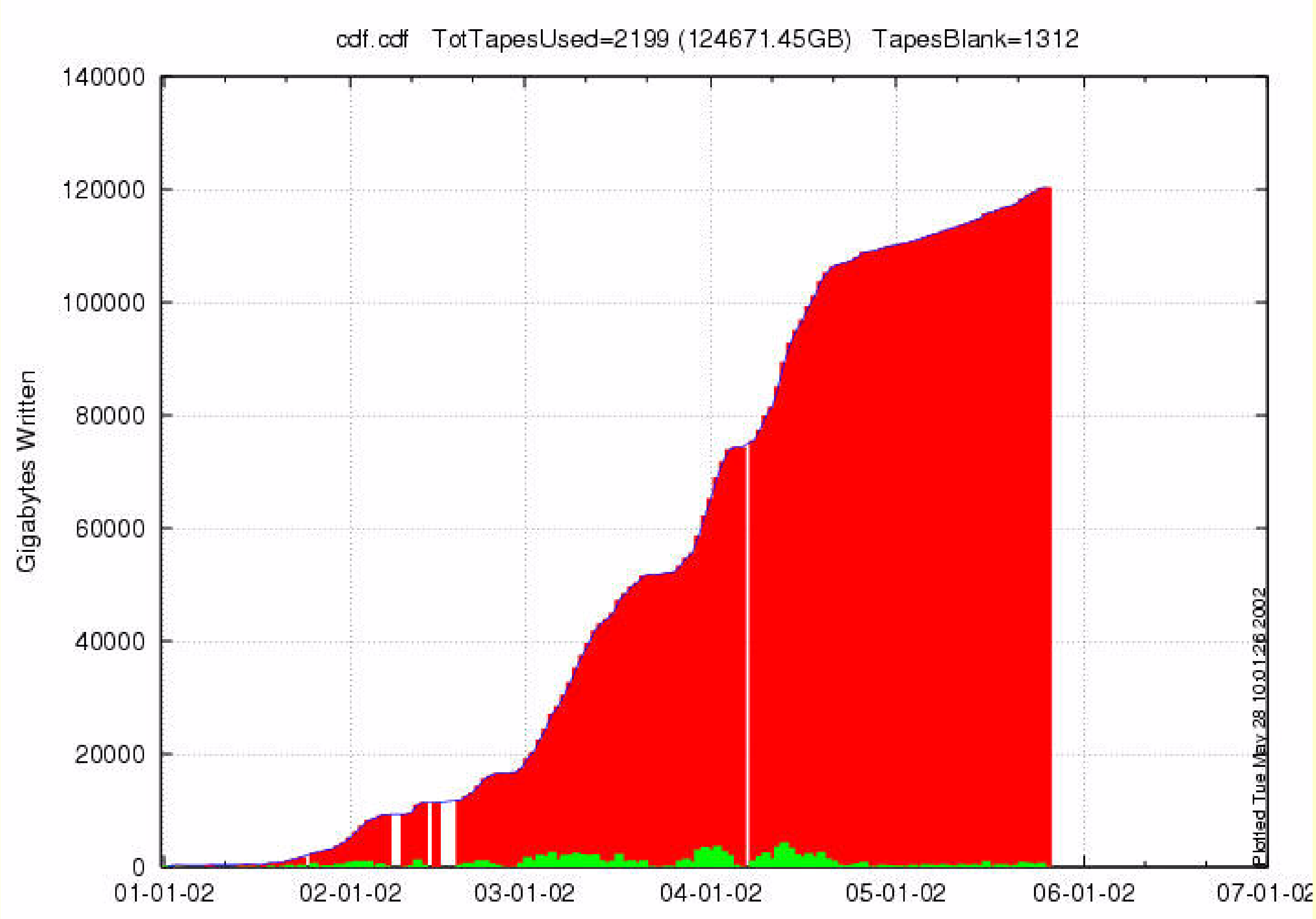
- Separate facilities for each experiment.
  - Storage, production farms, analysis.
- Data are in flat file format.
- RDBMS describes the files.
- Both are storage-system centric.
  - Moving away from data hits tape before processing.

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.

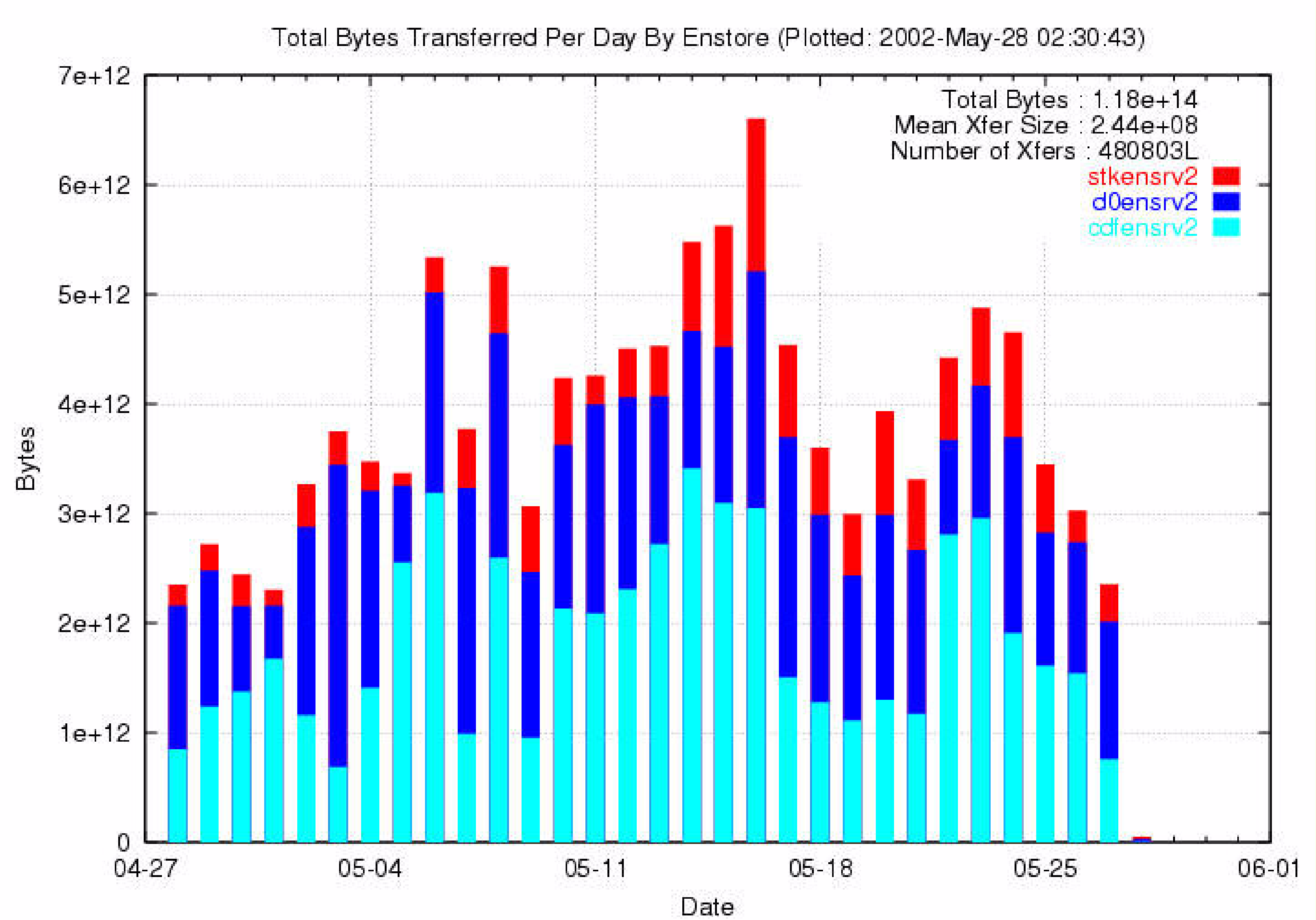
Permanent Storage

- 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.
- Sunken cost, capacity in AML/2
- Used STK silos available at a reasonable $$.
- Main usability criteria is tolerable error rate @ ~25 TB of data movement/day
- Unendorsed – Disk replacing tape.


cdf.cdf TotTapesUsed=2199 (124671.45GB) TapesBlank=1312
Gigabytes Written
140000
120000
100000
80000
60000
40000
20000
0
01-01-02
02-01-02
03-01-02
04-01-02
05-01-02
06-01-02
07-01-02
Plotted Tue May 28 10:01:26 2002

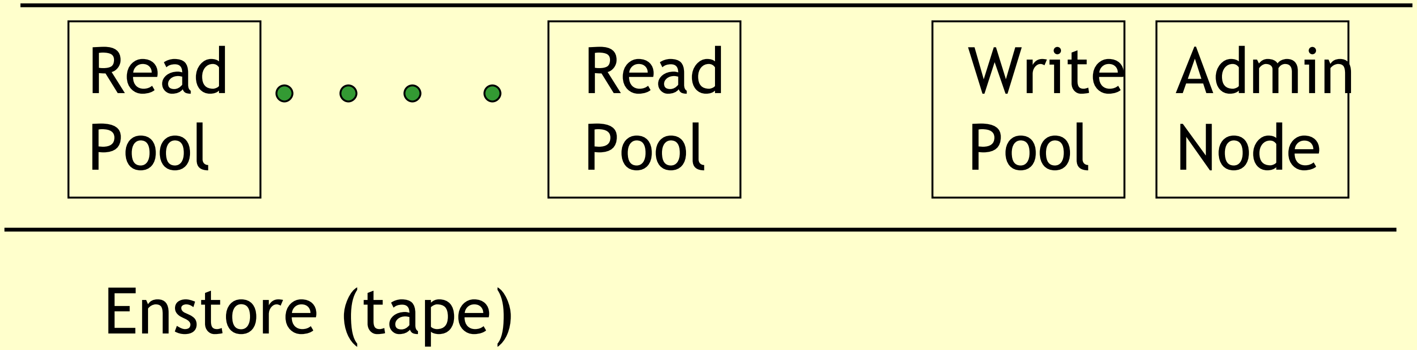
Data Movement -- Tape

- 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-May-28 02:30:43)
Total Bytes : 1.18e+14
Mean Xfer Size : 2.44e+08
Number of Xfers : 480803L
stkensrv2
d0ensrv2
cdfensrv2
Bytes
7e+12
6e+12
5e+12
4e+12
3e+12
2e+12
1e+12
0
04-27
05-04
05-11
05-18
05-25
06-01
Date

dCache

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

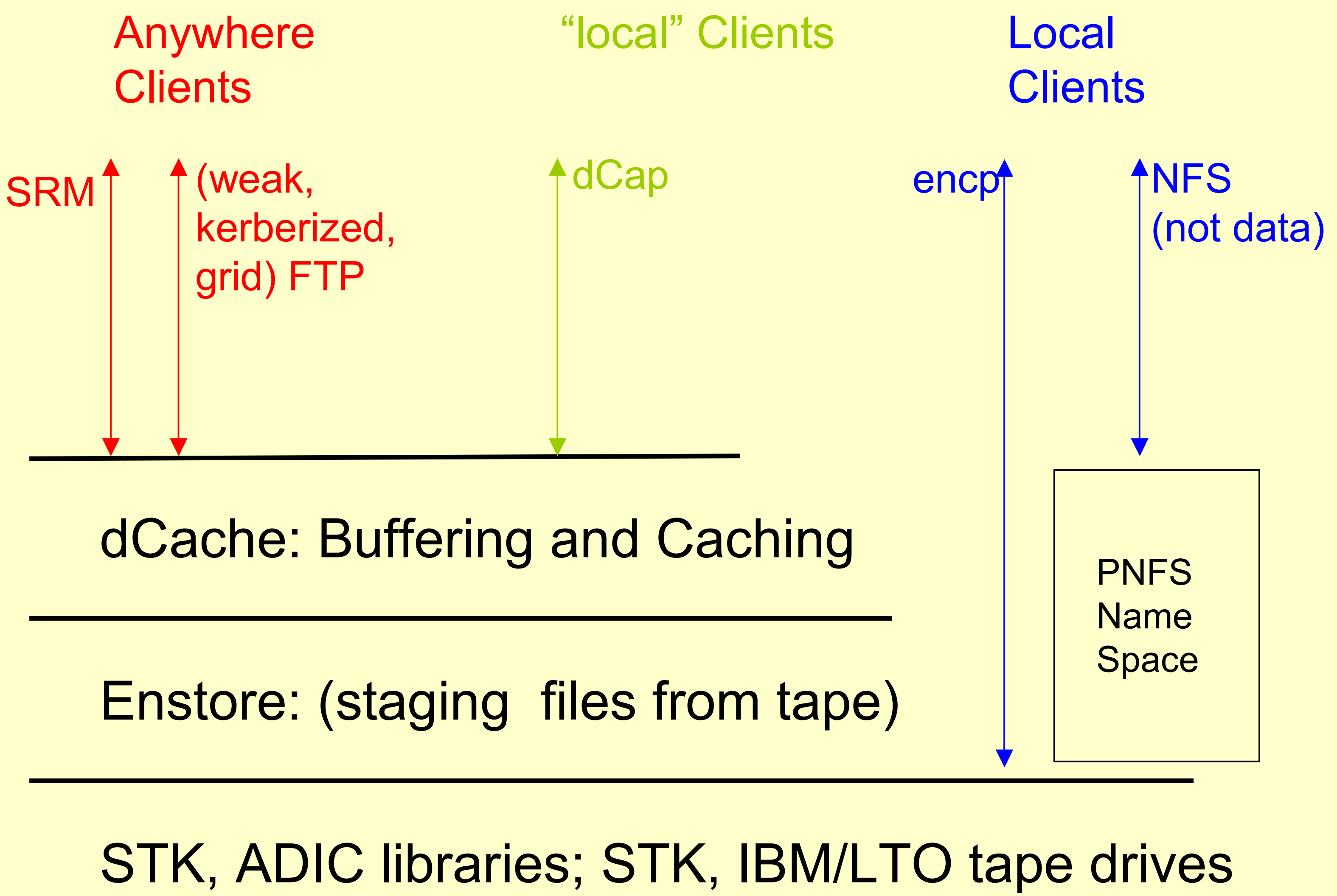


Read Pool . . . . Read Pool

Write Pool

Admin Node

Enstore (tape)



Anywhere Clients

"local" Clients

Local Clients

SRM

(weak, kerberized, grid) FTP

dCap

encp

NFS (not data)

dCache: Buffering and Caching

PNFS Name Space

Enstore: (staging files from tape)

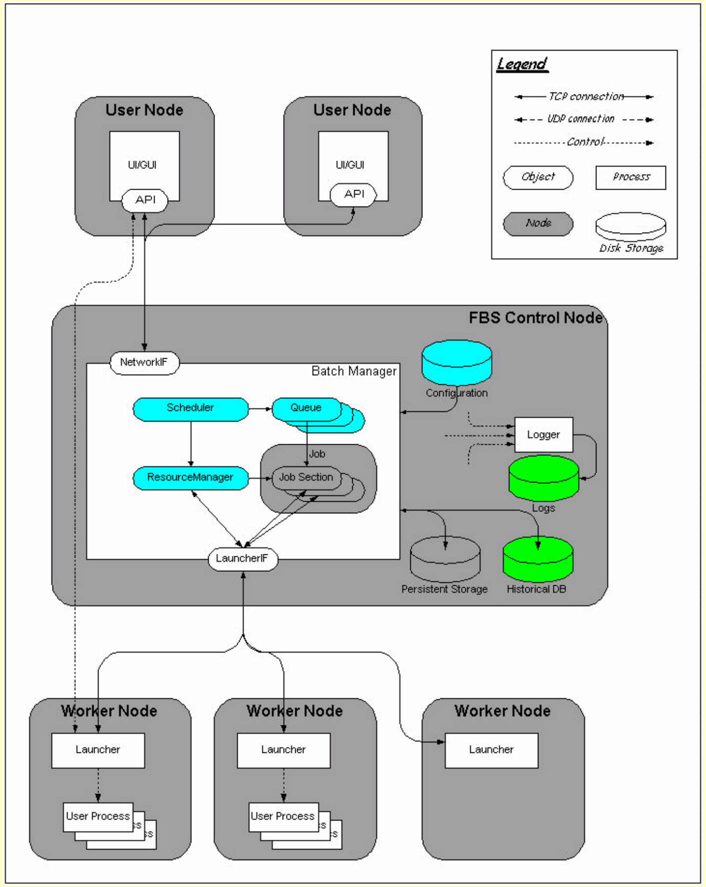
STK, ADIC libraries; STK, IBM/LTO tape drives

Oracle DBMS

- Run II has used a model of meta-data in a relational DBMS.
- DBMS is Oracle.
  - CDF/SAM is an example of the payoff.
- The run II DBMS systems have high availability requirements.
    - Current practice is to use platforms and tools that the vendor feels is a high availability environment.
    - These platforms may not be available at commodity prices.
      - ODS have a project to investigate linux as an oracle platform.
      - http://fncduh1.fnal.gov/supportdb/linux_strategy.html

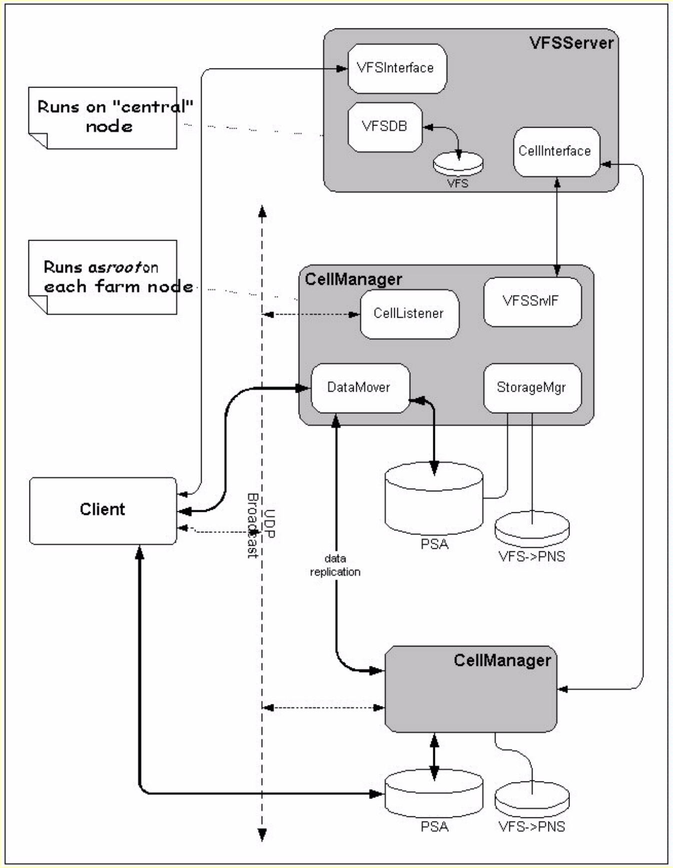
Production Farming

- Is done on linux farms of whitebox duals.
- Is supported by locally written infrastructure
  - FBS – batch system
  - Data handling primitives (FCP, and such)
  - And…



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



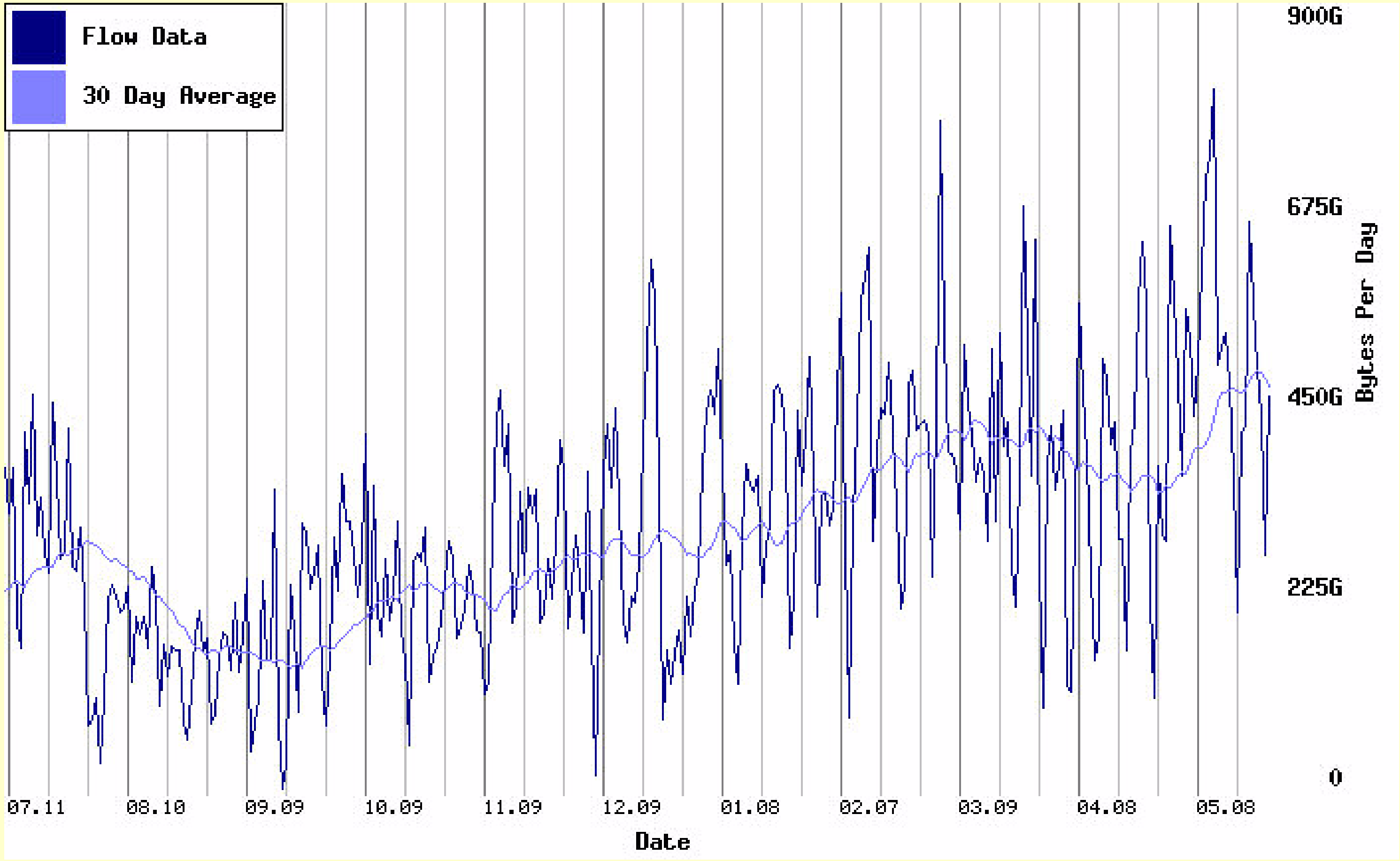
Recent System Developments

- Changing security environment
- Embracing Network for off-site data transfers. + advent of grid middleware.
- Investigating linux as opposed to SGI SGI for new procurements

Computer Security Strong Authentication

- Required to exec(), and "moral equivalent of FTP"
- FNAL infrastructure is Kerberos.
- PPDG Site AAA prototype to adapt to X509 *while things settle*.
  - FNAL CA
  - Kerberos password -> Kerberos ticket -> temporary cert signed by FNAL.
- Other concepts coming, VO, CASish things, etc.

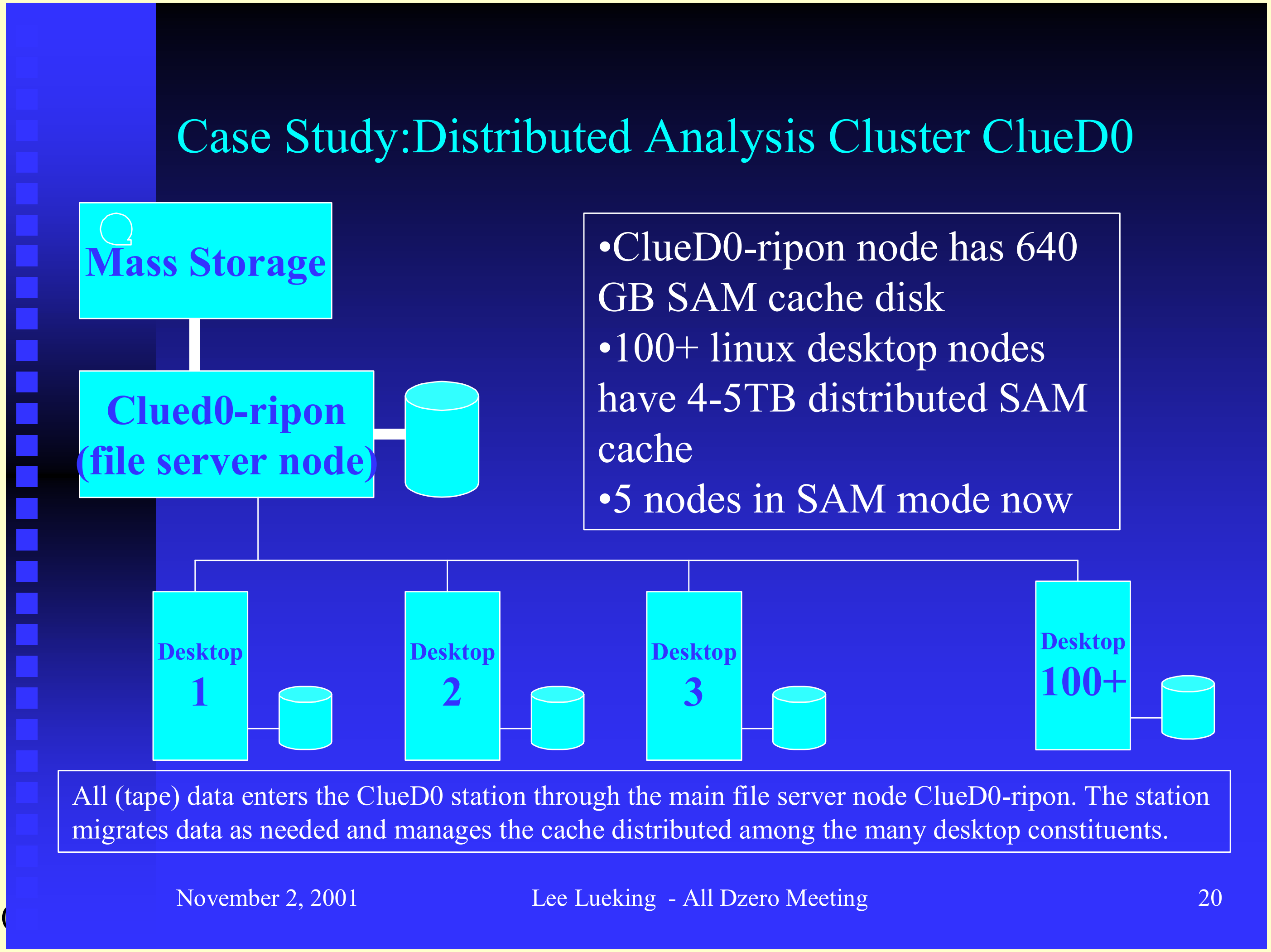
FNAL WAN throughput



Evolutionary System Directions

- Use of Grid techniques, sharing with wide area and local area caches
- Use of SAM to manage rolling staging on local disks of modest Linux boxes.
- Use of file-system like access directly over a competent local area network
- Use of excess disk in farms as a robust, semi persistent temporary area

Linux, Box, Staging middleware



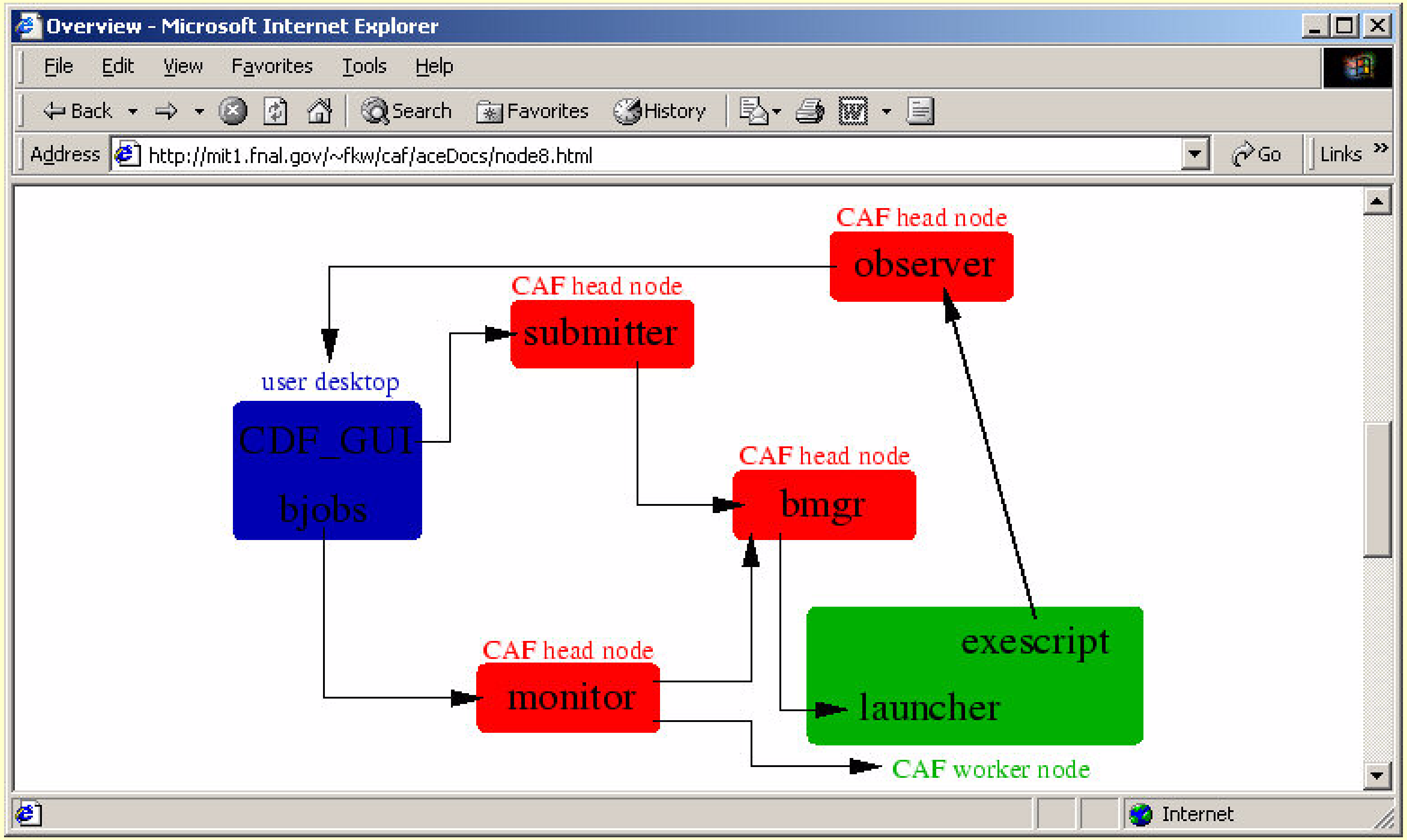
Case Study:Distributed Analysis Cluster ClueD0

- ClueD0-ripon node has 640 GB SAM cache disk
- 100+ linux desktop nodes have 4-5TB distributed SAM cache
- 5 nodes in SAM mode now

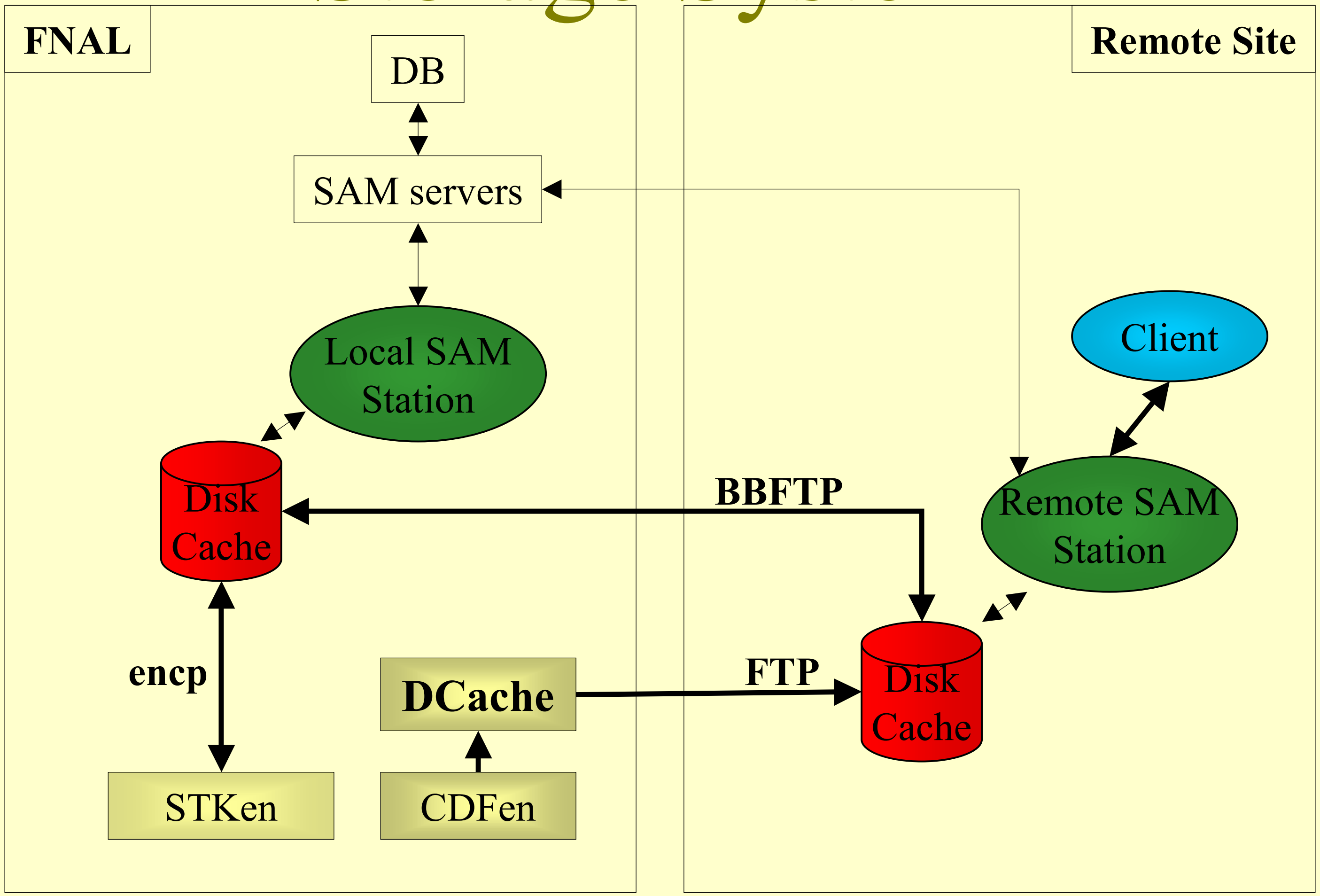
All (tape) data enters the ClueD0 station through the main file server node ClueD0-ripon. The station migrates data as needed and manages the cache distributed among the many desktop constituents.

November 2, 2001 Lee Lueking - All Dzero Meeting 20

File system access (CDF CAF).



Grid Techniques direct from Storage System



Fabric-awareness in the infrastructure

- FBS and other batch interfaces are (becoming) GRAM aware.
  - Feedback to grid schedulers is a developing area.
- Fermilab's Strong Authentication policy is being adapted to X509 certificates.
  - Site AAA project
- Grid FTP, SRM interfaces planned for the dCache.
- Understanding "monitoring"