Data Center Experience -- FNAL

D. Petravick

Philosophy

- Raw intellectual materials:
  - Experimenters
  - Computing professionals, et al.
  - Both groups have insight into the domain and its problems.
  - Both groups have technical capabilities, includign computer science capabilities.
  - Both groups create things
- Intellectual collaboration with experiments.

Principles

- Avoid the thin interface w/ experiment
  - Collborate deeply
  - Mutual learning
- Do the mundane things well.
  - Gives credibility.
- Scientific program is large:
  - CDF, D0, CMS are big dogs.
  - Many others viz MiniBoone, Theory, Minos…
  - These are important as well.
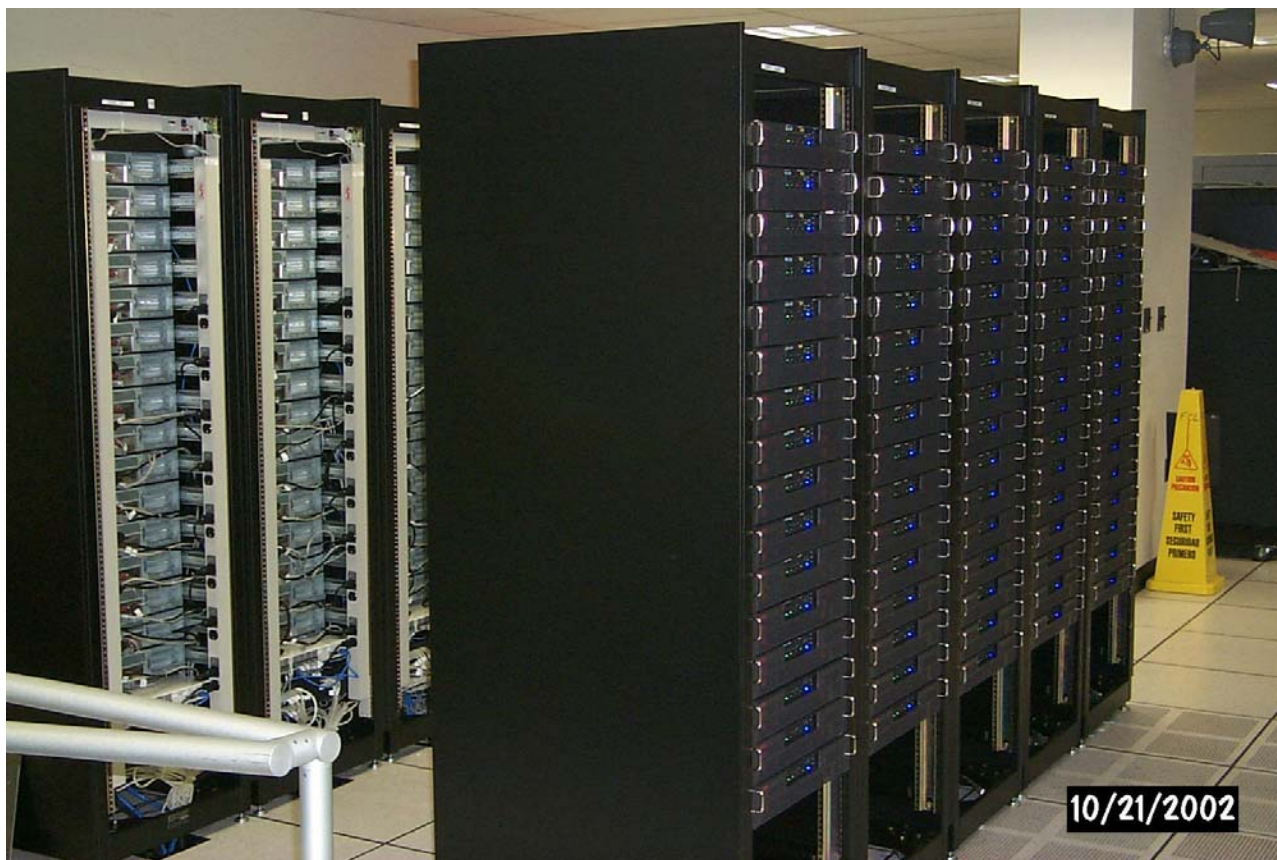
Important technical Entities

- Connectivity – Networks
  - Interest is exclusively Ethernet.
- Computational Systems – End nodes.
- Storage
- Facility middle ware.
- Experiment data handling system.
- Administration and operations support.

Connectivity

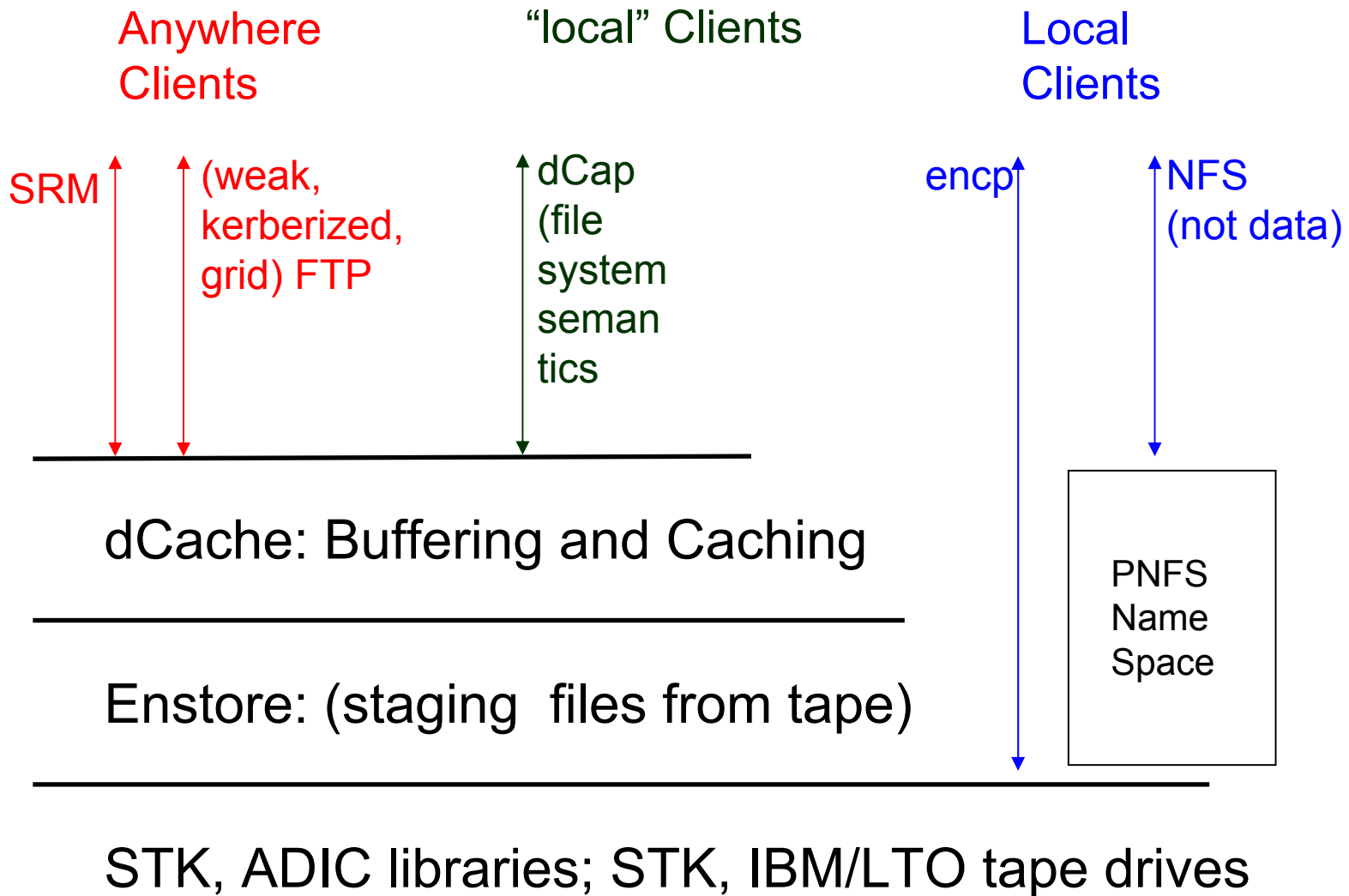- One dominant technology
  - Standard MTU Ethernet.
  - Interrupt coalesing
- Not adopted:
  - Fibre channel as a network
  - Large ethernet MTU.
- Cisco 650x as a workhorse.

Computational System

- Significant SGI presence
  - CDF actively removing
- Linux "white box" has been the focus for a long time.
- Concurrent skill base:
  - Coherent body of system administrator knowledge.
 - Fermlab RHL, kind of a config management

10/21/2002

10/21/2002

Anywhere
Clients

"local" Clients

Local
Clients

SRM

(weak,
kerberized,
grid) FTP

dCap
(file
system
seman
tics

encp

NFS
(not data)

dCache: Buffering and Caching

Enstore: (staging files from tape)

PNFS
Name
Space

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

Storage(disk)

- Main development directions
  - Re-use extra disk on farm nodes
  - Linux "TB file server"
- Ethernet Transport
- Fibre channel file systems tried, left behind.
  - Meta problem – distributed management of block level device.
- Some interest in OSD.
- Touchy integration (performance)
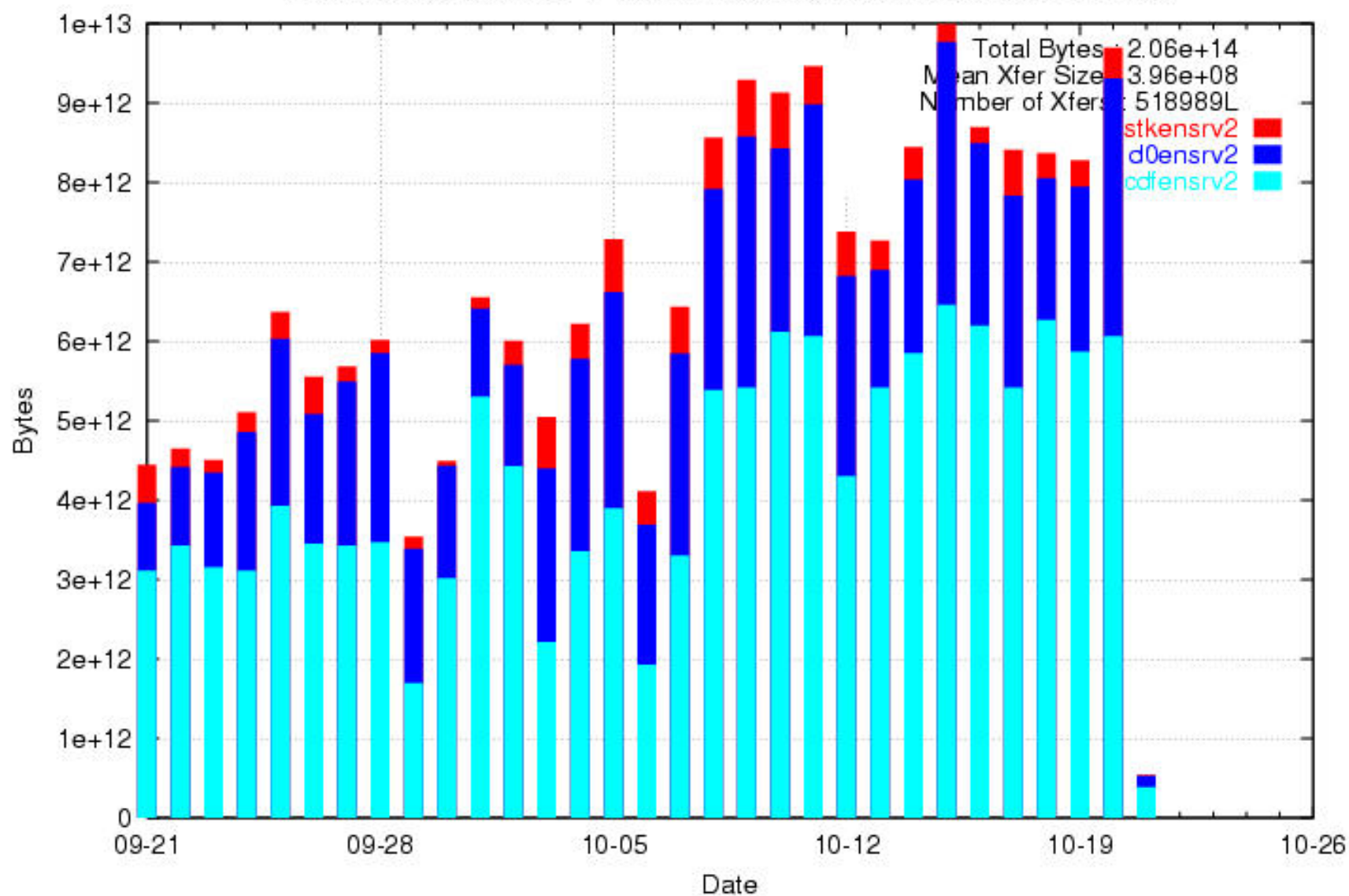
10/21/2002
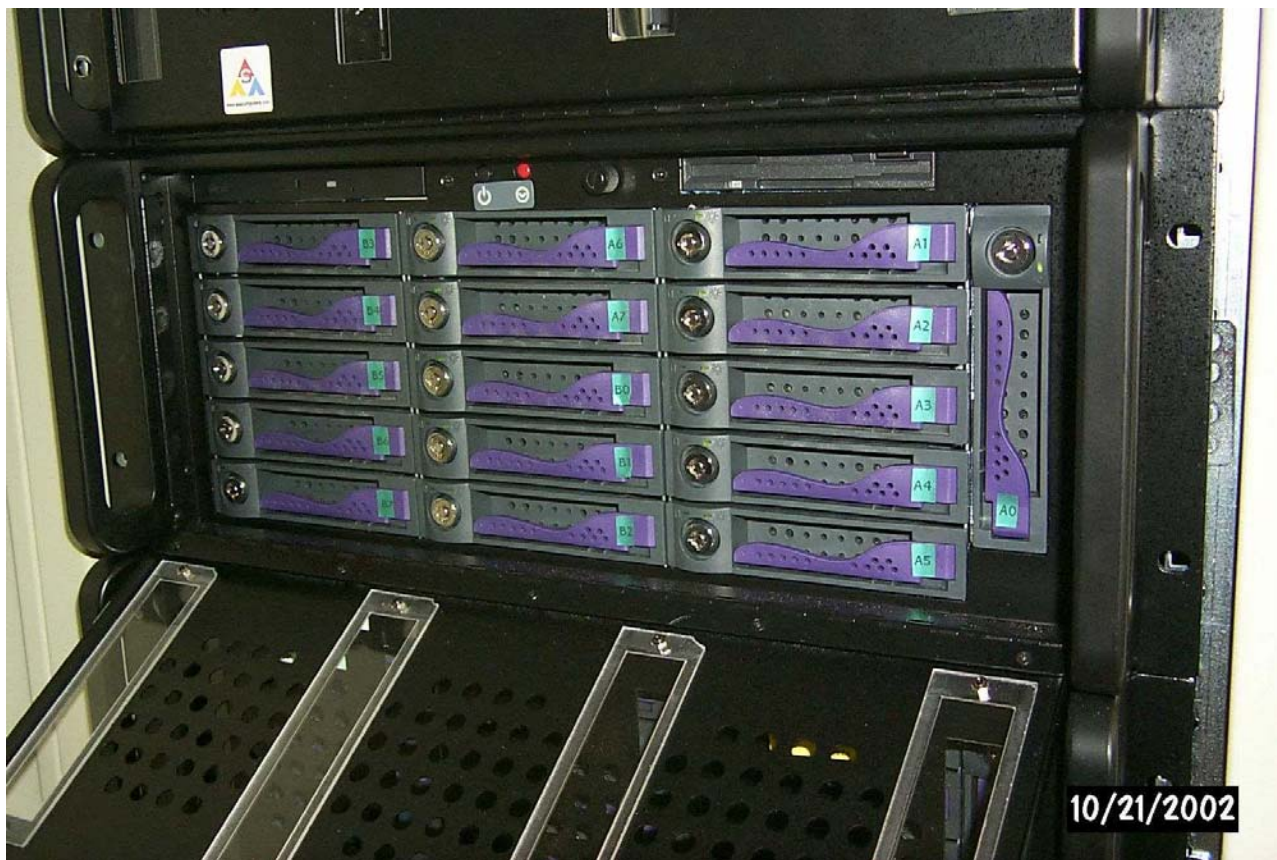
Storage(tape)

- STK 9940 in production
- IBM LTO + ADIC aml/2 in production
- STK T9940B (30 MB/sec, 200GB)
  - EPE
  - Being installed.
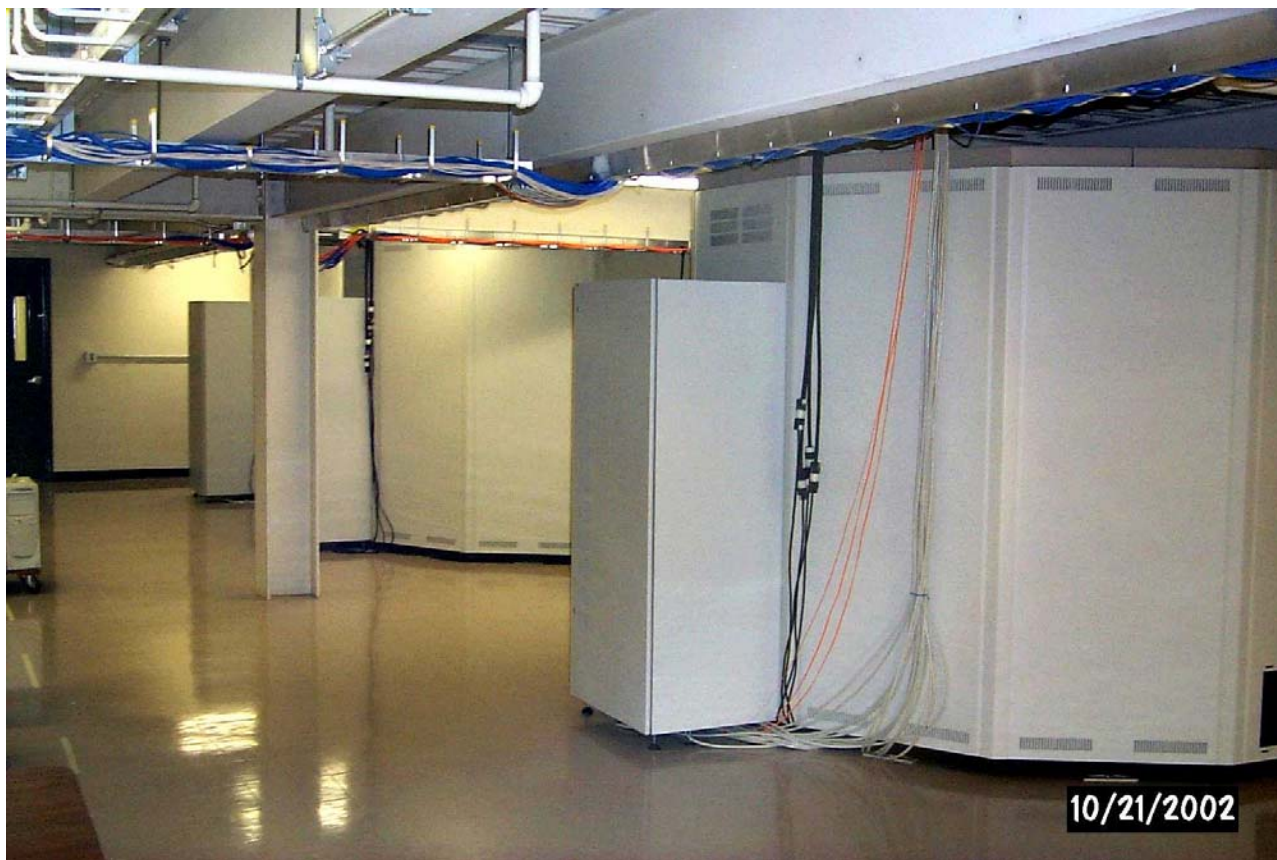
Storage and Data Movement Middleware - ENSTORE

- Staging from tape
- Very scalable, suitable for the campus and daq.
  - Best to source/sink the drive rate.
  - Simple minded authentication
  - Local protocols
- Scheduling and balancing suitable for data acquisition.

Total Bytes Transferred Per Day By Enstore (Plotted: 2002-Oct-21 02:30:42)

Total Bytes : 2.06e+14
Mean Xfer Size : 3.96e+08
Number of Xfers : 518989L

stkensrv2
d0ensrv2
cdfensrv2

Storage and Data Movement. – Dcache

- Functions
  - Rate adapt.
  - Failure model
  - More abstract interface than Enstore
  - Larger scope of service – wan as well.

Protocols

- Grid/wan <u>and</u> LAN access
- Types of access:
  - Staging
  - Intra file (I.E posix like) access
  - Management
- Desire to become symmetric with grid "storage element" (whatever _that_ is)

Congestion Controls

- Ethernet based systems must provide high level congestion control.

- Each pool serves only N streams to users.
    - Hope to move to rate-based controls.

- Each pool will only have M transfers to Enstore.

- We can apply discipline on the Enstore side.

FTP

- Weakly Authenticated FTP(1)
  - Augmented with MODE E
- Kerberized FTP
- Grid FTP
  - Custom server in dcache
  - Contributed protocol analysis at GGF

Storage Resource Manager(1)

- Example functions:
  - Pin, space reservations
- Implemented V1 of the protocol.
  - Our client works at Jlab.
  - Web services.
- Integrated into CDF AC++ framework.
  - For scaling.
  - Provides C Web services Client.
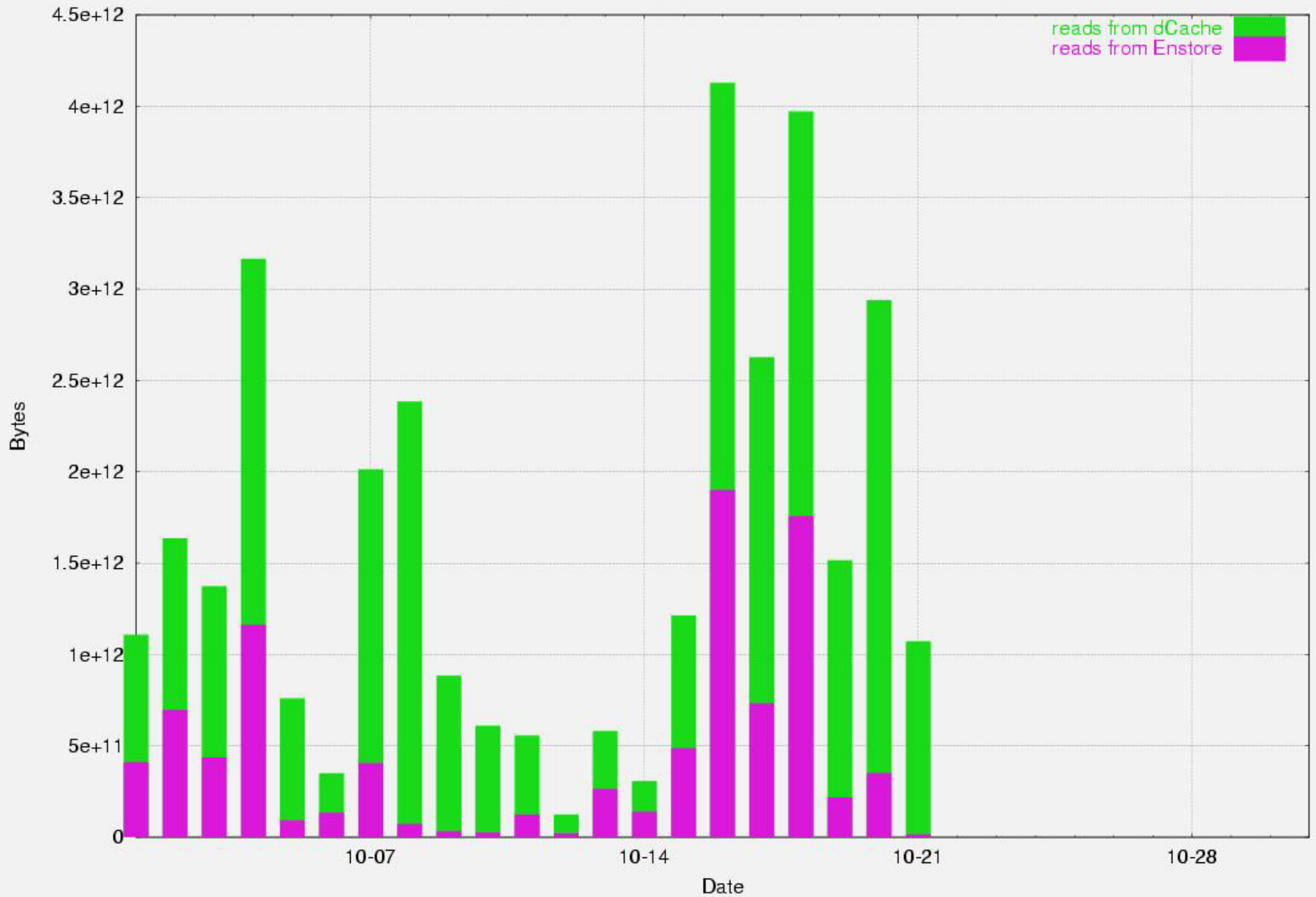- Preparing for V2 discussions.

Dcap

- The workhorse LAN protocol for Dcache.

- Intra file access as well as staging.

- Management functions like pinning exposed as well.

- Does not require PNFS mount.

Dcap

- Kerberos support in progress (DESY).

- Integrated into:
  - Root I/O (tDCacheFile).
  - Ac++.
  - Objectivity/AMS.
  - Large use at DESY.

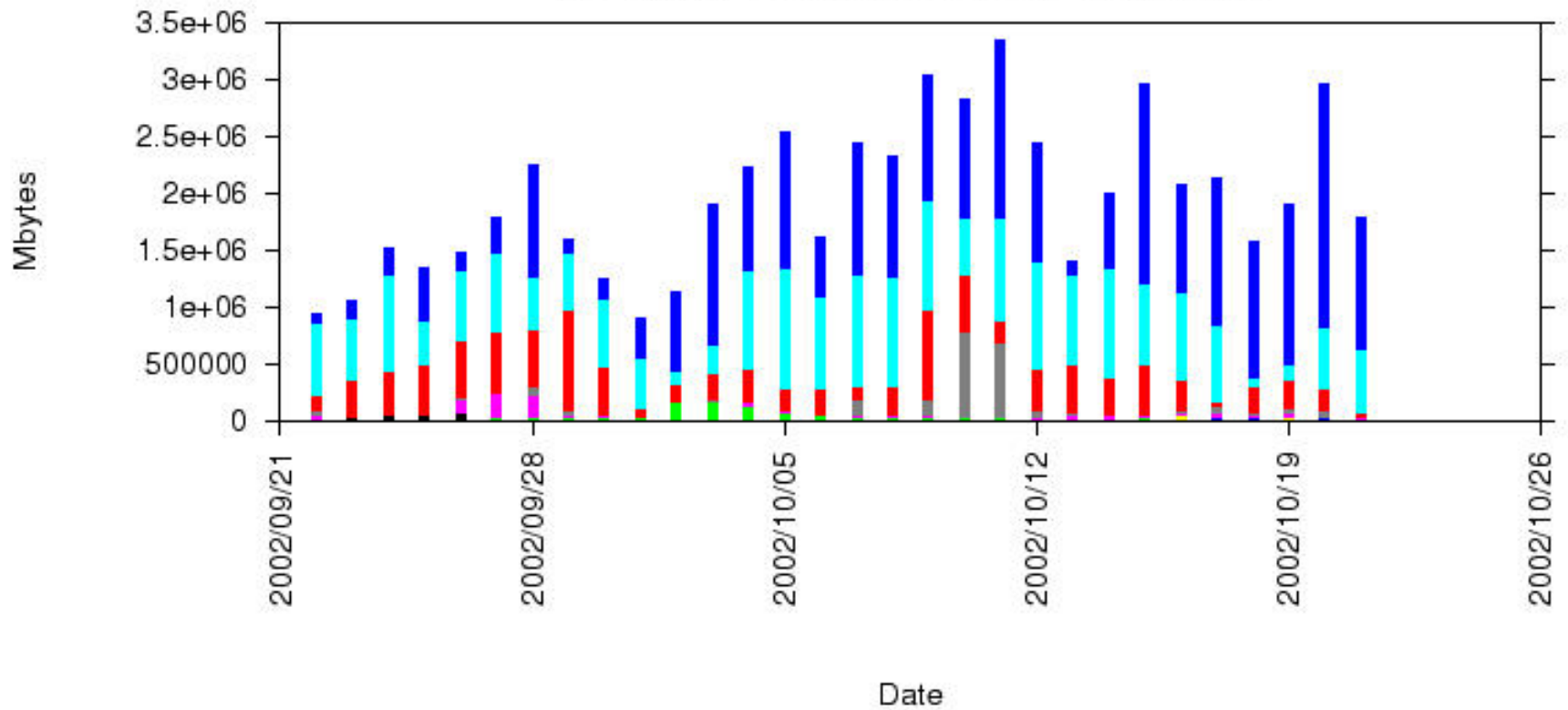Total Bytes Read Per Day (Plotted: Mon Oct 21 10:05:00 2002)

"Farm" Middleware

- Hopefully covered in Igor's talk.
- Broad Topic, elements
  - FBSNG (Farms Batch System)
  - Farms Tools (incl FCP for traffic shaping)
  - Dfarm (storage muxed on a farm)
- Detailed, close interface with experimenters.
  - Embody what is learned in software
  - Close collaboration with experimenters.
- Responsibility for deployment.

Experiment Data Handling

- SAM is the dominate middleware for CDF and D0.

- CMS looms large and has its own requirements.

- Many other experiments too,
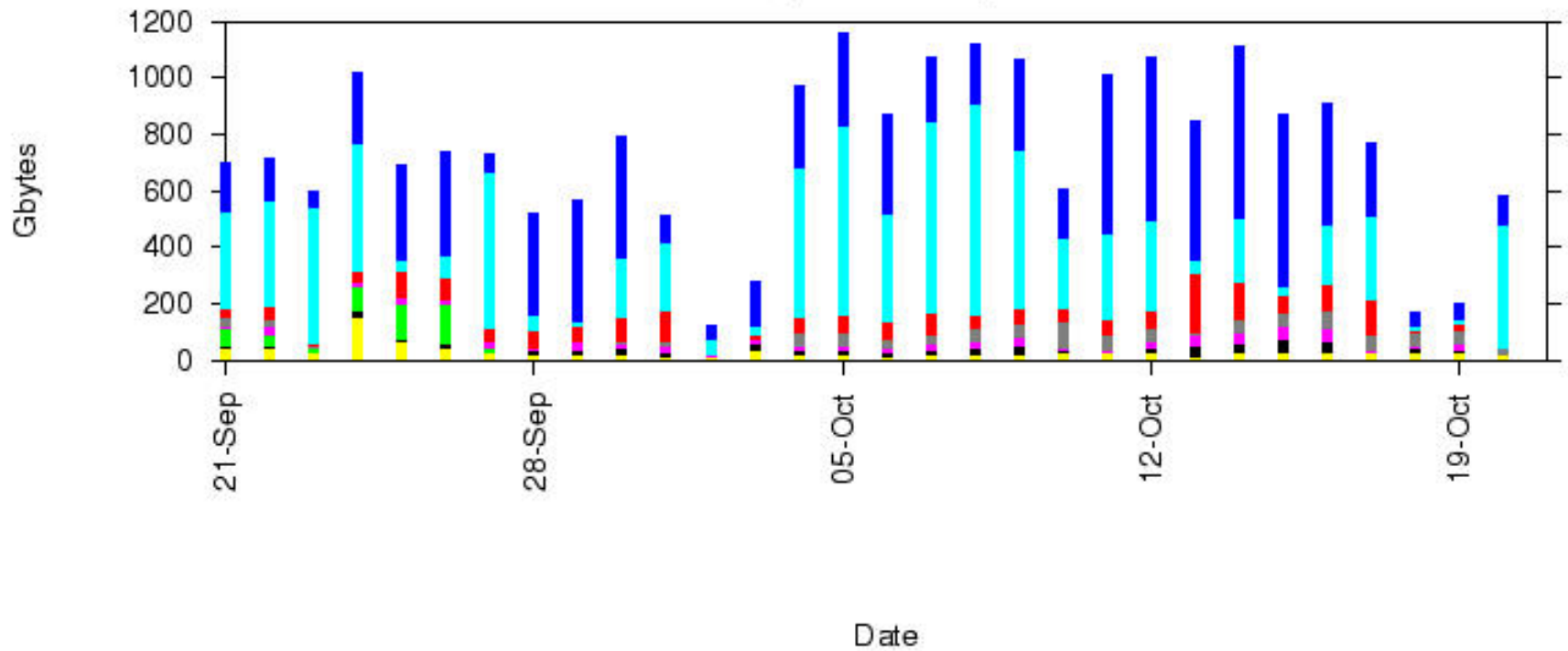 – Drives the need for facilities middleware.

Inter-Station - Total Mbytes Outbound - Last 30 Days

Incoming Gbytes per Data Tier
21-Sep-2002 - 21-Oct-2002
(D0 Production)

Data Tier

reconstructed ▮ (blue) simulated ▮ (gray) virtual-thumbnail ▮ (black)
raw ▮ (cyan) thumbnail ▮ (magenta) other ▮ (yellow)
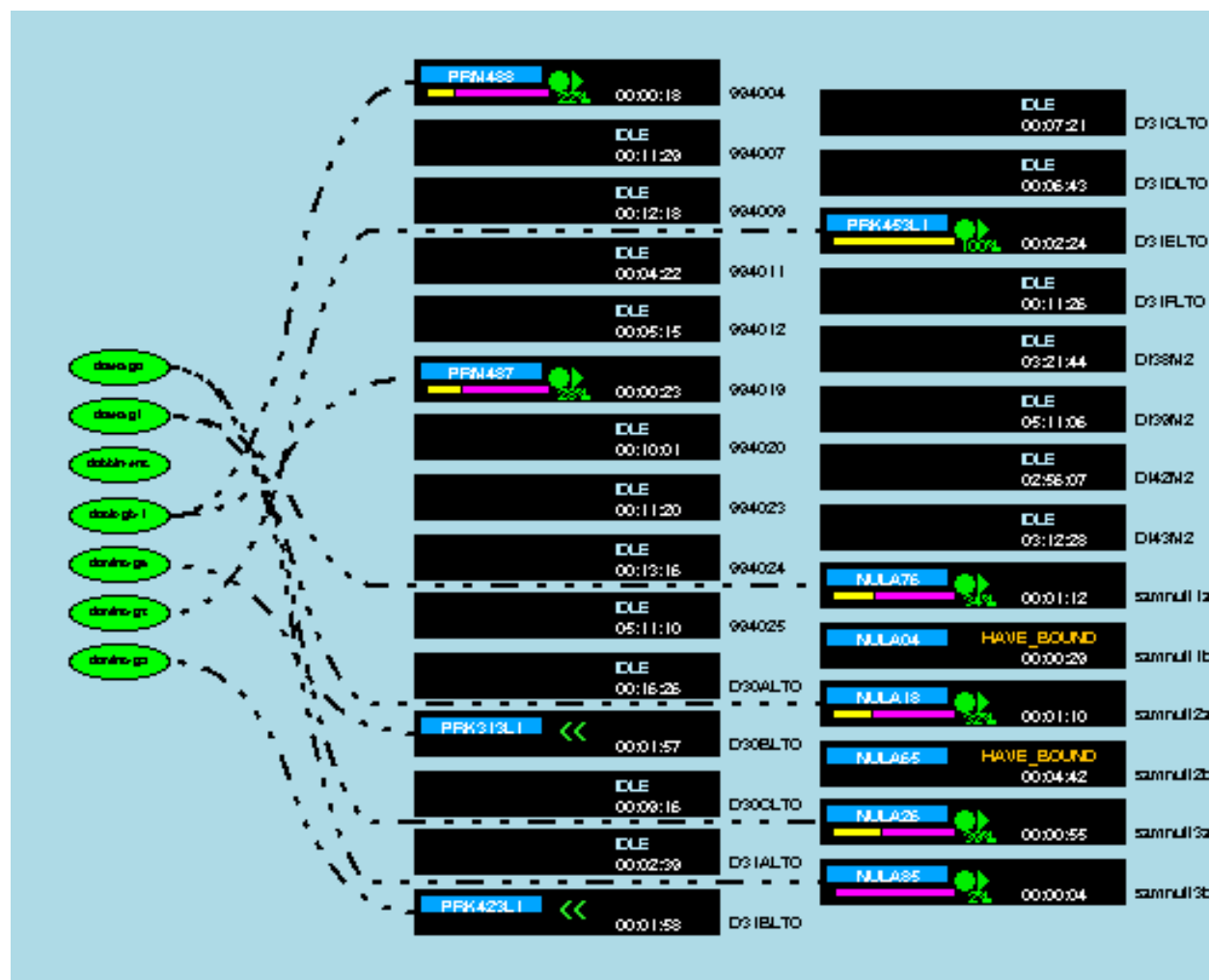root-tuple ▮ (red) filtered-reco ▮ (green)

Administration and operations support.

- Running a system, delivering a service is very important.

    OS and system availability.

    Facility middleware availability.

    Data Handling middleware availability.

Summary

FNAL facility provides comprehensice suite of tools and services to the FNAL scientific community

Community is varied.

Substantial Intellectual value, and assistance to the community.

Successful deployment of Run II experments

With key intellectual input.