Enstore

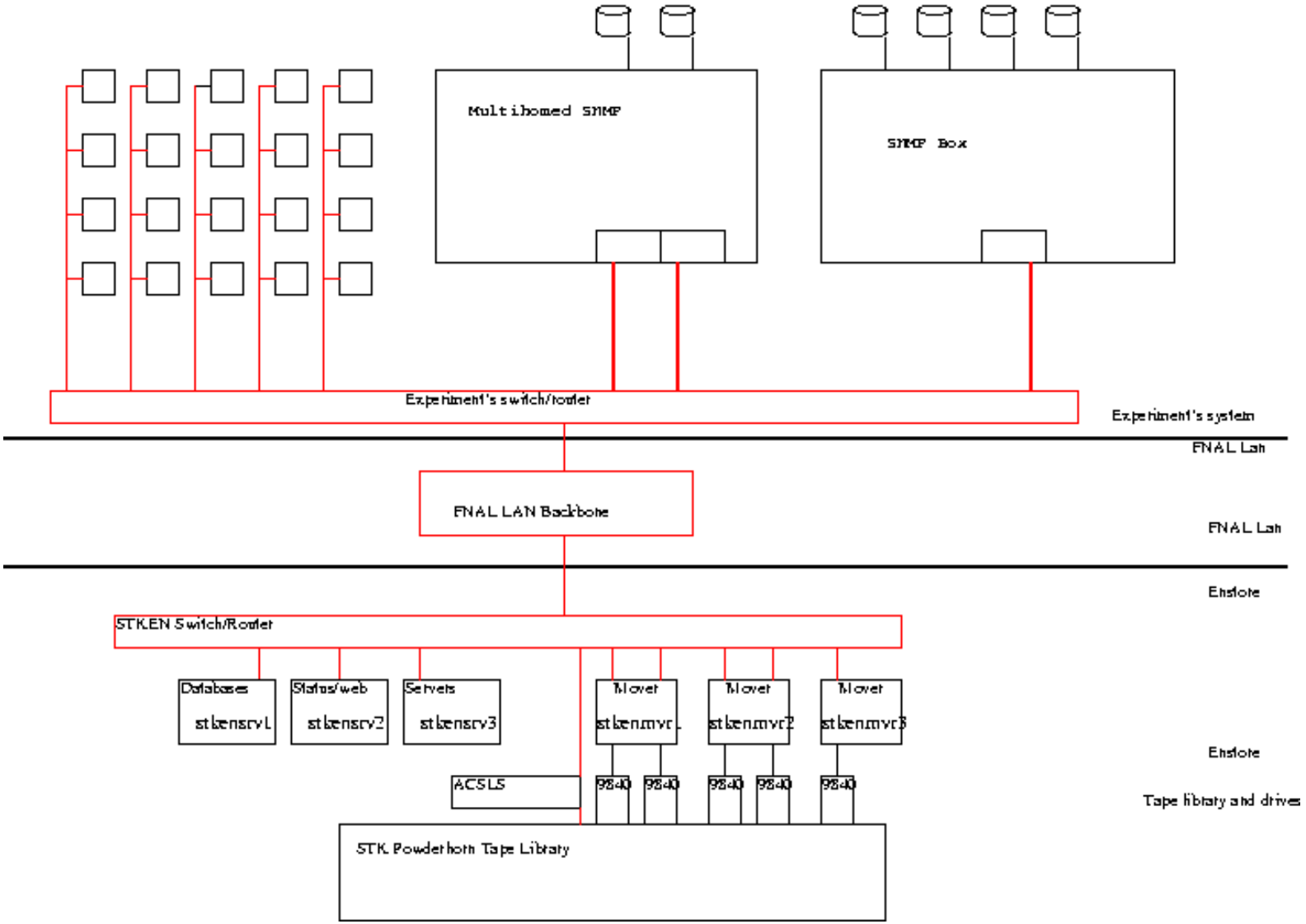
Service Envisioned

- Primary data store for experiment's large data sets, accessible via LAN
- High fault tolerance - ensemble reliability of a large tape drive plant, availability sufficient for DAQ
- Allow for automated tape libraries and manual tapes
- Catalog files in a distributed virtual filesystem (PNFS)
- Organize files efficiently on tape and expose information about tape organization to users
- CD will operate all the tape equipment
- Easy administration and monitoring

Hardware for Early Use

- 1 each - STK 9310 “Powderhorn” silo
- 5 each - STK 9840 “Eagle” tape drives
  - 11.5 MB/second
  - used at BaBar, CERN, RHIC
- 1500 - STK 9840 tape cartridges
  - 20 GB/ cartridge
- Linux Server and Mover computers
- FNAL standard network

Typical Overall System Incorporating Enstore



Service for First Users

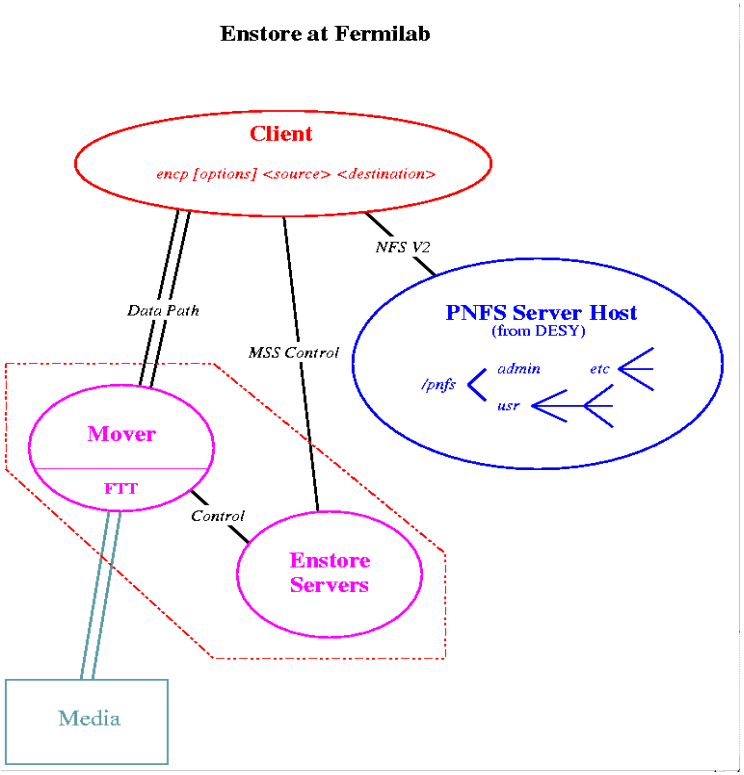
- Software in production (4 TB) for D0 Run II AML/2 tape library: 8MM, DLT drives
- STK system:
  - Only working days, working hours
  - Small data volumes ~1 TB for trial use
  - May have to upgrade LAN, network interfaces
  - Must be willing to point out bugs and problems
  - New hardware => small chance of data loss

Ease of Use

- Experiment can access tape as easily as a native file system
- PNFS mounted like NFS, viewable with UNIX commands (e.g. “ls”)
- Data is accessed with “encp” (Enstore Copy)
- Command syntax is similar to the UNIX “cp” command
  - `encp /scratch/qcd* /pnfs/theory/myproject`
- Delegate volume management to experiments
- Web-based monitoring

Basic Structure

- PNFS to name tape files using UNIX-like paths, served with NFS 2 transport
- Servers to schedule, configure, manage
- Movers to transport data and assure data integrity



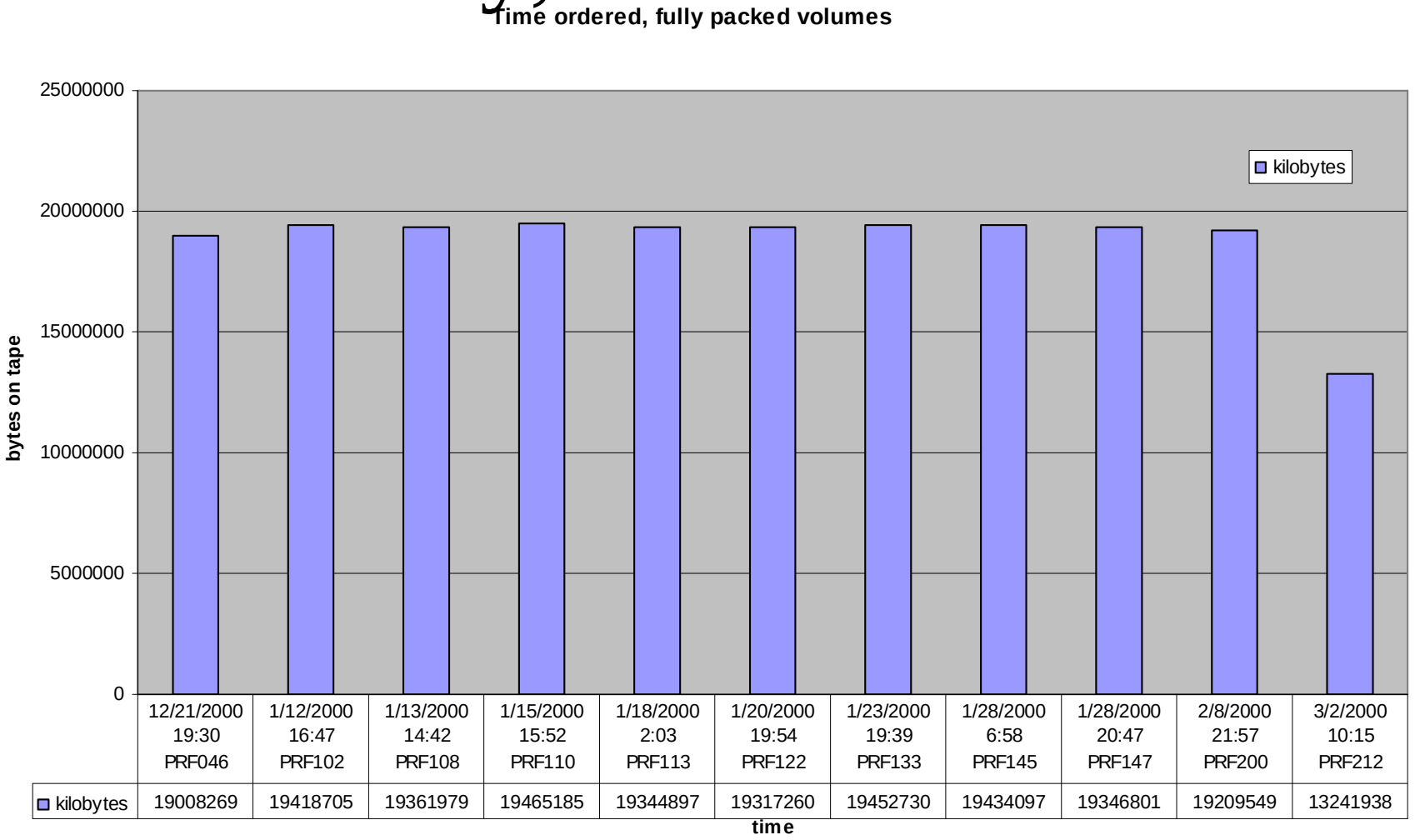
Volume Principles

- Cluster related files on the same tapes
  - Enstore provides grouping primitives
- Allow for more tapes than robot can hold
  - Enstore provides quotas in tapes and robot “slots”
- Allow users to generate tapes (and metadata) outside our system
  - Enstore “volume import” provides tools to do this
- Allow tapes to leave our system and be readable with simple tools
  - Enstore can make tapes dumpable with cpio

Grouping on Tapes

- Grouping by category
  - “File families”: only files of the same family are on the same tape
  - A family is just an ASCII name
  - Names are administered by the experiment
- Grouping by time
  - Enstore closes volume for write when the next file does not fit
- Constrained parallelism
  - “Width” associated with a file family limits the number of volumes open for writing, concentrates files on fewer volumes
  - Higher width allows bandwidth into a file family to exceed the bandwidth of a tape drive

File family, width=1 over time



	12/21/2000 19:30 PRF046	1/12/2000 16:47 PRF102	1/13/2000 14:42 PRF108	1/15/2000 15:52 PRF110	1/18/2000 2:03 PRF113	1/20/2000 19:54 PRF122	1/23/2000 19:39 PRF133	1/28/2000 6:58 PRF145	1/28/2000 20:47 PRF147	2/8/2000 21:57 PRF200	3/2/2000 10:15 PRF212
kilobytes	19008269	19418705	19361979	19465185	19344897	19317260	19452730	19434097	19346801	19209549	13241938

Tape Details

- Files do not stripe or span volumes
- Implementation details:
  - Tapes have ANSI VOL1 headers
  - Tapes are file structured as cpio archives, one file to an archive, one filemark per archive
- You can remove tapes from Enstore and just read them with GNU cpio (gives a 4GB limit right now)
- ANSI tapes planned, promised for D0

Library Manager

- Manages a homogeneous collection of tapes
- Accepts, queues and prioritizes encp client requests
- Dispatches work to Movers
- Will reorder work queues to minimize tape mounts and tape motion
- Identifies failed tape volumes

PNFS (DESY)

- Provides a directory tree to name files as you wish
- “Tags” provide per-directory defaults for creating files, which the experiment can administer (e.g. file family)
- “Layers” hold file metadata (e.g. volume label)
- Provides “volume map” directories where files are organized by physical location:
  - /pnfs/*mountpoint*/volmap/*family*/*volume*/
  - Filenames of form *partition_block_filemark* (easily sortable)

PNFS:UNIX features

- Implemented using PNFS from DESY. NFS v2 “Transport”
- “Almost all” UNIX Utilities work, ls, find
- Standard utility reads/writes fail by design
- Many files in a directory is a poor choice “by design”

- $ pwd
- /pnfs/sam/mammoth/mcc99_2/in2p3

- [$ du -sk
- 267171544 .

- $ ls -al sim.pmc02_in2p3.pythia.qcd_pt20.0_skip5800_mb1.1av_200evts.299_1138
- -rw-r--r-- 1 sam root 250184748 Nov 30 17:25 sim.pmc02_in2p3.pythia.qcd_pt20.0_skip5800_mb1.1av_200evts.299_1138

- rm sim.pmc02_in2p3. pythia.qcd_pt20.0_skip5800_mb1.1av_200evts.299_1138
- rm: sim.pmc02_in2p3 .pythia.qcd_pt20.0_skip5800_mb1.1av_200evts.299_1138: Permission denied

- $ cat sim.pmc02_in2p3.pythia.qcd_pt20.0skip5800_mb1.1av_200evts.299_1138 > /dev/null
- cat: sim.pmc02_in2p3.pythia.qcd_pt20.0_skip5800_mb1.1av_200evts.299_1138: Input/output error

PNFS:defaults for new files

- Metadata “tags” are associated with directories
- Accessed by the “enstore pnfs” command.
- Inherited on “mkdir”.
- Initial tag on initial directory given by ISD dept
- Administered by the experiment

- [petravic@d0enmvr17a in2p3]$ enstore pnfs --tags .
- .(tag)(library) = sammam
- .(tag)(file_family) = in2p3
- .(tag)(file_family_width) = 1
- .(tag)(file_family_wrapper) = cpio_odc

PNFS: File Metadata

- Describes an existing file
- Accessed by the “enstore pnfs” command
- Set by encp when the file is created

- $ enstore pnfs --info sim.pmc02_in2p3.pythia.qcd_pt20.0_10000evts_skip5800_mb1.1av_200evts.299_1138
- bfid="94400431100000L";
- volume="PRF020";
- location_cookie="0000_000005442_0000004";
- size="250184748L";
- file_family="in2p3";
- map_file="/pnfs/sam/mammoth/volmap/in2p3/PRF020/0000_000005442_0000004";

Removing Files

- Files may be removed using “rm”.
- User can scratch tape when all files on it are rm’ed. [enstore volume --delete]
- User can use a recovery utility to restore files up until the time the volume is scratched. [enstore file --restore]
- Files are recovered to pathname they were created with.

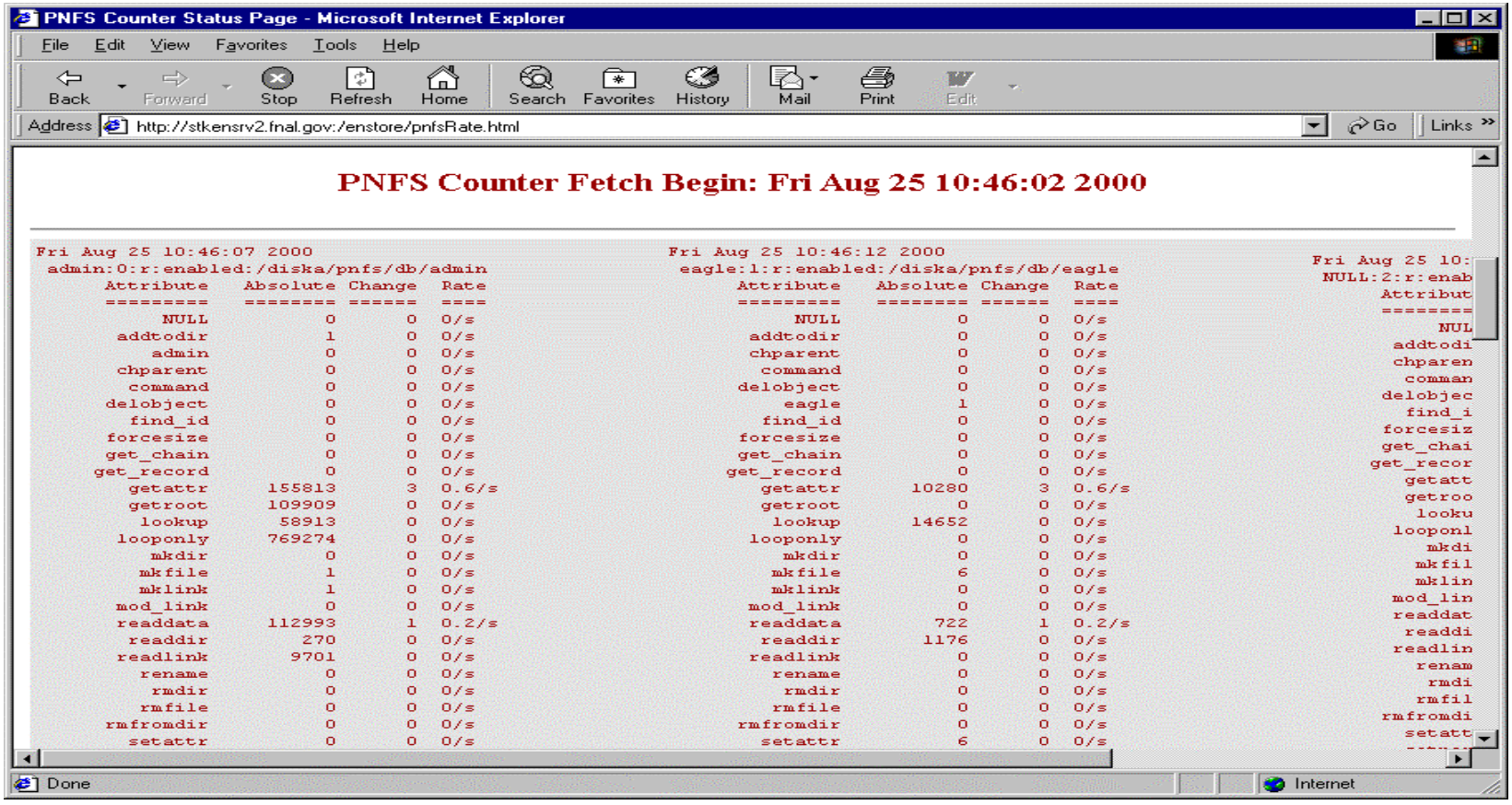
PNFS implementation (DESY)

- Mountd
- “nfsd”
- shared memory bkplane
- dbservers (gdbm)

System Admin's View of PNFS

- NFSv2
  - is the transport for the name space
  - gives authentication and authorization for access to metadata UID/GID + host based control on export
- Model FSTAB entry:
  - stkensrv1:/theory /pnfs/theory nfs user,intr,bg,hard,rw,noac 0 0
- user File data is NEVER transmitted via NFS
- NFS statistics at: http://stkensrv2.fnal.gov:/enstore/pnfsRate.html
- showmount is broken:
- hppc> showmount -e stkensrv1
- showmount: Can't decode result

http://stkensrv2.fnal.gov:/enstore/pnfsRate.html



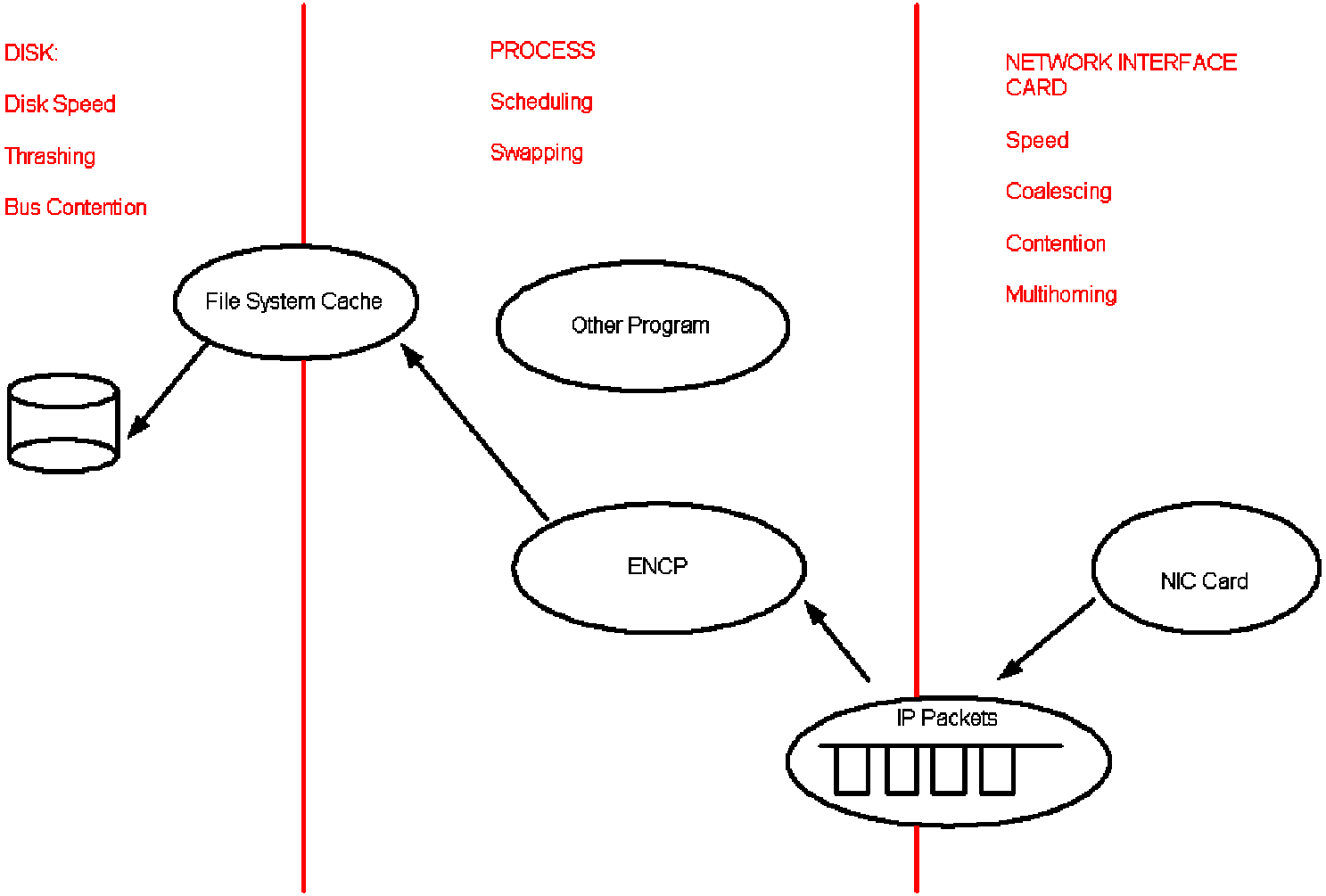
Sharing the Central Enstore System

- We make mount point(s) for your experiment
  - Host-based authentication on the server side for mounts.
  - Your meta data is in its own database files.
  - Under the mount point, UNIX file permission apply.
  - Make your uids/gid's uniform! (FNAL uniform UID/GIDS).
  - file permissions apply to the tag files as well.
- "Fair Share" envisioned, for tape drive resources.
  - Control over experiment resources by the experiment
- Priorities implemented for Data Acquisition.
  - Quick use of resources for the most urgent need

System Integration

- Hardware/system:
  - Consideration of upstream network.
  - Consideration of your NIC cards.
  - Good scheduling of the staging program.
  - Good throughput to your file systems.
- Software configuration
  - Software built for FUE platforms
    - Linux, IRIX, SunOS, OSF1

Elements of Good Throughput



Checking….

- Enstore monitor:

- <fsgi02> setup -q stken encp
- <fsgi02> enstore monitor
- Trying ('stkensrv1.fnal.gov', [], ['131.225.13.1']) Success. Network rate measured at 1.092 MB/S
- <fsgi03> enstore monitor
- Trying ('stkensrv1.fnal.gov', [], ['131.225.13.1']) Success. Network rate measured at 11.3 MB/S

- Recent rates available at: http://stkensrv2.fnal.gov:/enstore/active_monitor.html

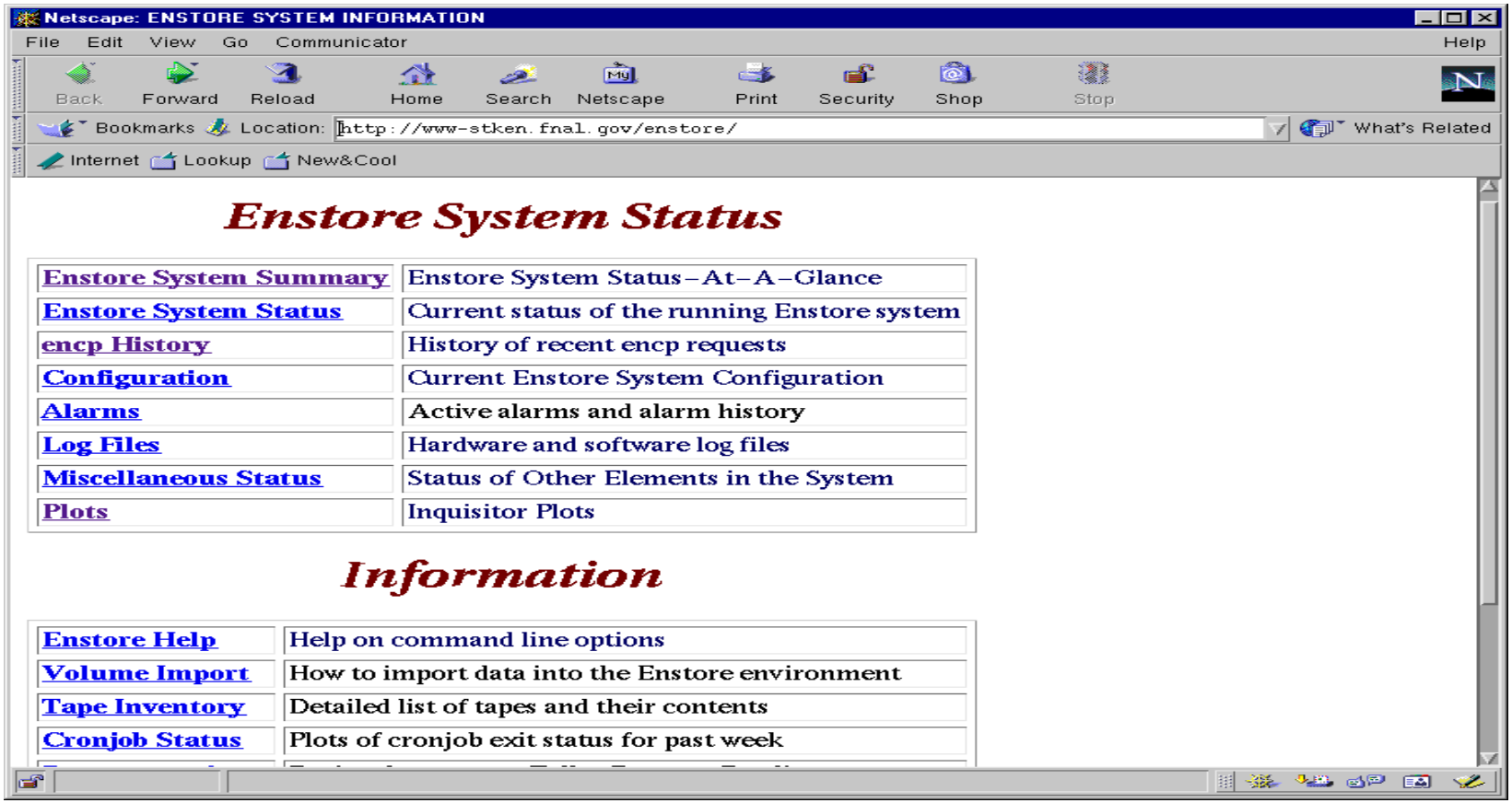
- Read a file to /dev/null

- <fsgi03> encp --verbose=1 /pnfs/sdss/opdb/2000-Jun-16/d1/data1241_6_1.photo.DB /dev/null
- Submitting read requests cumt= 0.399256944656
- Q'd: /dev/null 96116582800000L bytes: 115146752 on VO0116 0000_000000000_0000008 dt: 0.085179 cumt=0.485039
- Receiving data for file /dev/null cumt= 0.611217021942
- read_from_hsm /pnfs/sdss/opdb/2000-Jun-16/d1/data1241_6_1.photo.DB -> /dev/null: 115146752 bytes copied from VO0116 at 1.54 MB/S (8.68 MB/S) mover=eagle12.mover drive_id=9840 drive_sn=None drive_vendor=STK cumt= 71.847064 {'media_changer' : '….

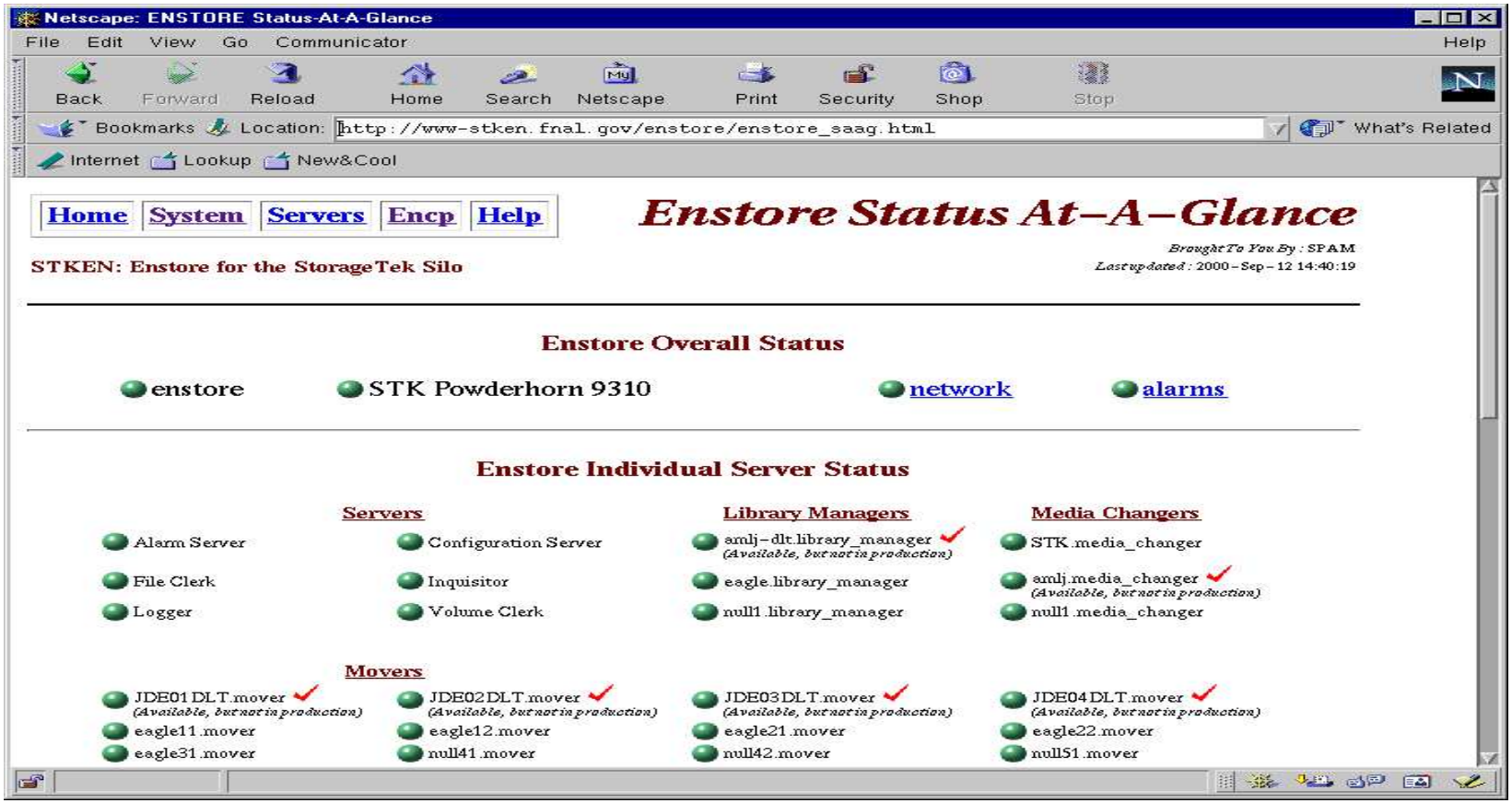
http://stkensrv2.fnal.gov/enstore/

- Source of interesting monitoring info
- Most updates are batched.
- Can see
  - recent transfers
  - the system is up or down
  - what transfers are queued
  - more

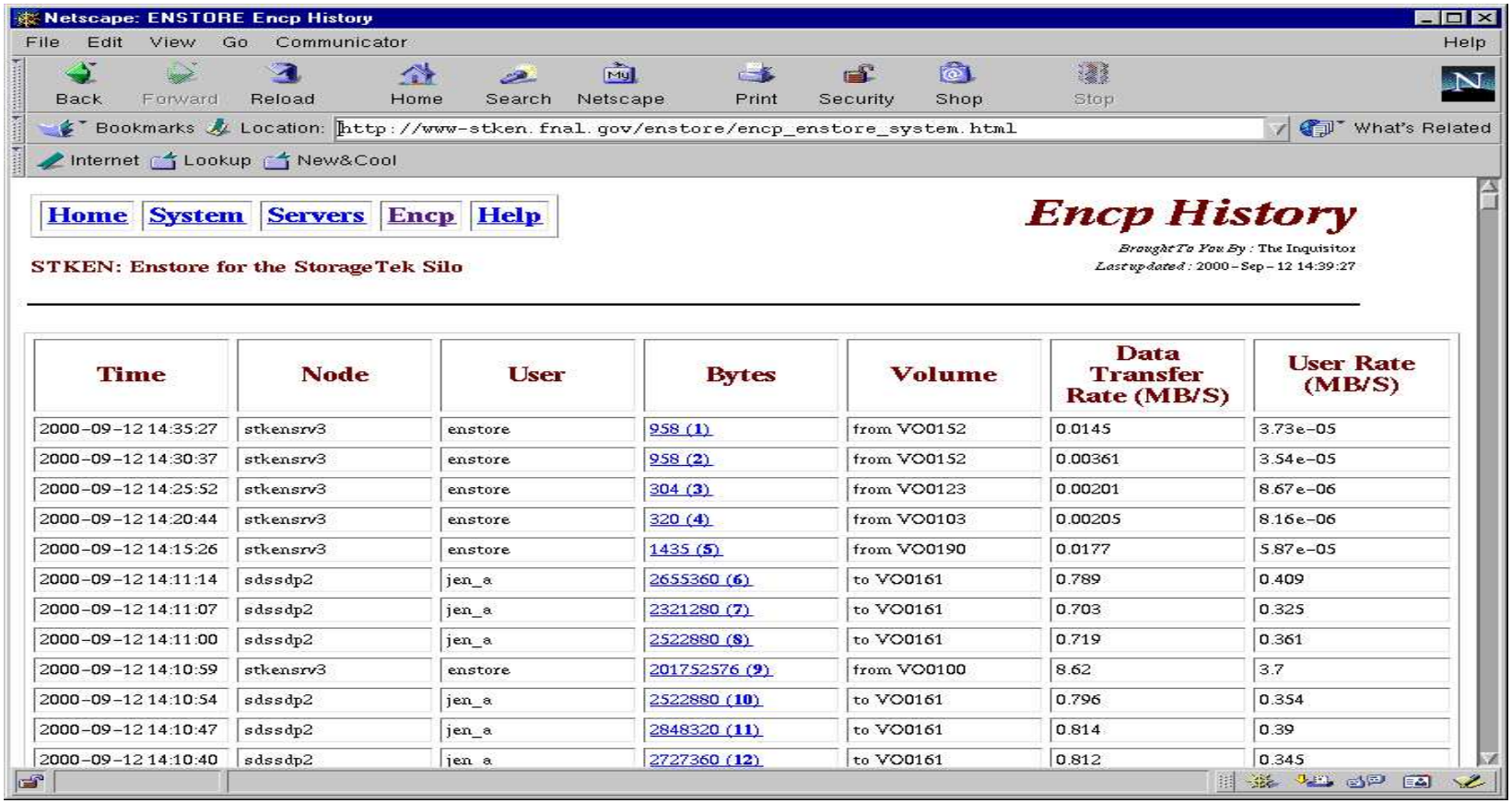
http://stkensrv2/enstore/



System Status: Green == Good

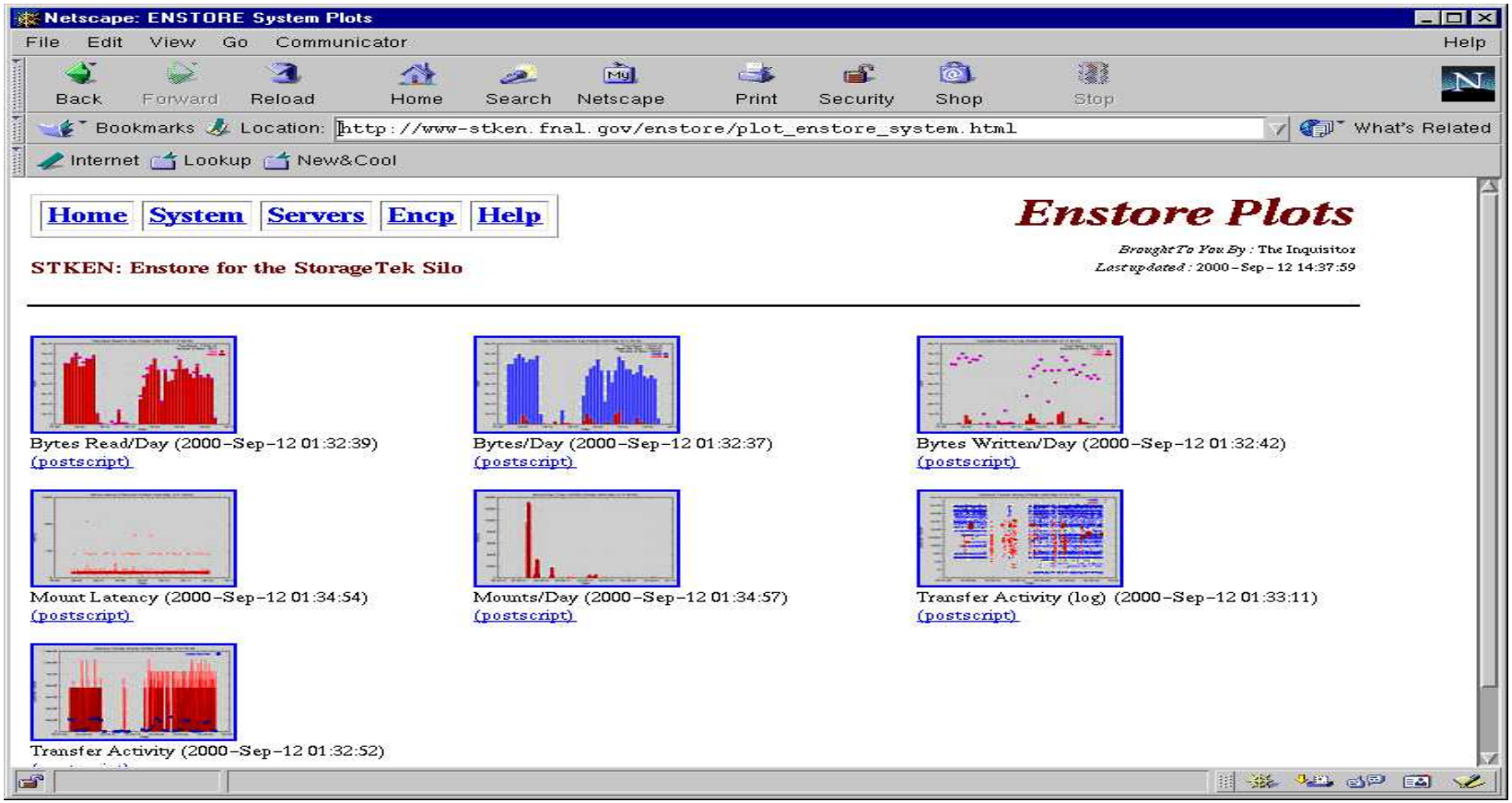


History of Recent Transfers



Time	Node	User	Bytes	Volume	Data Transfer Rate (MB/S)	User Rate (MB/S)
2000-09-12 14:35:27	stkensrv3	enstore	958 (1)	from VO0152	0.0145	3.73e-05
2000-09-12 14:30:37	stkensrv3	enstore	958 (2)	from VO0152	0.00361	3.54e-05
2000-09-12 14:25:52	stkensrv3	enstore	304 (3)	from VO0123	0.00201	8.67e-06
2000-09-12 14:20:44	stkensrv3	enstore	320 (4)	from VO0103	0.00205	8.16e-06
2000-09-12 14:15:26	stkensrv3	enstore	1435 (5)	from VO0190	0.0177	5.87e-05
2000-09-12 14:11:14	sdssdp2	jen_a	2655360 (6)	to VO0161	0.789	0.409
2000-09-12 14:11:07	sdssdp2	jen_a	2321280 (7)	to VO0161	0.703	0.325
2000-09-12 14:11:00	sdssdp2	jen_a	2522880 (8)	to VO0161	0.719	0.361
2000-09-12 14:10:59	stkensrv3	enstore	201752576 (9)	from VO0100	8.62	3.7
2000-09-12 14:10:54	sdssdp2	jen_a	2522880 (10)	to VO0161	0.796	0.354
2000-09-12 14:10:47	sdssdp2	jen_a	2848320 (11)	to VO0161	0.814	0.39
2000-09-12 14:10:40	sdssdp2	jen_a	2727360 (12)	to VO0161	0.812	0.345

Summary Plots



Checklist to use Enstore

- Be authorized by the computing division.
- Identify performant disks and computers.
  - Use “bonnie” and “streams”
- Provide suitable network connectivity.
  - Use “enstore monitor” to measure.
- Plan use of namespace, file families.
- Regularize UIDs and GIDs if required.
- Mount the namespace.
- Use encp to access your files.

Support while using Enstore

Trouble reports to helpdesk

- Helpdesk@fnal.gov
- x2345

Information from us:

- Status pages on the web
- Stk-users@fnal.gov mailing list
  - We require at least one person from your group

- Very draft users guide.