

Fermilab Mass Storage

Enstore,
dCache and
SRM

Michael Zalokar
Fermilab

What are they?

- Enstore
 - In-house, manages files, tape volumes, tape libraries
 - End-user direct interface to files on tape
- dCache
 - Joint DESY and Fermilab, disk caching front-end
 - End user interface to read cached files, write files to enstore indirectly via dCache
- SRM
 - Provides a consistent interface to underlying storage systems.

Software

- 3 production systems of Enstore

- RunII:

- D0
 - CDF

- Everyone else:

- MINOS, MiniBooNE, SDSS, CMS, et. al.

- 3 production systems of dCache and SRM

- CDF

- CMS

- Everyone else



Requirements

- Scalability
- Performance
- Availability
- Data Integrity

PNFS

- Provides a hierarchical namespace for users' files in Enstore.
- Manages file metadata.
- Looks like an NFS mounted file system from user nodes.
- Stands for “Perfectly Normal File System.”
- Written at DESY.



Enstore Design


- Divided into a number of server processes
 - Scalability is achieved by spreading these servers across multiple nodes.
 - If a node goes down, we can modify the configuration to run that nodes servers on a different node. This increases availability while the broken node is fixed.
- Enstore User Interface: `encp`
 - Similar to standard UNIX `cp(1)` command
 - `encp /data/myfile1 /pnfs/myexperiment/myfile1`

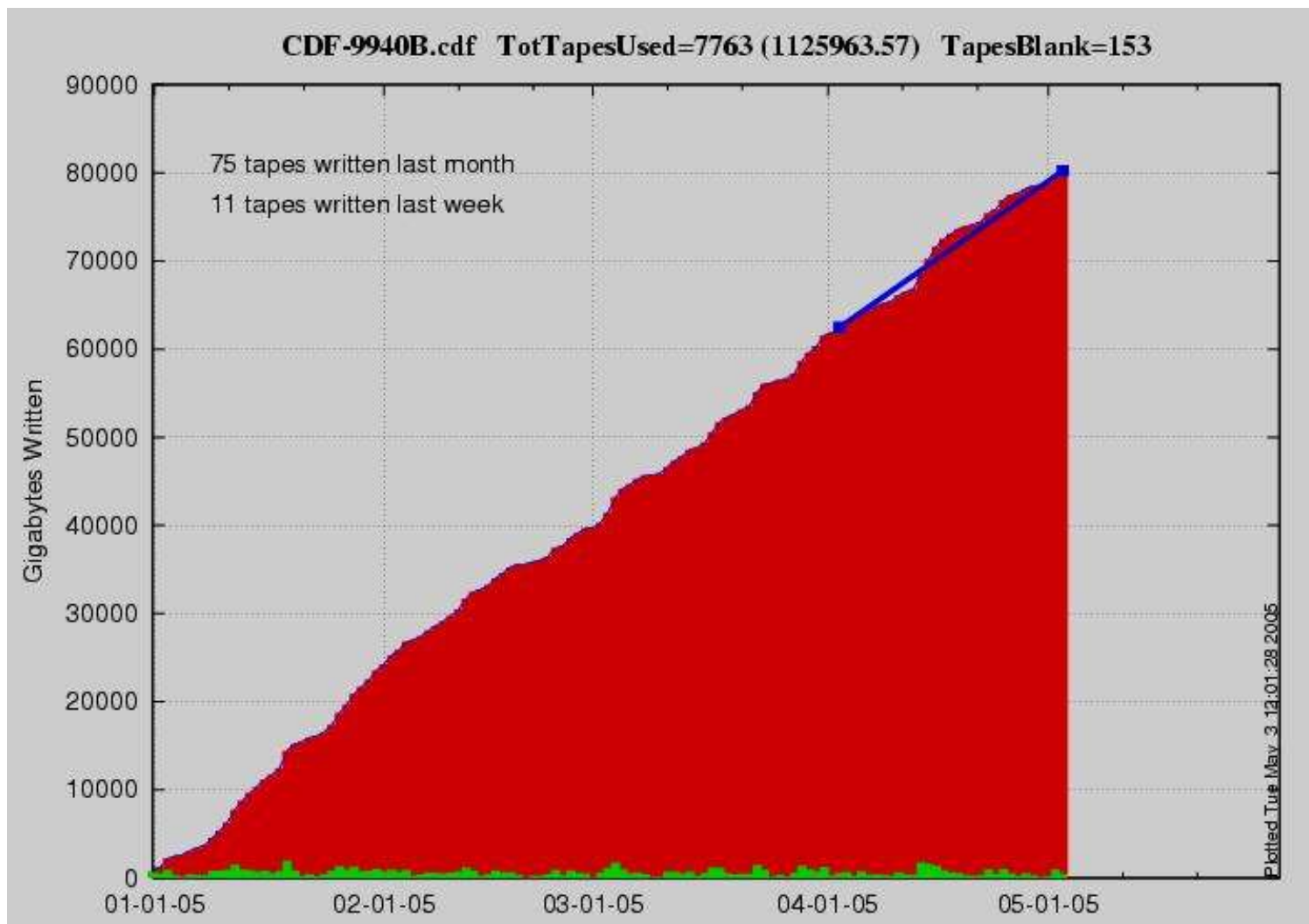
Hardware

- Robots
 - 6 StorageTek Powderhorn Silos
 - 1 ADIC AML/2
- Tape Drives:
 - LTO: 9
 - LTO2: 14
 - 9940: 20
 - 9940B: 52
 - DLT (4000 & 8000): 8
- 127 commodity Linux PCs



Enstore Monitoring

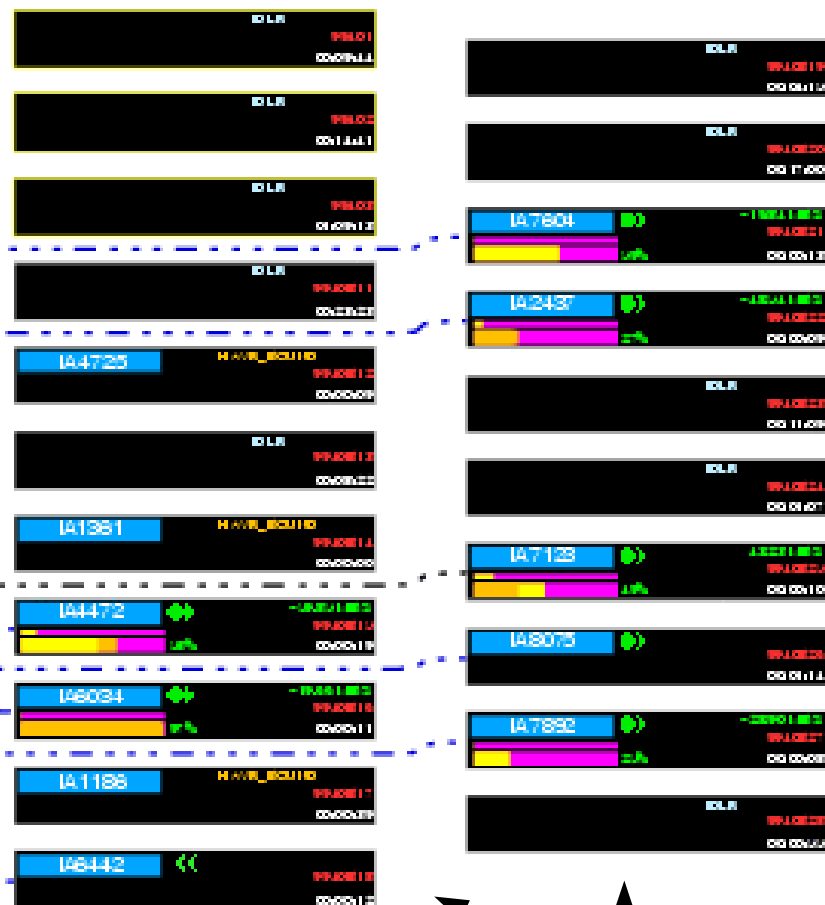
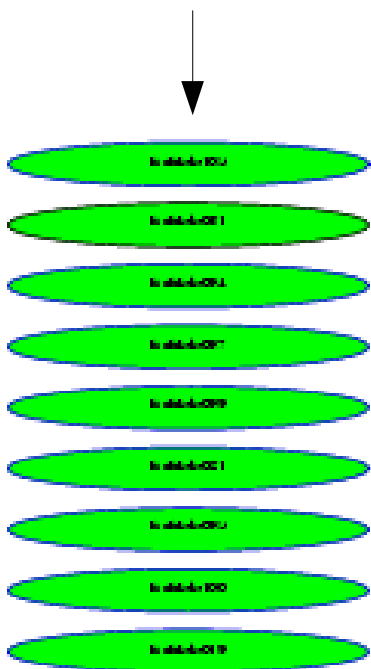
- Web pages for current server statuses 
- Cron Jobs
- Plots for resource usage
 - Number of tapes written
 - Number of tape drives in use
 - Number of mounts
 - And much more...
- entv (ENstore TV)



- X-Axis is time since January 1st 2005 until present
- Y-Axis is number of gigabytes written
- Includes summary of tapes written in last month and week

ENTV

- Client nodes



- Real time animation

- Tape & drive information

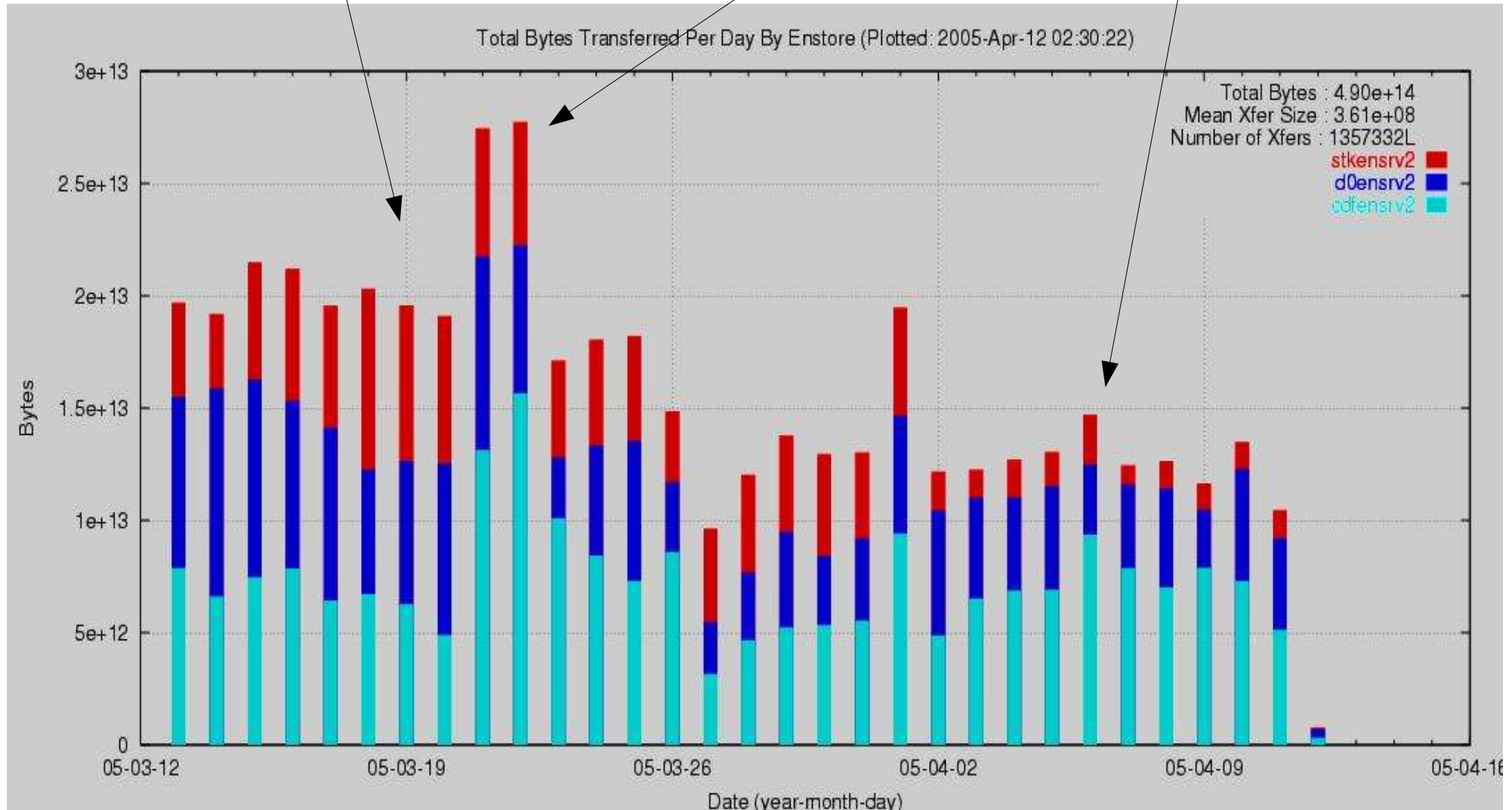
- Current Tape
- Instantaneous Rate

By the Numbers

- User Data on Tape:
 - 2.6 Petabytes
- Number of files on tape:
 - 10.8 million
- Number of volumes:
 - ~25,000
- One Day Transfer Record
 - 27 Terabytes

Performance: 27TB

- Two days of record transfer rate
- CMS Service Challenge in March (In red)
- Normal usage



Lessons Learned

- Just because the file transferred without error, does not guarantee that everything is fine.
 - With Fermilab's load we see bit error corruption.
- Users will push the system to its limits.
 - Record 27TB transfer days were not even noticed for three days.
- Just having a lot of logs, alarms and plots is not enough. They must also be interpretable.



dCache

dCache



dCache

- Works on top of Enstore or as standalone configuration.
- Provides a buffer between the user and tape.
- Improves performance for 'popular' files by avoiding the need of reading from tape every time a file is needed.
- Scales as nodes (and disks) are added.



User Access to Data in dCache



dCache *dCache*

- srm – storage resource manager
 - srmcp
- gridftp
 - globus_url_copy
- kerberizedftp
- weakftp
- dcap – native dCache protocol
 - dccp
- http
 - wget, web browsers



dCache

dCache Deployment

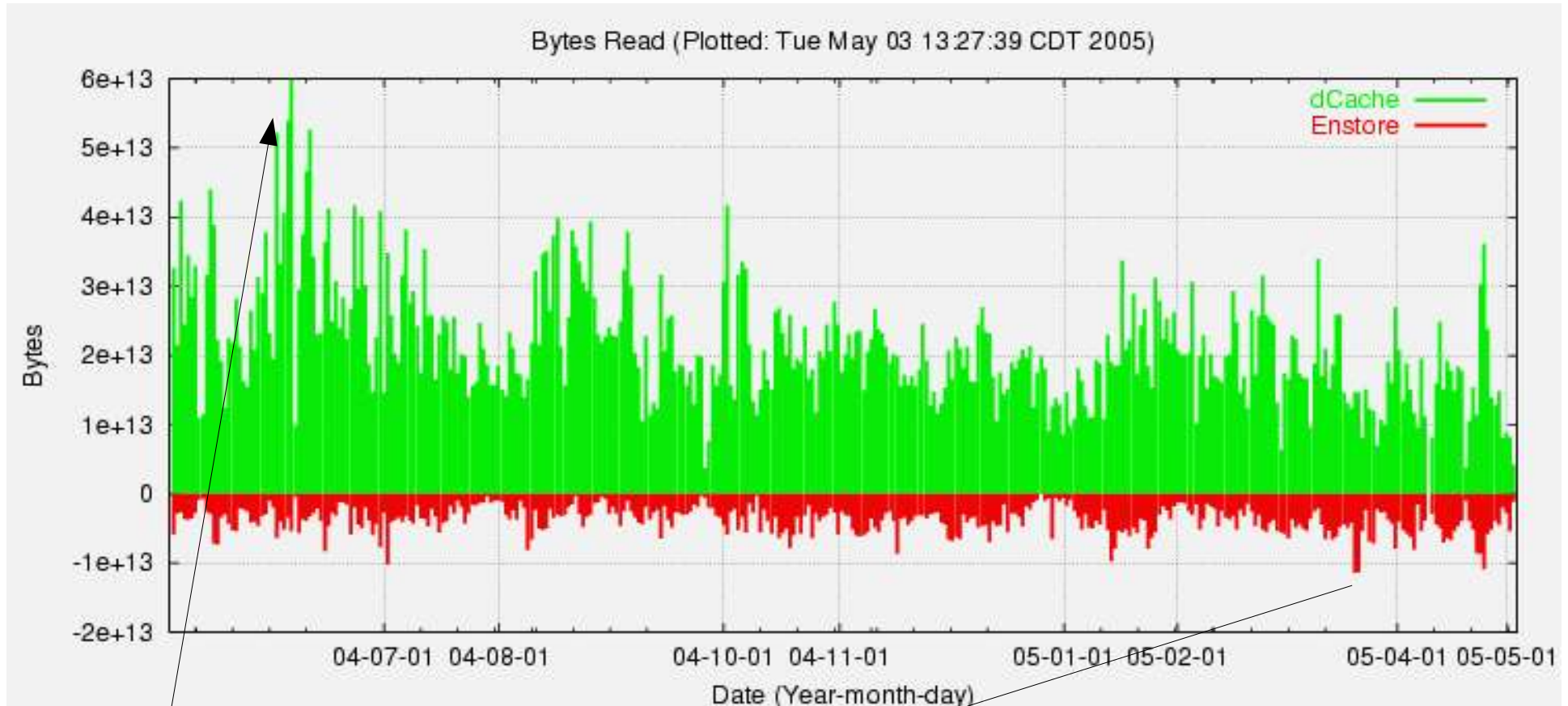


dCache

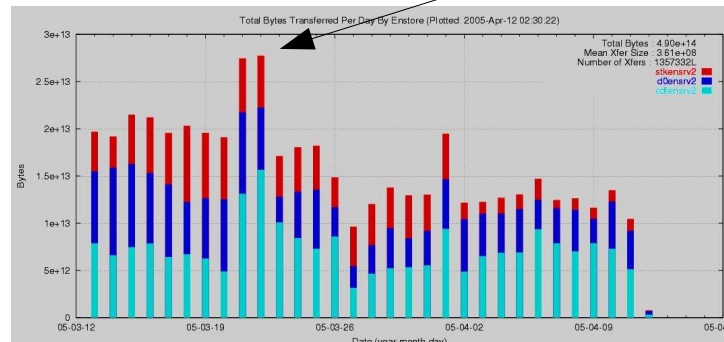
- Administrative Node
- Monitoring Node
- Door Nodes
 - Control channel communication
- Pool Nodes
 - Data channel communication
 - ~100 pool nodes with ~225 Terabytes of disk



dCache Performance



Record transfer day of 60GB.
This is for just one dCache system.





Lessons Learned

- Use the XFS filesystem on the pool disks.
- Use direct I/O when accessing the files on the local dCache disk.
- Users will push the system to its limits. Be prepared.

Storage Resource Manager

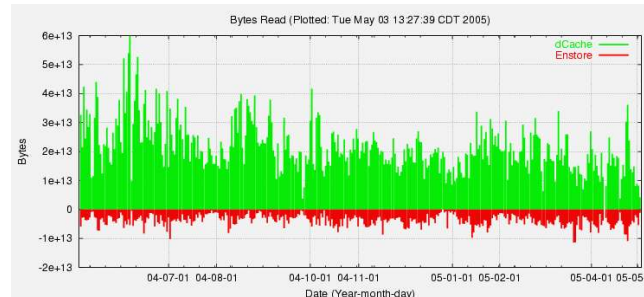
- Provides uniform interface for access to multiple storage systems via SRM protocol.
- SRM is a broker that works on top of other storage systems.
 - dCache
 - Runs as a server within the dCache.
 - UNIXTM filesystem
 - Standalone
 - Enstore
 - In development

CMS Service Challenge

- 50 MB/s sustained transfer rate
 - From CERN, through the dCache to tape in Enstore
 - On top of normal daily usage of 200 to 400 MB/s
 - Rate throttled to 50 MB/s
- 700 MB/s sustained transfer rate
 - From CERN to dCache disk

Conclusions

- Scalability
- Performance
- Availability
 - Modular design
- Data Integrity
 - Bit errors detected from scans.



Requirements are achieved.