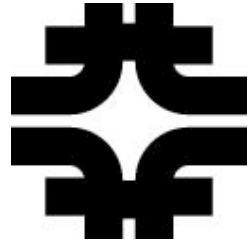


IO R&D LHC perspecitve

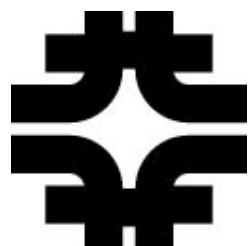
D. Petravick

Fermi National Accelerator Lab



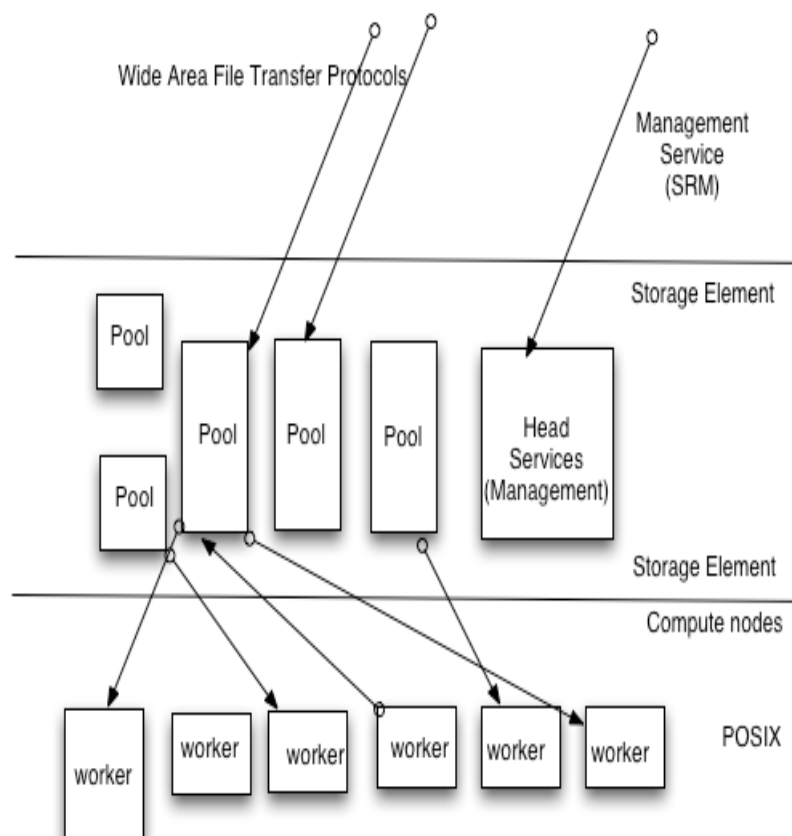
Overview

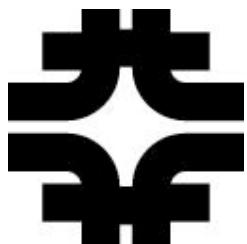
- HEP is an statistical science that studies particle collision event records < 1 MB containing records of a few KB.
- HEP DM systems deals with distilling, ordering records, partly for efficient access.
- HEP I/O
 - is multi stream capacity IO.
 - “file systems” are community written.
 - IDE, Linux, IP SAN, WAN
 - World wide interoperation requirements (LGC grid)



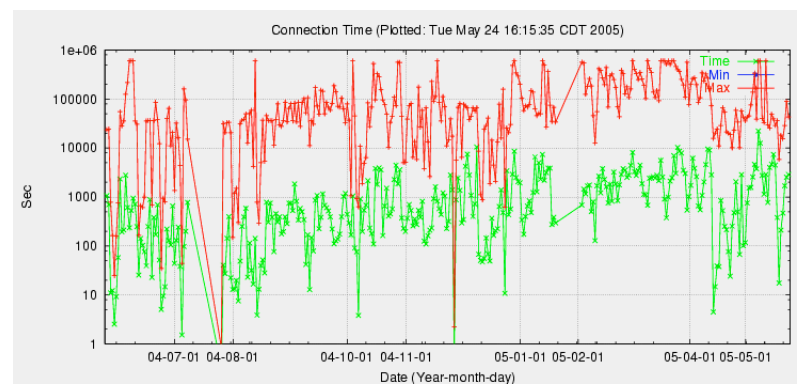
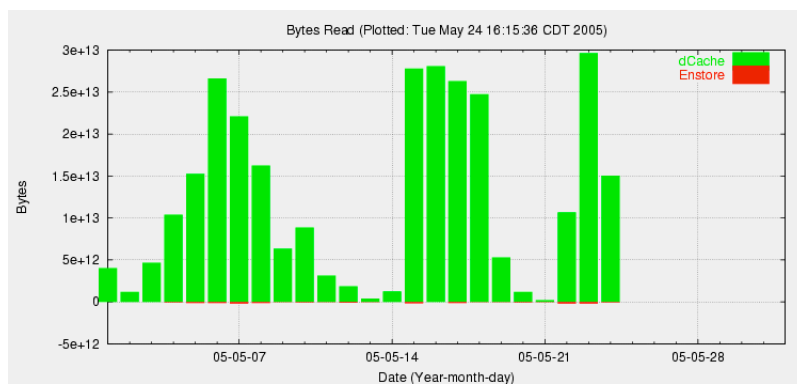
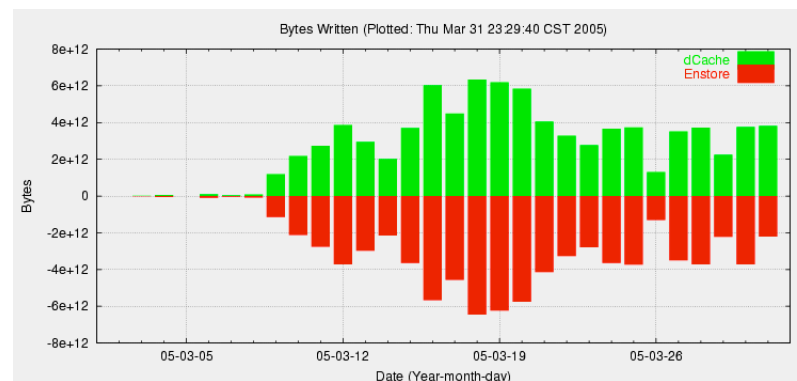
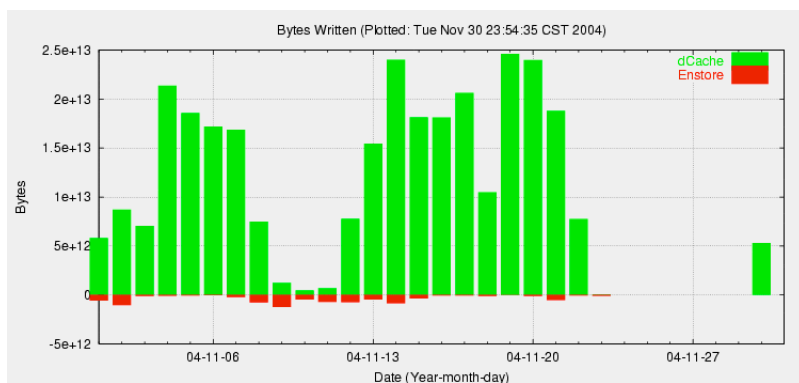
OSG Storage element

- Storage Resource Manager interface.
 - storage space management
 - data movement resource management
- File transfer protocols (GridFTP as “lingua franca”)
- POSIX-like IO on the worker nodes.



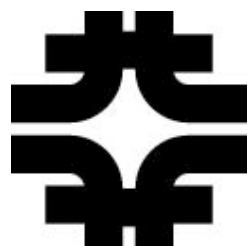


SE performance

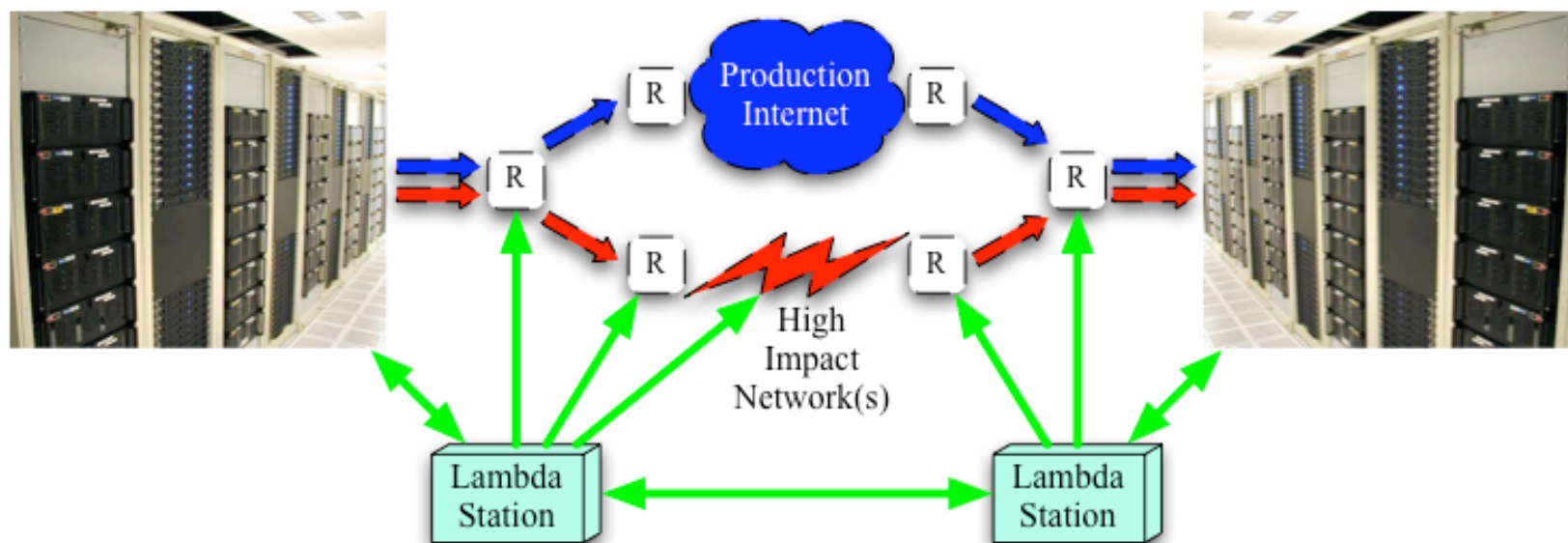


8/16/05

DALLAS HEC IO -- D. Petravick,
FNAL

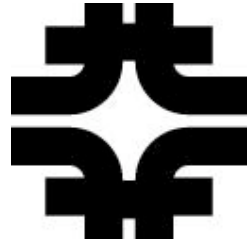


Practical Integration



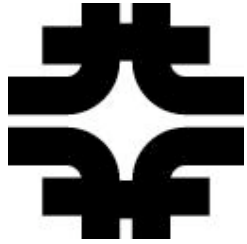
8/16/05

DALLAS HEC IO -- D. Petravick,
FNAL



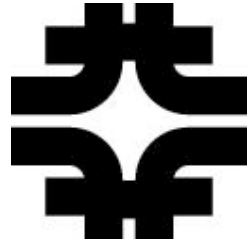
R&D topics (1)

- Integration to Wide Area.
 - Including ckt oriented “high impact” networks.
 - At large bandwidth*delay, (worldwide)
 - Retaining commodity last mile.
 - Service controlled firewalls, service informed IDS,
 - “VO perimeter”
- Internal scheduling (Data movement)
 - IP based SAN --> shaping to reduce last port congestion.
 - kernel resources for TCP @ large latency
 - (yet more) automatic replication for performance.



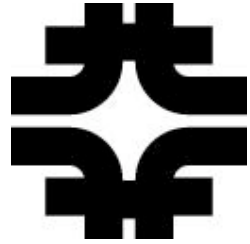
R&D Topics(2)

- Grid/Batch Systems R&D
 - Proper interface to upper level including grid batch schedulers.
 - Manage Storage Space (e.g. n-file transfer will complete)
 - Manage data movement resources.(do not always start service on request)
- Scale-breaking wide area interactive use.
 - Multi terabyte transactions in under an hour.
 - Interactive data sharing.



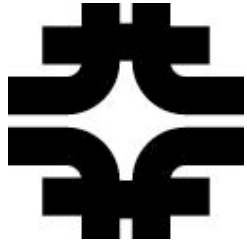
R&D (3)

- Standardize (or at least articulate deviants)
 - Subset POSIX I/O apropos for write-once files.
 - + Added interfaces for prestige, pin, groupings...
 - + Relaxed timings for file systems meta-data --
 - allow for more easliy scalable name spaces, etc.
 - + Communicate User Behavior Assertions (e.g. not too many small files)
 - + X509 + GRID permission model.



R&D topics (4)

- Reliability and fault detection.
 - HEP's lust for commodity equipment implies storage system detects, mitigates faults in disk/kernel file systems, etc.
- True SAN
 - Infiniband potential for commodity non IP SAN.



R&D Topics (5)

- Novel Storage Devices
 - Tape as the basis for a permanent archive.
 - Performance of tape induces a streaming requirement that is not naturally in the problem.
 - Tape -- love it or leave it?
 - Small (random) object IO, memory in the storage hierarchy I.E. access w/ file system metaphors)?
HEP data is Object oriented w/potentially very small gain-size
 - Performance of disk induces a streaming requirement which is not naturally in the problem.
 - Disk -- love it or leave it?