Data Management

CERN IT Department

Plans and Architectural Options for Physics Data Analysis at CERN

D. Duellmann, A. Pace

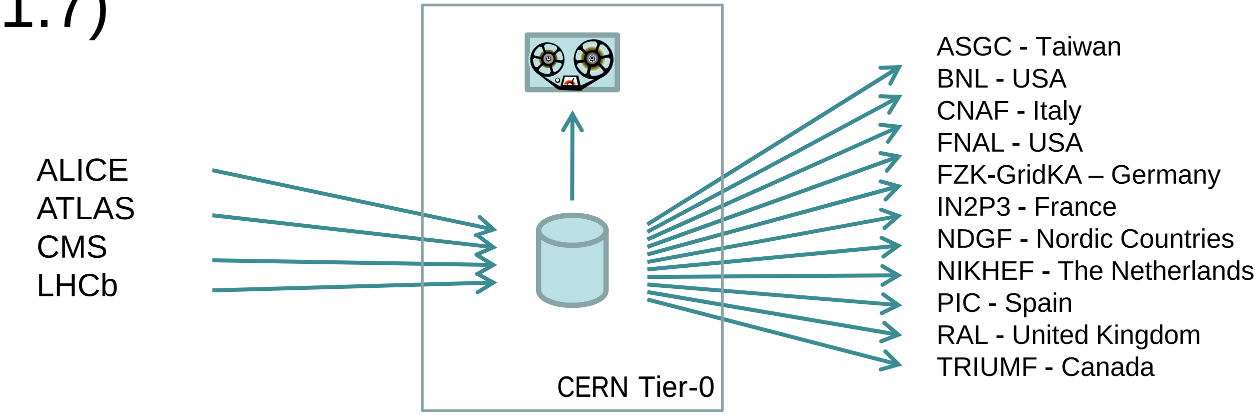
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Agenda

- Short summary of
  - Data Management Software Status for the Tier-0
  - Ongoing developments for Tier-0 activities
- Requirements for Analysis received so far
- Current areas of developments and plans

Software Status for the Tier-0

- For all Tier-0 data operations, CERN is using CASTOR for
  - Migration of LHC data to tape
  - Replication of the LHC data to Tier-1 sites
- CCRC’08 Data Challenges have validated this architecture which is now in production (version 2.1.7)



- ... but few improvements are necessary as Tier-0 data must also be accessible for analysis at CERN

Ongoing developments for Tier-0

- During 2008, several simplifications and improvements were made to CASTOR

  1. Improvements on Security, SRM and Monitoring
  2. Improvements in the tape handling area
  3. Reduced latency when accessing files already on disk and improved concurrent access

- See D.Duellmann at Hepix Fall 2008
  http://indico.twgrid.org/contributionDisplay.py?contribId=121&sessionId=87&confId=471

Security, SRM and Monitoring

CERN IT Department

- Strong authentication now implemented

Today (on v2.1.8)

  - End user access to CASTOR has been secured (nameserver, disk cache, protocols)
  - Support both PKI (Grid certificates) and Kerberos

Spring 2009

    - Consolidation ongoing, Performance optimizations planned
  - See Frohner, Garcia, Ponce at Hepix Fall 2008 http://indico.twgrid.org/contributionDisplay.py?contribId=133&sessionId=90&confId=471

Mid 2009

- Castor Monitoring and SRM interface being improved
  - See De Witt, Lo Presti, Pokorski, Rekatsinas at Hepix Fall 2008 http://indico.twgrid.org/contributionDisplay.py?contribId=171&sessionId=90&confId=471

Increasing Tape Efficiency

Today (on v2.1.8)

- New tape queue management supporting recall / migration policies, access control lists, and user / group based priorities

Mid / End 2009

- New tape format with no performance drop when handling small files

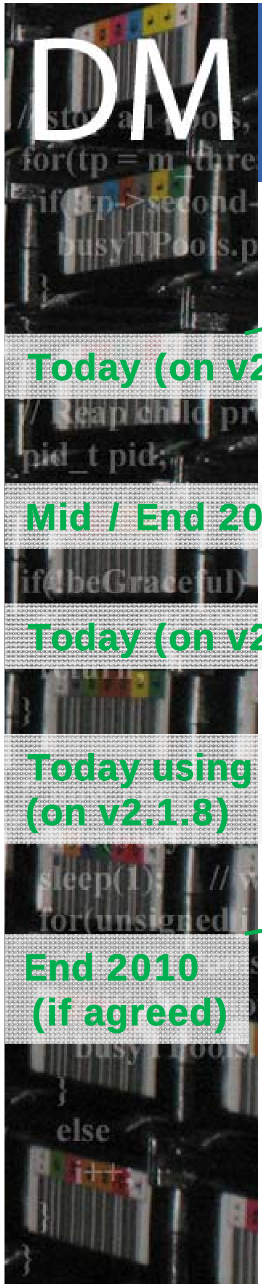
End 2009

- Data aggregation to allow managed tape migration / recall to try to match the drive speed (increased efficiency)
- See Bessone, Cancio, Murray, Taurelli at Hepix Fall 2008
  http://indico.twgrid.org/contributionDisplay.py?contribId=181&sessionId=90&confId=471
- See plan and timelines at
  https://twiki.cern.ch/twiki/bin/view/DataManagement/CastorTapeSubsystem

The access latency problem...

- Not a problem for the Tier-0 operation
  - – Scheduled I/O guarantees Tier-0 data streaming to tape and to Tier-1 which are not affected by the high LSF latency
- ... but a concern for interactive data access and analysis
  - – Unacceptable to wait seconds to open a file for read when it is already on disk

Reduced access latency



Today (on v2.1.8)

Mid / End 2009

Today (on v2.1.8)

Today using FUSE (on v2.1.8)

End 2010 (if agreed)

- Removed the LSF job scheduling for reading disk files
  - Latency reduces from seconds to milliseconds (like a file system)
  - Plan to remove job scheduling also for write operation
- Focussed to offer direct access to storage on Castor disk pools
  - Direct access offered with the XROOT interface which is already part of the current service definition
  - Excellent integration with root
  - Mountable file system possible
  - Additional access protocols can be added (eg. NFS 4.1)
- Performances can be increased by improving the name server
- See A.Peters at ACAT 08

http://indico.cern.ch/contributionDisplay.py?contribId=171&sessionId=29&confId=34666

New requirements for analysis

- Working groups on analysis requirements led by B. Panzer (Jul – Oct ‘08) for Tier-0, by M. Schulz (Nov’08 -) for Tiers-1/2
- (simplified) requirements collected so far
  - Analysis made on disk pools only
    - Tape access limited to managed and aggregated recall / migration requests from the experiments
    - No end-user recalls or direct access to tapes
  - Strong demand for direct file access
    - No unnecessary latencies
    - Mountable file system accessible from both interactive desktops and batch systems
  - Integration with the existing physics analysis tools (grid middleware and root)
  - Secure / manageable (Authentication, ACLs, Quota, Accounting, etc ...)
  - File catalogue and name server consistency

Areas being investigated @ CERN

- Deploy a new file system
  - High performance file system access
  - Mainstream software and services
- Offer direct file access to Castor files
  - Limited duplication of data
  - Consistency guaranteed and data management possible
  - Performance impacted by name server lookup

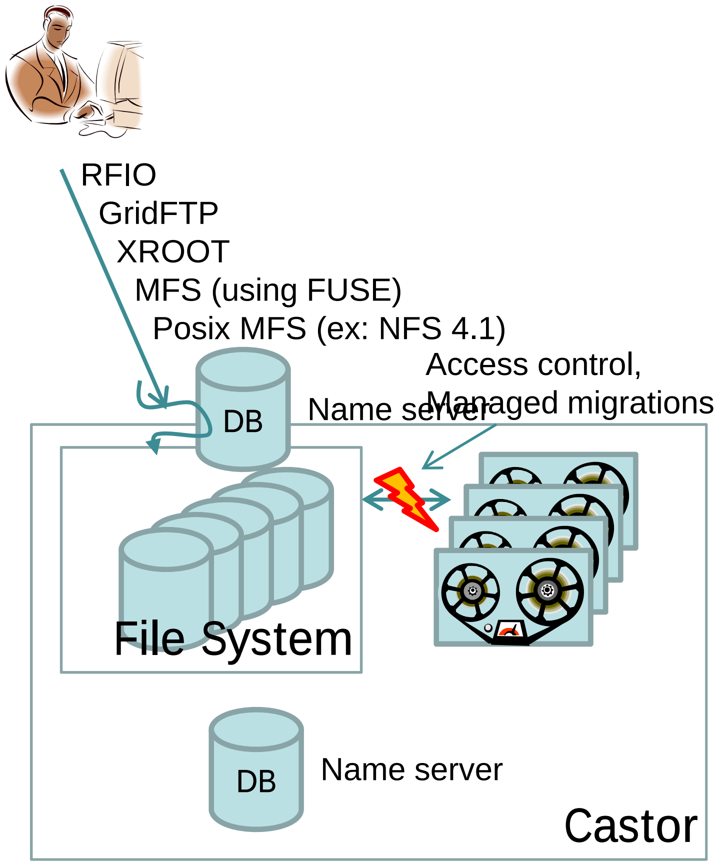
Area 1: A new file system

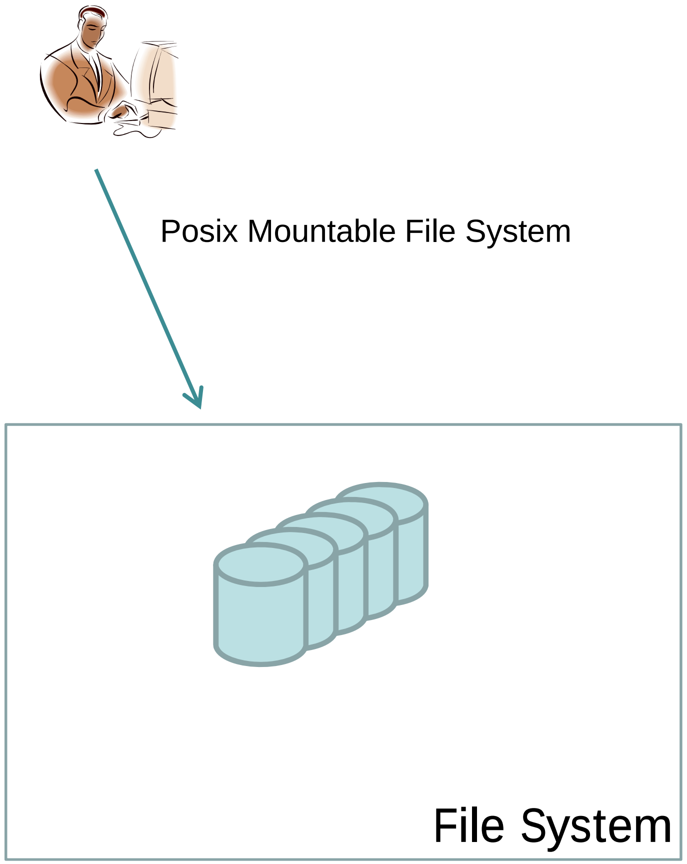
- High performance, highly scalable
  - Gpfs, Lustre, NFS...
- No end-client software necessary (posix) for *basic* file operations
- Long-term dependencies if *advanced* (specific) file operations features are exposed to end users
- Areas of research
  - Integration with existing infrastructure
  - Wide Area Transfer / Global namespace
  - Data replication (replicated pools / hot spots)
  - Tape integration
  - integration with the file catalogue
  - SRM interface (based on StoRM ?, BeStMan ?)
- Realistic timeline: End 2010 for the file system

Area 2: Castor direct access

- Direct file system access to Castor files with similar performance in term of requests/sec and latency
  - A new file system could also be considered as a storage technology for Castor file servers
  - But ... name server lookup generates performance overhead
- Access protocol
  - XROOT, RFIO, GridFTP (available now)
  - Mountable file system possible using Fuse (available now)
  - NFS 4.1 (as a future option)
- Name server / file catalogue consistency ensured
  - Global namespace
  - Same logical filename across pools and tapes
    - Integration with tapes (in place)
    - data management possible
- No long-term dependencies: separation between access protocols and storage technologies.

The options under investigation





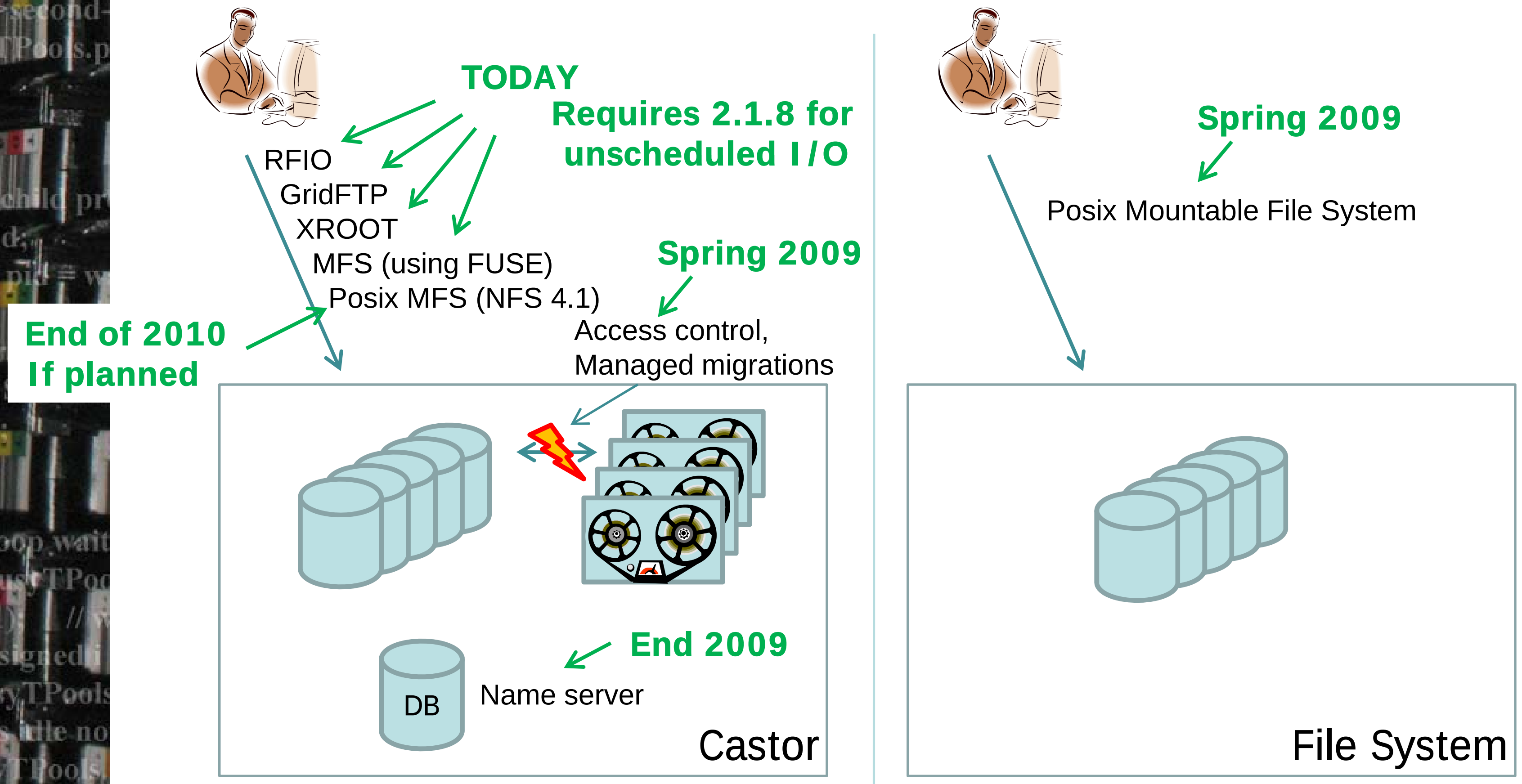
- Technology independent data protocol
- Centrally managed data solution

Storage technology change transparent (Lustre / GPFS / ...)
Internal architecture change transparent (Tape / Disk / Flash ...)
End-to-end consistency (eg checksums)

- Higher performance
- Data managed by the experiments

Namespace on tape differs from namespace on disk
Two files with same name can have different content
Two files with different name can have same content
IT-organized data migrations more heavy

Possible timescale



- Base functionality available today
- Spring 2009 – Security, Access control and Quotas
- End of 2009 – improved name servers
- End of 2010 – Native file system access as a new service

- Spring 2009 – Prototype
- End of 2009 – Visible from lxplus
- End of 2010 – Running as a new service

No service approved yet
These dates are only effort estimations

Conclusion

- Mountable file system is considered a strategic direction
  - Either with Castor or with a new file system
- Feedback welcome to understand to what extent experiments require a unique name space between files on disks and files on tape
- The two approaches implies different levels of consistency and data management automation experiments expect from IT

Questions / Discussion