CERN IT Department

Data Management Architecture Team Status Report

Dirk Düllmann

Physics Service Meeting,

19 June 2008

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

Topics

- short term - ongoing
  - security and access Control
  - repack strategy
- proposed projects
  - integrated monitoring
  - DB deployment model & schema consolidation
  - investigate “file system” options for Castor pools
  - evolve architecture to face future requirements for analysis (quota, opens/sec, small files)
  - tape efficiency: dataset clustering/aggregation, tape lables, disk cache for tape
  - client protocol consolidation (eg xrootd, rfio)
- vision/guesses - new media/scaling
  - storage hierarchy and changing media roles
  - federated storage and meta data components

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

Security and Access Control

- Two separate issues
  - both required on short term
- Experiments (ATLAS,CMS) :
  - insure that only production users can access tape (or get significant tape priority)
  - implemented via black/white lists
- but also risk of “standard” security issues
  - strong authentication, prevent unauthorised modification/deletion of data -> incident may happen at any time
  - development plan presented (Rosa @ Castor F2F)

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

End-to-end Monitoring

- Lack of integrated monitoring
  - high deployment costs and developer involvement in operations
  - lack of performance metrics to steer development priorities or validate proposed optimisations
  - lack of usage metrics for experiment mgmt to prioritise resource use
- Many metrics have already been added in recent Castor releases
  - Catalog of missing metrics being defined with deployment
  - Dedicated db schema for monitoring being defined and populated
    - for now from (often several joined) DLF records
- Propose to setting up web interface to access monitoring plots and resources consumption reports (similar to DM DB monitoring)
  - file lifetime and size distribution on disk
  - cache efficiency, hit/miss rate by user, time, fpath, fsize and tape
  - request clustering by user, time, fpath, fsize and tape
  - weekly/monthly: tape cost(mounts) per user and top users

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

Logging / Monitoring Transport

- Propose to replace direct DB connection to DLF DB from castor daemons by the use of syslog
  - reliable log transport and aggregation
    - following the initiative in FIO tape area
  - standard server infrastructure and client interface
    - eases integration of log messages from other DM components (xrootd, FTS, LFC, etc)
- Log-extractor process will then filter/extract pre-defined message
  - into the existing (rotating) DLF DB
  - into a longer term monitoring database schema
- Propose to integrate with the work already in progress by the tape area in FIO

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

Database use in CASTOR

CERN IT Department

- CASTOR is a database centric system
  - goal: all persistent state is kept in DB, daemons are stateless
- Currently each component in a separate DB clusters
  - One node active, one node idle fail-over
  - Should take advantage of full database consistency management (eg DB constraints) -> s/w being reviewed
- Propose to merge of name server, stager, logging/monitoring state in one DB cluster per experiment
  - would allow to implement cleanup/sync/correlation tasks in a more efficient/reliable way
  - use DB node dedication to insure component resources and limit cluster interconnect traffic
  - Operational experience from physics DB applications at CERN and CASTOR deployment at T1
- Propose to include DBA experts in the development projects
  - increase query plan stability, add DB side monitoring
  - fully exploit DB provided consistency

Data Management

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

DB Schema Review / DB<->Disk Consistency

CERN IT Department

- Many similarities between stager and SRM request handling
- Propose to integrate SRM requests with stager & drop separate DB
  - Similar approach works well in DPM
  - Schema changes required
  - Additional SRM load to stager seems manageable
- Review disk server <-> DB synchronisation
  - Significant DB load for the name service created by asynchronous synchronisation with disk pools content
    - Does DB have sufficient I/O headroom during backups?
    - Can we achieve (sufficient) consistency without shipping continuously full disk server inventories?
  - log inconsistency and eliminate causes
  - trigger synchronisation by volume inconsistency

Data Management

Software Process & Testing

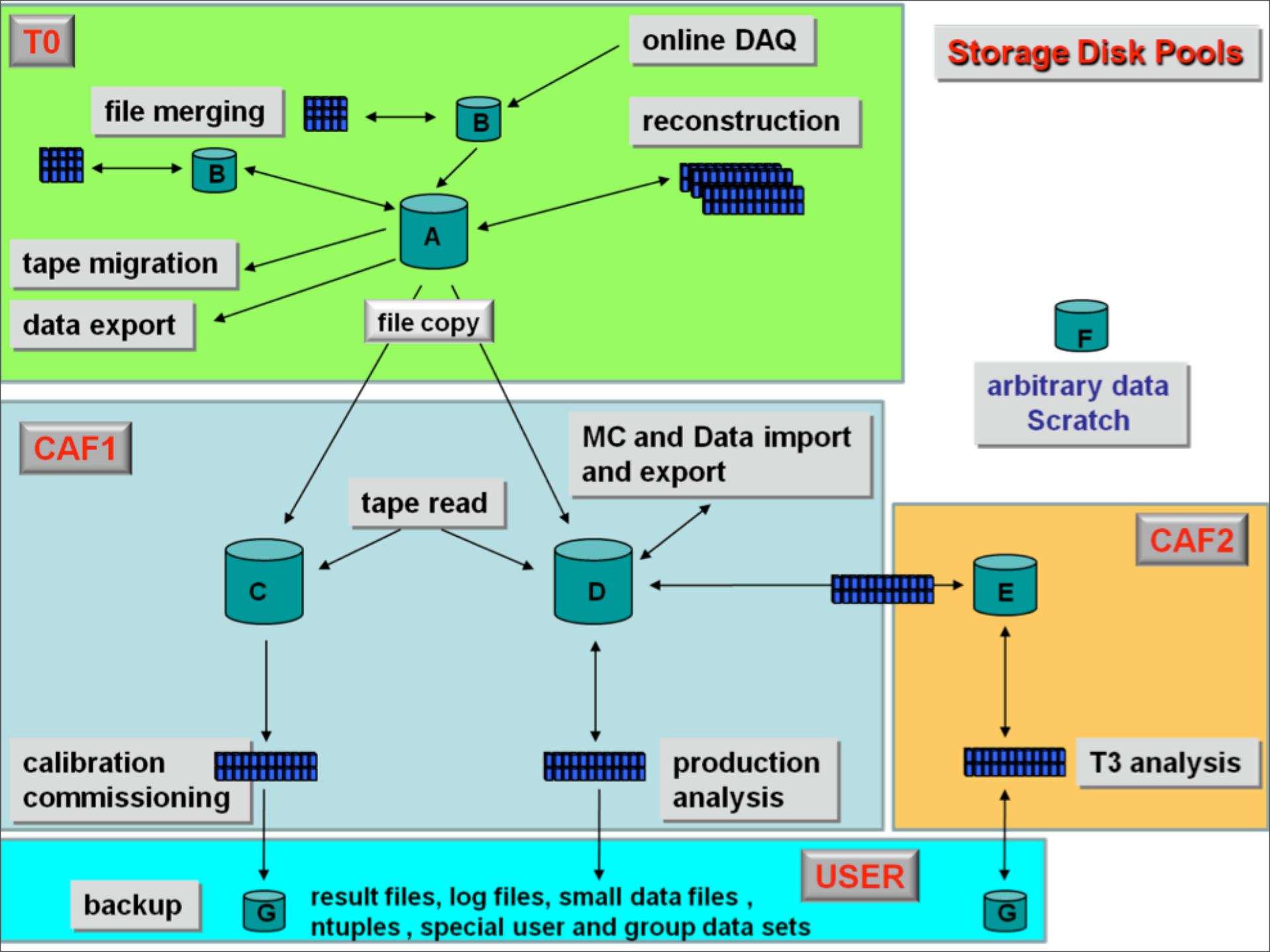
- Opportunities for (even small) changes after LHC start-up will decrease
  - maintain trust into release testing
  - test coverage needs to be extended to SRM
    - add main SRM use cases - eg access tape files before release
    - CERN team should perform pre-release tests also for SRM
- Component tests in addition to big-bang / stress tests
  - some development effort for (simple) component emulators / mock-ups
    - eg: run functional stager tests with emulated tape sub-system (with emulated tape errors)
  - may also allow to run component-wise scalability tests
    - eg: how far would the stager scale if scheduling bottleneck would be removed?

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

Analysis Requirements

- Requests from experiments for significant analysis setups now exist
  - came in late (but hardly as a surprise)
- Known/predicted requirements and a straw man model are summarised in a recent paper by Bernd
  - Shared Analysis Area (read-only, tape backup)
  - User Analysis Area (small files, on-disk copies, no tape)
- Impact for storage system
  - support for high file open rates (model dependent 100-1000/s)
  - need disk quota
- XROOTD seems well placed as efficient client protocol
  - back-end options:
    - castor by-pass a la ALICE
    - "fast" scheduling (yet to be demonstrated)
    - disk only
- Need to be closely involved in experiment activities to shape arising analysis models

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


T0
online DAQ
Storage Disk Pools
file merging
B
reconstruction
B
A
tape migration
data export
file copy
F
arbitrary data
Scratch
CAF1
MC and Data import
and export
tape read
CAF2
C
D
E
calibration
commissioning
production
analysis
T3 analysis
USER
backup
G
result files, log files, small data files ,
ntuples , special user and group data sets
G

Data Sets and End-to-End Clustering

- Experiments production and transfer systems are often based on “data set” concept
  - list of files which are main higher level granule of processing and data access
  - experiments ‘unfold’ into individual files for tape and inter-site transfers and afterwards ‘collect’ again
- Investigate if storage and transfers components could take direct advantage
- Data management system could internally aggregate data set files into a series of “data clusters”
  - subset of a user data set with defined minimum and maximum aggregation size (eg a few GB)
  - containing file meta-data (eg tar, filesystem image)
- Propose to perform aggregation between user pool and dedicated disk for tape and site transfers
  - No additional I/O introduced on user pools

Data Management

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

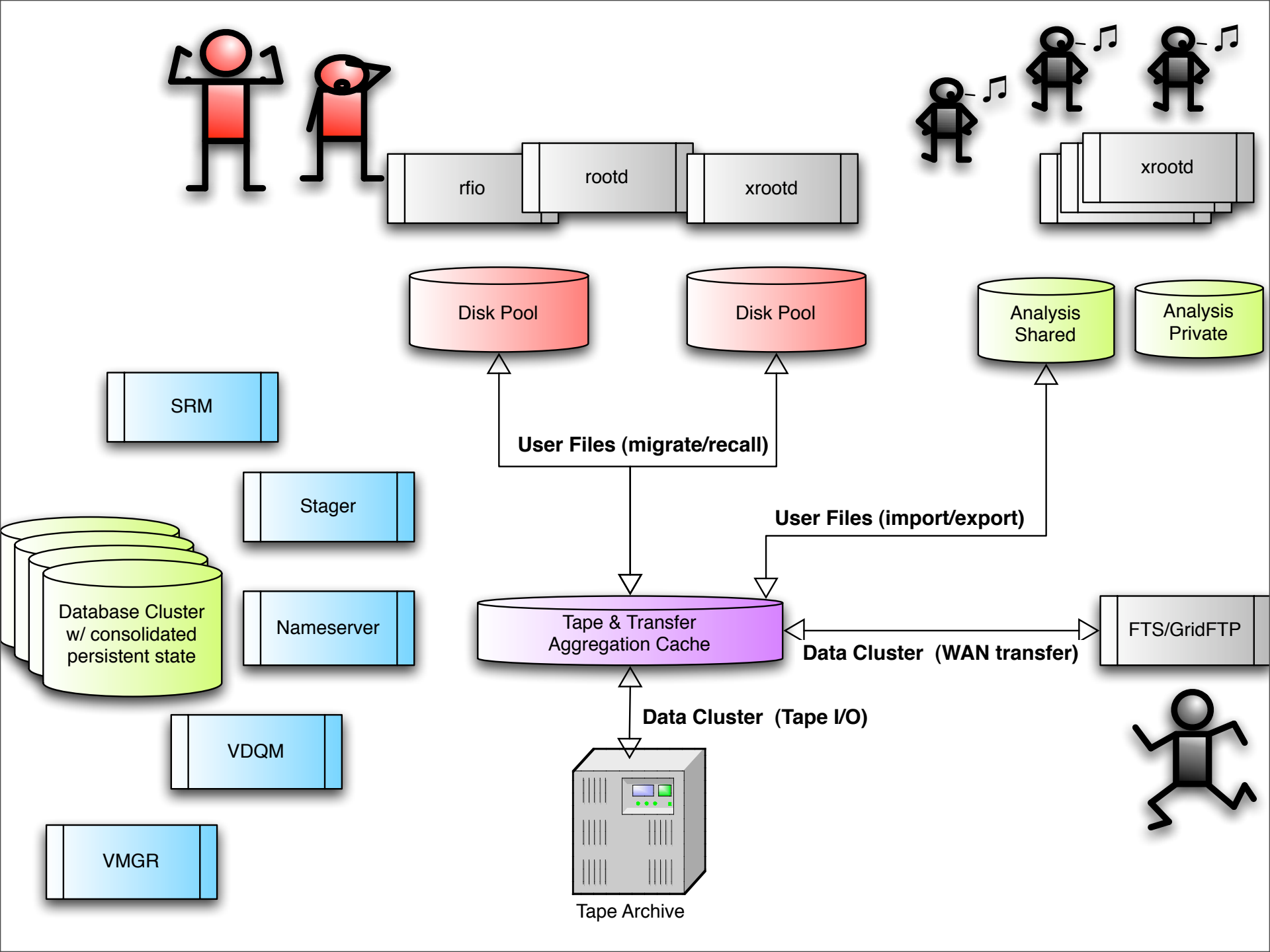
Advantages of Data Clusters and additional Disk Buffer

CERN IT Department

- for Tape transfers
  - avoid small file performance problems limiting efficient use of tape drives and robots
  - maintain clustering on tape as defined by user
  - insure tape drive transfer bandwidth
- for Site transfers
  - Smaller number of entities to track
  - De-coupled from competing user I/O
  - Fewer atomic catalogue updates

Data Management

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


rfio
rootd
xrootd
xrootd
Disk Pool
Disk Pool
Analysis Shared
Analysis Private
SRM
User Files (migrate/recall)
Stager
User Files (import/export)
Database Cluster w/ consolidated persistent state
Nameserver
Tape & Transfer Aggregation Cache
FTS/GridFTP
Data Cluster (WAN transfer)
Data Cluster (Tape I/O)
VDQM
VMGR
Tape Archive

Longer Term Issues & Strategy

- Changing media roles and new media types will come with more significant challenges
- A multi-level "disk" hierarchy will be required
  - to use active power management on archive disks (as tape replacement)
  - to integrate solid state disks (as user side cache)
  - staging/migration will still be there, but more in a more generalised way between storage tiers
- Latency of meta-data operations and data movements will need to be re-balanced with associated storage media-speed and latency
  - eg files in solid state disk will need low latency meta-data and transfers handling

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

A Scenario - Layered Storage Components

CERN IT Department

- Propose to prototype more layered and component based architectures
  - Storage blocks are autonomous and maintain local meta-data consistency
  - Provide a generic interface/protocol for data access and transfer
  - Individual storage blocks are federated and layered on top of each other
    - rather than managed via single/global services (eg global DB name space)
  - Low layers may deal with dataset clusters
    - good match for bulk transfer endpoints
  - Higher ones with user files
    - end user random file access
  - Avoid global scheduling with detailed internal information from different layer may face scalability and manageability problems
- Still many open questions, which need real life monitoring input and prototype studies to advance.

Data Management

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

Summary

- Several development projects have been identified to improve stability and performance of DM components
- Assume that this will continue to be background work
  - foreground being short term bug tracing / fixing
  - but need to move from brainstorming mode to project work
- Need to agree on relative priorities and required resources
- Plan to organise IT/DM accordingly

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

Topics Again...

- short term - ongoing
  - security and access Control
  - repack strategy
- proposed projects
  - integrated monitoring
  - DB deployment model & schema consolidation
  - investigate “file system” options for Castor pools
  - evolve architecture to face future requirements for analysis (quota, opens/sec, small files)
  - tape efficiency: dataset clustering/aggregation, tape lables, disk cache for tape
  - client protocol consolidation (eg xrootd, rfio)
- vision/guesses - new media/scaling
  - storage hierarchy and changing media roles
  - federated storage and meta data components

Data Management

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