Improvement Options for LHC Mass Storage and Data Management

Dirk Düllmann

HEPIX spring meeting @ CERN,

7 May 2008

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

Outline

- DM architecture discussions in IT Data Management group
  - Medium to long term data management strategy
  - Current CASTOR status and development
    -> Sebastien's talk later in this session
- Starting point for a status review
  - Known scalability limits
  - Expected access pattern and technology changes
  - Deployment and maintenance cost
- Main areas for improvement reviewed so far
  - System diagnostics and scalability model
  - Tape - efficiency and performance
  - Disk (or rather non-tape) storage hierarchy
  - Database use and internal consistency
- Still several items to cover for a complete priority list and plan

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

Working Group Mandate

CERN IT Department

- Review medium to long term evolution of data management s/w components now concentrated in DM group
  - CASTOR, DPM, LFC, FTS, ...
- Review possible architectural changes and prioritise them by their expected possible impact on
  - scalability
  - deployment cost
  - s/w manageability
- Driven by
  - LHC schedule and development resource constraints
  - commonalities between LHC experiment production models and emerging analysis models on the experiment side
  - existing (and possibly new) performance metrics from the running service
  - operational experience from deployment teams at CERN Tier 0 and its major partner sites

Data Management

Operational Mode

- Architecture team has been formed with
  - 2 key developers from CASTOR team
  - 2 key developers from Grid Deployment
  - open invitation to group and CASTOR project management
  - try to avoid big committee approach
- Currently meet twice a week
  - currently mostly brainstorming and data/idea collection mode
    - complemented with short conceptual prototypes
  - invite/consult key people from main stake holders
    - site deployment
    - experiment responsible for main areas (CDR, production, analysis)

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

Timeline

- Structure existing ideas and evolution options and summarise their impact on scalability and performance based on existing (and projected) use patterns
- Estimate the required development effort and propose a medium term development plan
- First round should conclude in summer with an agreed priority list of significant developments
  - Merge/rebalance with existing shorter term development plan

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

Starting point

- Development for many key components of the data management system now concentrated in IT-DM group
  - CASTOR, DPM, LFC, FTS, GFAL/lcg-util
  - Maintain close relationship to deployment in other CERN groups
- CCRC showed promising results for the tested data flows
  - well co-ordinated, large volume streams
  - tape & transfer centric at T0 and T1s
  - CASTOR development/deployment teams have managed to stabilise most the core system after a period of fire-fighting
- User data analysis plans arising inside experiments
  - resources for/from experiments soon becoming available
- Now is a good point in time to consolidate/complete the DM strategy
  - avoid trap of optimising the system only by write use cases and penalise read/analysis performance
  - collect missing elements and remove scalability problems
  - integrate experiment and site experience and plans

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

Expected Medium to Long Term Changes

CERN IT Department

- Experiment focus will soon move to efficient analysis and extraction of physics results
  - less co-ordinated and resource limited mode
  - CAF and Analysis use cases to be defined with experiments
    - mainly disk based
  - storage performance need to be kept for high rate data ingestion
- Several technology shifts are coming up...
  - New file systems (eg Lustre, ZFS, NFS 4.1) and more efficient client protocols (eg xrootd)
  - Jim Gray et al: tape is dead, disk is tape, flash is disk, flash is cool
  - Power managed disk libraries
- ...and will need to be integrated to take advantage
  - more storage hierarchy above tape
  - possibly replacing any non-archive tape use
- Management of evolution during LHC lifetime will be key to avoid expensive surprises

Data Management

Known Scalability Limitations

CERN IT Department

- File sizes and tape format
  - Still room to improve tape/robot efficiency arising from small files
  - Concatenation can/should done by experiments, but very large user level files may not be optimal for several reasons
    - increased I/O load during (user level) concat operations
    - decreased efficiency in user side caches (as files are cache global cache granule)
  - The storage system itself may be able to help avoiding the strong coupling between user files and efficient storage entities.
- File clustering on tape
  - File set striping to tapes - good for writes, but source of inefficiency for reads
  - Need an end-to-end clustering design that maintains the experiment knowledge about most likely read pattern into account
- I/O scheduling in CASTOR has been proven to work for critical data path to tape and tier 1 site
  - but implementation limits file access to O(20)/s
  - introduces significant complexity/coupling across storage layers -> maintenance/ deployment cost
  - can not assume that this will be sufficient for analysis use cases

Data Management

Maintenance and Deployments Issues

CERN IT Department

- Insuring a consistent state between different parts of the CASTOR system could save significant user and deployment time
- Overall system complexity and difficulties to obtain consistent and correlated diagnostics needs significant deployment effort
- A validated scaling model based on reliably measured metrics could allow to predict system performance wrt changing user requirements and planned h/w expansions.
- Need to establish the key metrics and provide consolidated tools to analyse/predict the overall system behaviour
  - Some work already took place: tape area, CCRC and CASTOR development
  - This needs to be now consolidated and obtain direct support by main CASTOR software
  - Analysis of this metrics should drive the prioritisation of developments
- Propose to work jointly with CASTOR deployment on a scaling model and its validation from production data and dedicated scalability tests

Data Management

Database use in CASTOR

CERN IT Department

- CASTOR is a database centric system
  - Moving towards all persistent state inside the database
  - All components should be stateless - eg daemons can restart after failure
  - Component scalability and availability is achieved via shared DB state and multiple component instances as required
- Current database use at CERN splits component state into several DB clusters (see Nilo's talk)
  - could take more advantage of full database consistency management (eg DB constraints) -> s/w being reviewed
  - merge of name server and stager state into a RAC cluster would allow to implement tedious deployment cleanup/sync tasks in a more efficient way
  - Could still use DB cluster node dedication to insure component resources and limit cluster interconnect traffic
- Operational experience from experiment DB applications at CERN and CASTOR deployment at RAL

Data Management

Data Sets and End-to-End Clustering

CERN IT Department

- 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
  - often data sets are actually 'unfolded' into individual files for tape and inter-site transfers and afterwards 'collected' again by experiment software
- Investigate if storage and transfers components could take more direct advantage of exploiting the established concept
- Data management system could internally aggregate data set files into a much smaller series of "data clusters"
  - subset of a user data set with defined minimum and maximum aggregation size (eg a few GB)
  - containing local meta-data (eg tar, filesystem image)
- Propose to perform the necessary aggregation between user pool and dedicated disk pool for tape and site transfers
  - No additional I/O introduced on user pools

Data Management

Medium Term Advantages of Data Clusters and additional Disk Buffer

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

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

A Disk Pool for Analysis

- Main Assumptions
  - Analysis will (have to) be mainly disk based
    - Large fraction of relevant data will be online and read-only with tape archive
    - Resource allocation managed by physics working group responsibles
    - Some additional writable area for private user files with disk based redundancy
  - Work patterns are
    - expected to be very different from CCRC activity
    - still only vaguely defined and may only stabilise after one year of LHC running
    - we may have to support a larger number of file accesses than achievable with current CASTOR set-ups
- Option prototyped: start from a disk centric set-up, taking advantage of several simplifications
  - Use a cluster wide file system for implementing the file name space and meta data handling (eg ACL handling and quota)
  - Complement this by a low overhead protocol for user access, which adds missing authentication and basic load balancing
  - Interface to tape archive via explicit data-set import and export

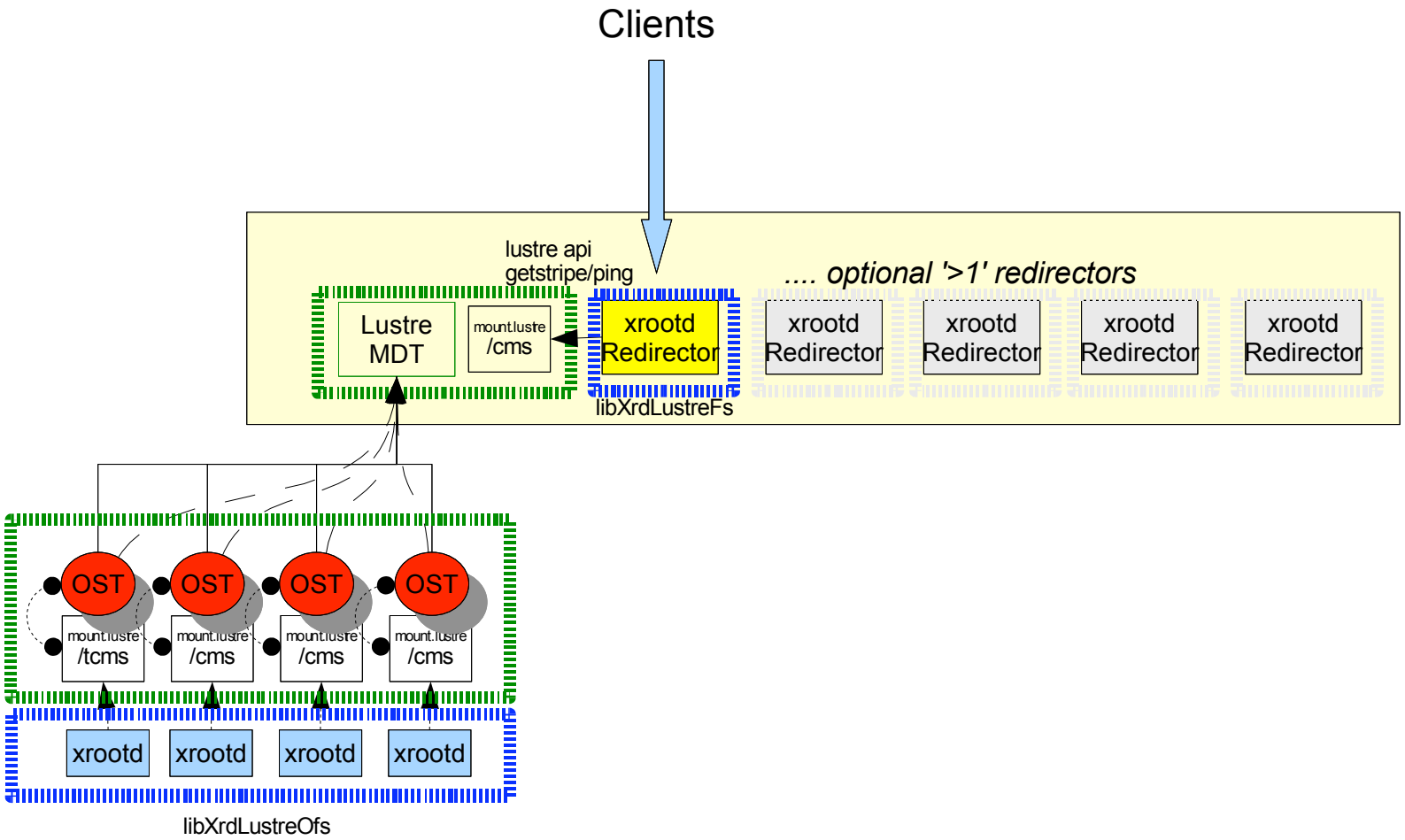
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

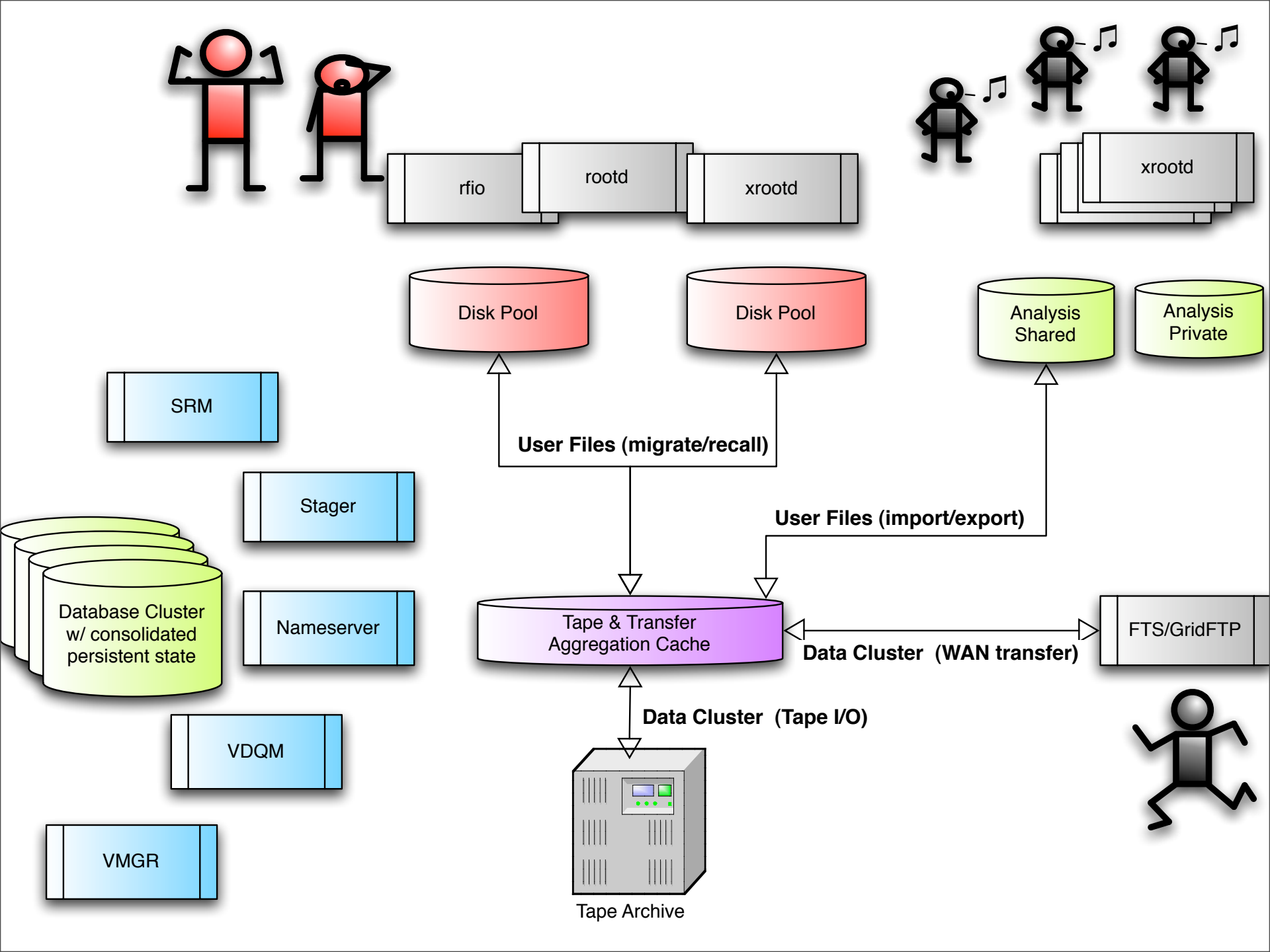
Xroot over lustre

Andreas.Joachim.Peters@cern.ch

XrootLustre Setup

Standard




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
Data Cluster (WAN transfer)
FTS/GridFTP
Data Cluster (Tape I/O)
VDQM
VMGR
Tape Archive

Summary

- DM architecture group has just started
  - So far in collection mode seeking input from deployment teams, experiment users and external sites
- Some areas of improvement are being proposed and/or prototyped
  - consolidated system for diagnostics with a scalability model behind
  - aggregation and clustering based on data sets with a dedicated tape/transfer cache
  - review of database deployment model and schema to improve consistency
  - prototype for an 'analysis pool' based on lustre complemented by xrootd
- Current proposals mainly address medium term changes with limited impact on existing deployment
  - longer term evolution is taken into account to be prepared for a less tape centric approach
- Will continue to discuss evaluations and development options with other storage development teams and the deployment/user community

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