CASTOR developments status and plans

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

Outline

- CERN's Data Management strategy
  - Reminder on the context
  - Ongoing projects
  - Long term vision
- Handling the analysis with CASTOR
  - Overview of the xroot approach
- Release agenda
  - Latest releases and new features
  - Timelines for upcoming releases

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

Context reminder

- Data Management group one year old
  - Contains CASTOR, DPM, LFC, FTS, Oracle services for physics...
  - Groups expertise in data management in order to benefit from the synergies
- Task force from January to June 2008
  - Defined the strategy of the group
  - Reviewed the design choices of CASTOR
  - Incorporated the analysis case in the CASTOR roadmap

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

Output of the task force

- Overall the CERN CASTOR instances are running fine
- Few weak points were identified where improvements are necessary
  - These may explain some Tier 1 issues
- 5 projects have been launched to address them
  - Tape efficiency, repack
  - File access protocols and data latency
  - Security
  - SRM and Database schema
  - Monitoring

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

Tape efficiency

- Improvements needed for tape access
  - Recall efficiency
  - Migration efficiency of small file
    - Mandatory for CERN's repack exercise
- Three areas where developments are expected :
  - File aggregation on tape
    - e.g. Tar/zip files of 10GB
  - New tape format
    - less tape marks, no metadata, ...
  - Repack strategies
- Deployment of VDQM 2

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

Tape efficiency status

- Repack
  - Several functional enhancements, including post repack actions (e.g. pool movement) and throttling (max nb of tapes/files)
  - In production since December, now stable
  - See also Tim's talk on Thursday
- VDQM 2
  - Provides user-based priorities (using stager admin tools) and tape-based priorities (using VDQM admin tools)
  - Redundant setup possible (two servers, common DB)
- VMGR
  - "offline" status for libraries: problematic ones can be set offline.
    - recall and migration requests will remain on hold
- More details in German's presentation
  - Especially on post 2.1.8 features (tape format, aggregation, ...)

Access protocols and latency

- CASTOR needs to support analysis activity
  - Small files, many concurrent streams
  - Mainly disk, with aggregated tape recalls
  - Low latency for file opening
- The XROOT protocol and I/O server have been choosen to achieve this
  - With CASTOR specific extensions
- Practically
  - Strengthen direct file access using XROOTD
  - Think about direct write operations from XROOTD
  - Simplication of the scheduler/stager with reduced latency
    - Disk management is the responsibility of XROOTD
- More details later in this talk

Security

- Goal is to ensure that all software components can be deployed in secure mode
  - Every user is authenticated (authentication)
  - Every action is logged (accounting)
  - Every resource can be protected (authorization)
- Ensure complete interoperability of Grid and local users
- Allow coexistence of secure/insecure interfaces and offer migration plans
- In practice
  - Kerberos 5 or GSI authentication for all client connections (NS, RH, transfers)

Security status

- Implementation done last year
- Functionnality tests are all ok
- But stress tests have revealed a few issues
  - Kerberos has problems on SLC4
    - The default library is not thread safe
  - Kerberos may be slow on SLC5
    - If you do not disable the replay cache
  - GSI authentication may lead to RH dead locks
    - Investigations ongoing on locking issues, seem to be internal to the globus library
- Production version expected in 2.1.9 or in a later 2.1.8
  - But only supporting SLC5

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

SRM and Database schema

- Goal
  - Plan to combine the stager and the SRM software
    - Would allow also to merge the 2 databases
  - This may have an impact on the name server
- Today
  - SRM stays independent of CASTOR
  - Look at possible optimization of the CASTOR DBs
    - at possible simplifications (e.g. using DB link)
  - Review the nameserver interface for internal usage
    - Try to optimize DB usage
    - Try to see whether we can achieve needed performance for filesystem usage

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

SRM and DB schema status

- SRM 2.7 stabilized, 2.8 roadmap defined
  - improved logging
    - tracing of requests through their lifetime
  - time-order request processing
- Prototype of new nameserver API
  - for internal usage, with direct DB access
  - ensures a single round trip to DB
    - logic implemented in PL/SQL
  - stress tests have been run
    - number of stats/s goes from ~400 to ~10K (!!)
  - plans are to use this for repack and xroot daemons

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

Monitoring

- Goals
  - **Ease operation**
  - Cockpit of key indicators of the current status
  - Automated alarm before performances reach unacceptable values
  - See (and fix) the problems before the users see it
  - Real time, automatic detection of pathological cases
- Current state
  - A first implementation is ready for next 2.1.8 release
    - And passed extensive stress testing
  - See Dennis' presentation for more details

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

The next step

- 5 projects to address CASTOR weaknesses
- What about the future ?
  - New requirements may raise
    - Mountable storage
    - Data safety for disk only
  - New hardware may change the current picture
    - 10GB network interfaces
    - iSCSI storage
  - New technologies are coming
    - Solid state disks
    - Distributed file systems
- We need to be ready for these changes

DM's Long term vision

- Mountable storage is a plus for analysis
  - It simplifies the client software
  - But puts high pressure on the implementation
    - Think of a “configure;make;make install” in CASTOR
- The new generation of distributed filesystems is mature (e.g. Lustre)
  - Provide excellent performances
  - Candidate for analysis + a replacement for AFS
- Common solution to Mass Storage and analysis
  - Has advantages in terms of deployment effort
  - Provides backup for analysis data
  - Provides FS efficiency to disk cache of the MSS

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

Implications for CASTOR

- Test of mountable CASTOR
  - By taking advantage of XROOT + FUSE
  - To find out usage patterns and concerns
- Explore new, efficient solutions for disk only pools
  - Lustre
  - XCFS : Xroot Catalog FileSystem
- Improve our metadata efficiency
  - Nameserver needs to support FS load
- Pluggable tape backend
  - With small file support via aggregation

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

Ongoing explorations

- Mountable CASTOR via FUSE and XROOT
  - code exists
  - But the visible namespace is still debated
  - The nameserver 'slowness' is a major concern
    - Under study as mentioned
- XCFS prototype under stress test
  - XROOT Catalog FileSystem
  - The test is run over ~400 nodes
  - As a candidate for disk only analysis
- Lustre + Xroot prototype also under stress test
  - Lustre backend
  - Clients access only via the Xroot protocol

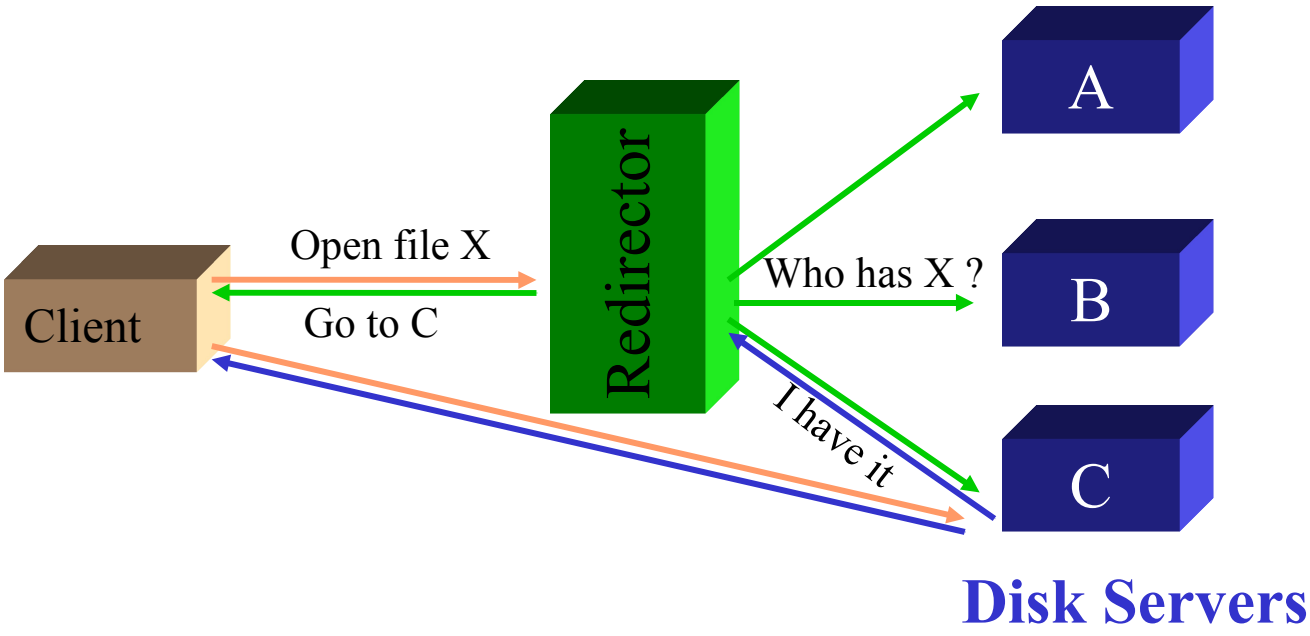
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Analysis in CASTOR : the XROOT approach

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

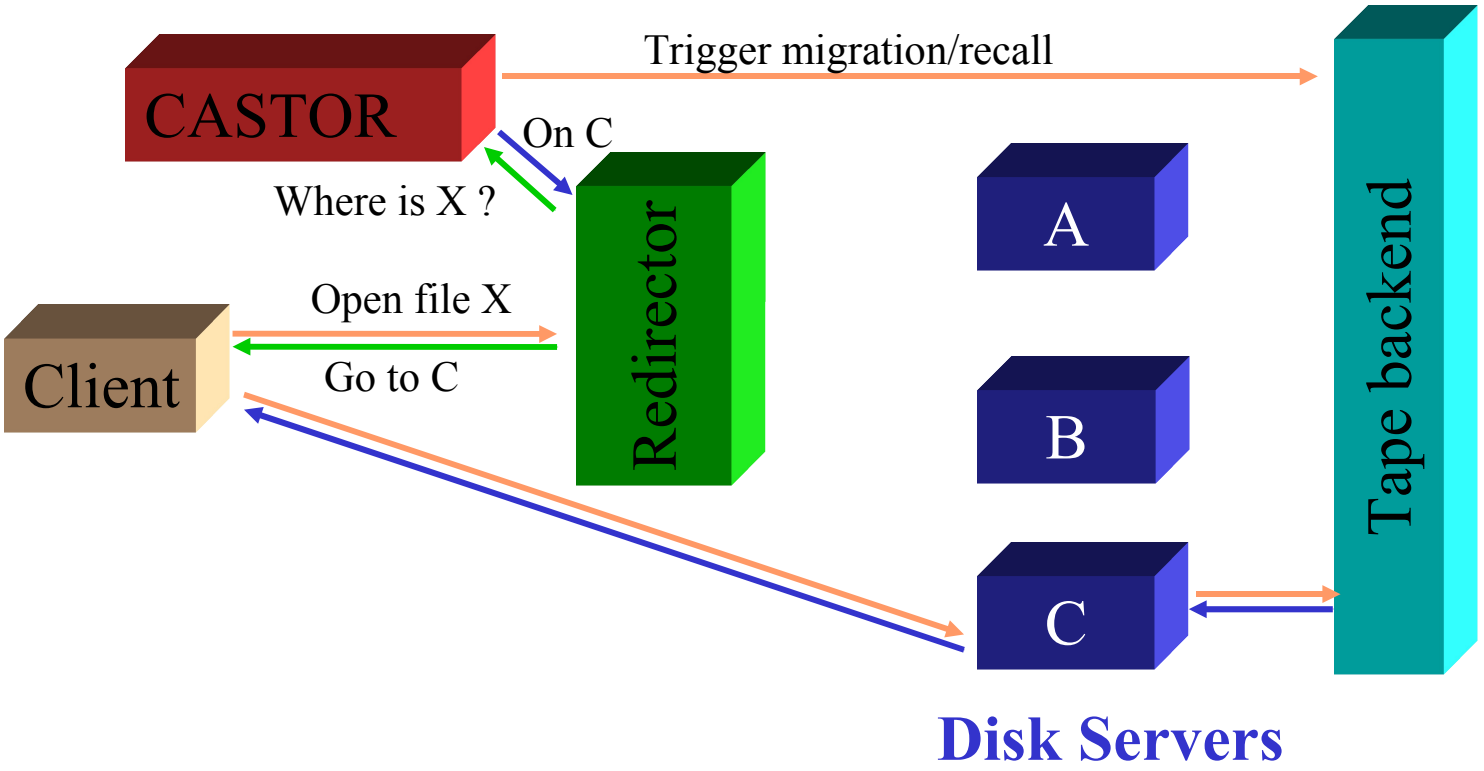
XROOT overview

- Client connects to a redirector node
  - This redirector finds out where the file is
  - It handles a cache of recent files for efficiency
- Client then connects directly to the node holding the data



XROOT in CASTOR

- Client connects to a redirector node
- The redirector asks CASTOR where the file is
- Client then connects directly to the node holding the data
- CASTOR handles tapes in the back



The gains

- Benefits from low latency of XROOT
  - 80ms per file opening (1-2s for CASTOR)
  - few ms if XROOT cache is activated
- Many connections per second (small files)
  - Scheduling of CASTOR can be dropped
    - Or kept for full scheduling
  - >700 connections per second if dropped
- And native xroot for bandwidth optimization
  - Can serve concurrently 100s of streams per node
  - Castor extensions allow bandwidth reservation for streaming clients (e.g. tape backend)

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

Practically

- New XROOT plugin in CASTOR
  - Tighter integration, aware of CASTOR concepts
    - e.g. service class, disabled disk server
  - Using CASTOR namespace, e.g. nameserver API
- Extensions of XROOT
  - Security (Globus, kerberos)
  - Stream scheduling on a disk server
    - Ability to dynamically lower throughput dedicated to users when a tape stream starts
  - Configurable redirector
    - Can use its cache or CASTOR (or both)
    - Can use its scheduling or CASTOR's (or both)

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

Current deployment (2.1.8)

- Stream scheduling not present
  - Under test
- Single configuration deployed
  - Only uses CASTOR cache
  - Only uses XROOT scheduling
- On a new CASTOR instance : c2cernt3
  - Dedicated to analysis of Atlas and CMS
  - Disk only, without SRM,FTS,...
- Other production instances will be upgraded
  - But existing pools will not go to XROOT soon

CASTOR releases status and agenda

CASTOR 2.1.7

- Is in maintenance mode
  - Stable, deployed in production at all sites
  - No new features included for quite long
  - Bug fix releases for major issues
- Will be phased out in the next few months at CERN
  - Repack and analysis setups already running 2.1.8
  - Plans to upgrade the other production instances in March/April
- Would not be supported anymore after the release of 2.1.9 according to current rules
  - i.e. sometime in Spring

CASTOR 2.1.8

- Now stabilized
  - First official release is 3 months old
  - For ~2 months on CERN's repack and analysis setups
  - No new features added (2.1.8 is a branch in CVS)
- Proposal of the CERN's operation team to build a 2.1.8-6 with important fixes/features backported and to switch to maintenance mode
  - Draining tools for diskservers (#37688)
  - stage_qry -s fix for available space (#27304)
  - Castor NS overwrite (#44799)
  - Enhanced test suite with xroot tests (#45506)
- 2.1.8-6 should be available by end of February
  - Deployed 2-4 weeks later on CERN production instances

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

Some features of 2.1.8 (1)

- Support for replication on close
  - And disk-n support coming in 2.1.8-6
- Improved GC logging
  - LastAccessTime, NbAccesses, GcWeight, SvcClass, GcType
- Basic user space accounting
- High level monitoring interface
- New implementation of stagerJob
  - improved logging and new xroot interface
- Enabled checksuming support in RFIOD
- Official release of VDQM2 (stateless, priorities)

Some features of 2.1.8 (2)

- Ordering of requests is now respected
- Support for OFFLINE libraries
- Deny stager_rm in DxT1 service class if file is not migrated and no other replica is entitled for migration
- End-to-end file check summing
  - New nssetchecksum command
- Support for forcing the ns host in the stager
  - The use of this feature drops support for multiple nameserver Dbs
- … and much much more, cf. release notes

CASTOR 2.1.9

- Current development version (head of CVS)
- Will include
  - Improved nameserver
    - Including dedicated direct DB interfaces
  - Further xroot integration (write case)
  - Revisited build infrastructure
    - Ability to build only client/tape part
- Timelines not yet very precise
  - Spring/Summer 2009
  - Deployment before LHC startup unlikely for T0 with current LHC scheduler

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

Summary

- Long/Medium term vision are now clear
  - Analysis, XROOT, filesystems
  - Explorations are ongoing with promissing results
    - Targetting 2.1.9 release
- analysis is coming,
  XROOT will play an important role
  - CASTOR 2.1.8 provides the initial functionnality
  - Efforts ongoing to improve efficiency
- For the LHC run 2009, the recommended CASTOR version is 2.1.8
  - Tier 1s are encouraged to upgrade before LHC starts
  - Especially since the first LHC run is expected to be long, until autumn 2010

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