CASTOR status and overview

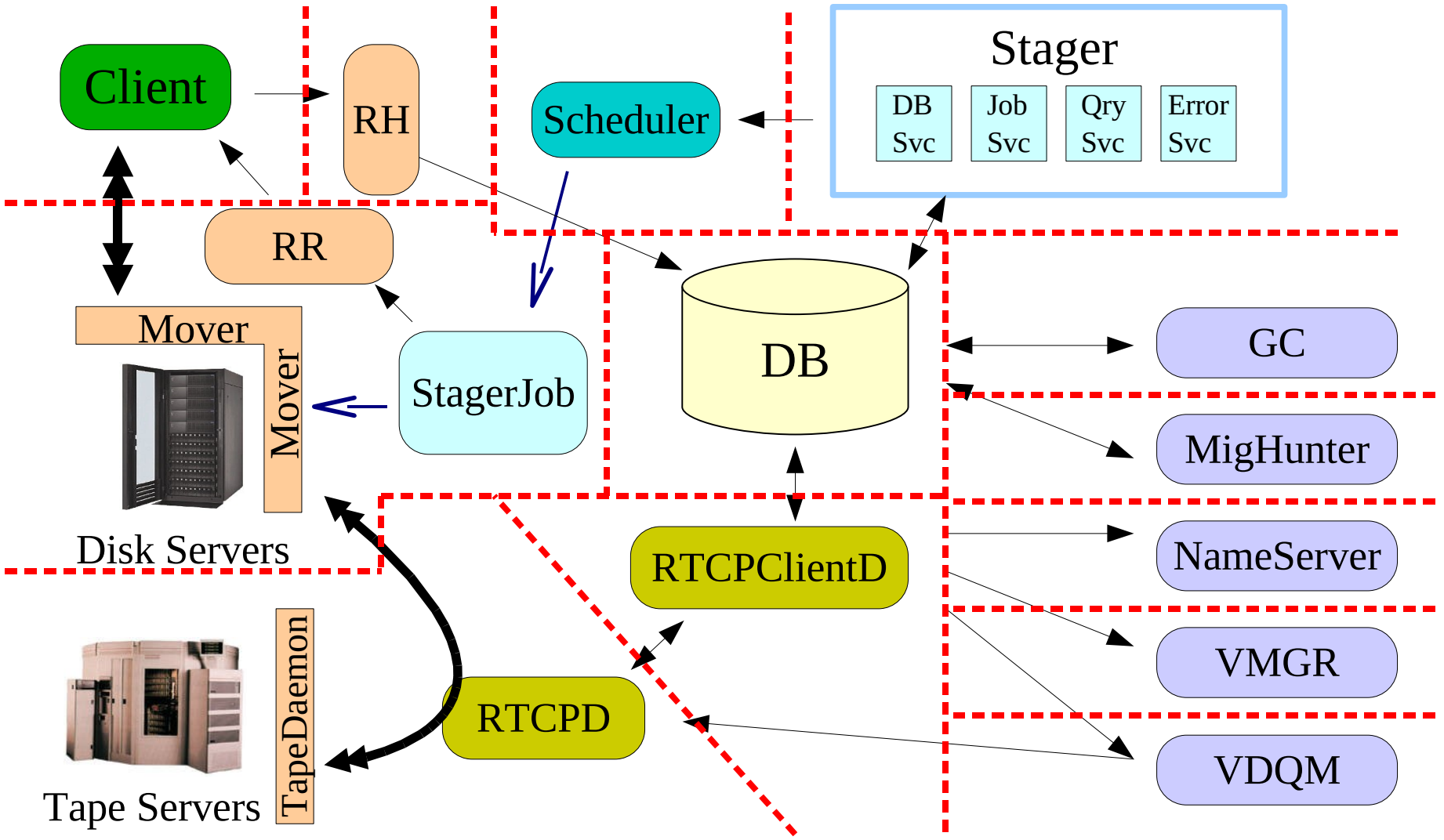
HEPIX, ROMA, April 6 2006

Sebastien Ponce, CERN - IT

Outline

- Quick Overview
- More details (Michael's "list of criteria")
  - Functionalities
  - Security aspects
  - Scalability
  - Flexibility
  - Configuration and Operation
  - Packaging and support

CASTOR 2 Overview



Main evolutions vs CASTOR 1

- Architecture changes
  - DB centric with stateless deamons
  - light clients
  - externalized scheduling and policies
- Tape optimizations
  - less mounts
  - possible optimization of the read order
  - guaranteed throughput
- New features
  - see next slides

Functionality (1)

- SRM conventions for client command set
  - get, put, prepareToGet, prepareToPut
- Transactions for input streams
  - prepareToPut, ..., put, ... , putDone
- New queries
  - support for user tags (-U option)
  - Incremental queries for newly staged files
    - stager_qry --getnext -U SC4
  - regular expressions and directory queries for staged files
    - stager_qry --regexp 'HEPIX.*2006'
    - stager_qry /castor/user/s/sponce/

Functionality (2)

- Pluggable policies
  - for recall, migration, garbage collection, I/O scheduling
  - Defined per disk pool but centrally written
  - Allows support for
    - volatile storage (GC, no migration)
    - durable storage (no GC, no migration)
    - permanent storage (GC, migration)
- Request priorities
  - scheduling through external pluggable scheduler
  - only LSF supported right now
  - single queue, priorities managed by admin

Functionalities (3)

- "Pluggable" protocols
  - rfio and rootd are integrated with CASTOR :
    - rfio://server:port//castor/cern.ch/...
    - root://server:port//castor/cern.ch/...
  - gridFTP v1 available on wan pools :
    - gsiftp://server:port//local/mnt/point/...
    - but not integrated, RFIO used in the background
  - xrootd being integrated
  - protocol addition is easy but no clean interface
- SRM v1 and v2 interfaces

Security aspects

- Authorization
  - per file ACLs at the namespace level
  - restricted access to diskservers (rootd, rfio, xrootd)
- Authentication
  - strong authentication under work
    - running currently at the prototype level
- Resiliency against hardware failures
  - any node can die with no major impact
    - relying on the DB for data, all deamons can be replicated
- Disaster recovery
  - regular backups of the DBs

Scalability

- Current numbers
  - 46M files, 4.6 PB of data (->average size is 100M)
  - 715 TB of disk space, up to 1M files staged
- Key factors
  - DB capability
    - no major concern, even with single CPU, midrange server
  - Scheduling
    - limited by LSF CASTOR plugin
    - currently limited to ~10 file accesses per second
    - under investigations
  - all daemons are stateless and can be replicated

Scalability tests

- CERN's internal challenges setup
  - setup
    - 50 diskservers (~240 TB disk)
    - 30 tape drives (3592, LTO3 and T10K)
    - 120 clients nodes
  - results
    - 2.2 GB/s incoming, 1 GB/s read back + 1.2 GB/s to tape
    - stable for 2 days, only limited by switches
- Practical metric
  - 60 MB/s per tape drive (40 MB/s for LTOs)
  - max 3 streams per server (2 Out, 1 In)

Flexibility

- Hardware running at CERN :
  - libraries : IBM 3584, StorageTek PowderHorn and SL8500
  - drives : IBM 3592, StorageTek 9940, LTO3, T10000
- Namespace administration
  - the namespace was always unique
  - no need to split but feasible (DB export and import)
- OS
  - only Linux supported on the server side (SLC3/4)
  - ia32, ia64, x86_64 architectures

Configuration

- Site customization
  - no recompilation, only config files
  - ~10 files and 40 lines to adapt
- 2 typical configurations
  - Tier 0
    - 7 to 9 central servers + 3 DB nodes + n diskservers
  - Tier 1
    - 4 central servers + 2-3 DB nodes + n diskservers
- CERN configuration
  - 6 instances running (4 exp, sc4, test)
  - all nodes are quattor managed

Operation

- Operation
  - with < 2 FTE : very restricted system
    - we don't expect small sites to run CASTOR2
  - 2 to 4 FTE : TIER 1 level, if some external expertise is provided besides (ORACLE, LSF)
  - Tier 0 runs with 6-8 FTE
- Admin tools
  - Web based centralized logging system
  - Set of scripts with man pages
  - Lemon sensors

Packaging and support

- Packaging
  - RPMs on the castor web site : http://castor.cern.ch
  - Release notes and update scripts provided
  - Presentations on the architecture and installation guide provided. User guide is being updated
- Support
  - 2 level of user support :
    - operation team : ~ 3 people
    - development team : ~ 2-3 people
  - we target Tier1s and have special agreements
- License
  - open source, moving from GPL to EGEE