Castor
status and plans

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

Sebastien Ponce, Hepix, May 7th 2008

Outline

- Quick overview of CASTOR
- Recent evolution and improvements
- Current state and setup
  - Tier 0
  - Tier 1s
- Some performance numbers
- Plans for the near future

CASTOR2 overview

- a mass storage solution targeting the CERN Tier 0 and the Tier 1s
- handles a tape back-end and a disk cache in the front-end
- it is the successor of SHIFT and CASTOR 1
  - was triggered by LHC needs
  - brings better scalability

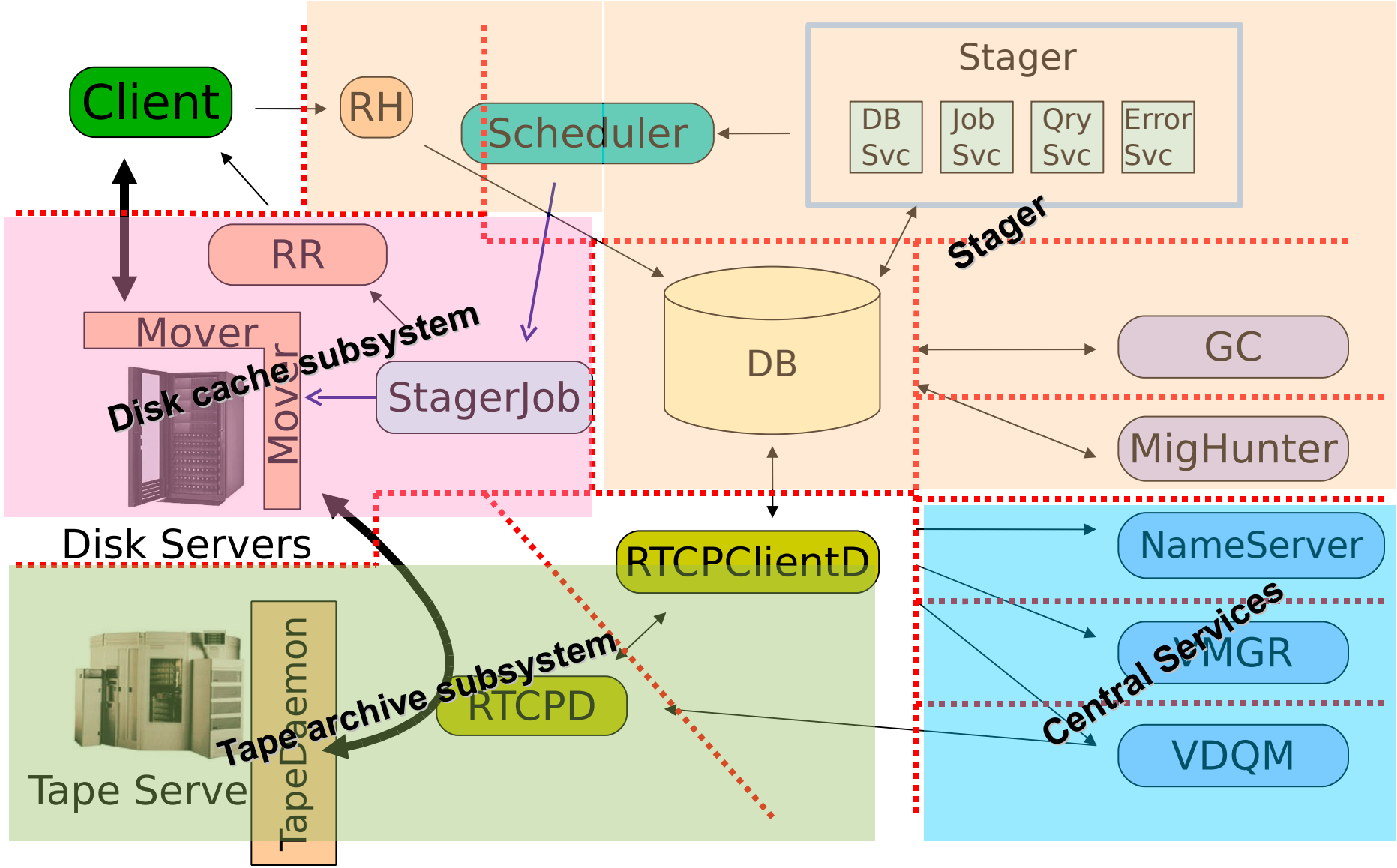
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Key points

- Database centric
  - using stateless redundant daemons
- 2 layers of storage
  - disk + tape
- a unique namespace
  - /castor/cern.ch/...
  - cross instances
  - but not cross sites

Castor 2 Architecture

CERN IT Department



Main features

- SRM like user interface
  - get/put/putDone
- Actual SRM 2 interface
  - developed and supported by RAL
- pluggable policies for decisions
  - migration, recall, GC, scheduling...
- pluggable protocols
  - supporting rfio, root, gridFTP, xroot

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

Latest improvements (1)

(Since 2006 spring Hepix, CASTOR 2.0.3)

- Rewrote Monitoring and I/O scheduling
  - allows better & faster scheduling
  - allows error recovery
  - added scheduling of internal replication of files
- full implementation of disk only pools
  - failure of incoming requests when pool is full was missing
  - targeted cleaning was missing

Latest improvements (2)

- Introduced pool level user restrictions
  - based on white & black list mechanism
  - permissions are based on request type, pool, user id and group id
- extensions of the policies
  - especially for tape side with stream, migration and recall policies
  - allowed great improvements of tape efficiency e.g. for migrations

Latest improvements (3)

- Disk level checksum of files
  - from the entrance of the system all way through
  - using extended attributes of the filesystem
  - allows to detect disk corruptions before migration
- internal component rewrite
  - repack, the tape copy mechanism
  - VDQM, the drive queue manager

Current situation

- 2 versions are concurrently supported
  - 2.1.6 series
    - the de facto standard
  - 2.1.7 series
    - the newest, deployed on Tier0 for Atlas and CMS
  - main differences are
    - additional consistency checks
    - scheduling optimizations
    - GC optimizations
    - better logging

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

SRM2 situation

- the CASTOR SRM2 interface is developed and coordinated by RAL
  - the CASTOR dev team considers it as an (important) external client
- Current production version is 1.3-21
  - in stabilization mode
    - only bug fixes and MoU agreed extensions will be introduced
  - being ported to most recent version of the CASTOR client library : 2.1.7

Deployment (Tier 0)

- 5 production instances of CASTOR
  - Alice, Atlas, CMS, LHCb & public
  - mix of 2.1.6 and 2.1.7 versions
  - common namespace and tape part
- Some numbers :

	nb nodes	disk space (TB)	av I/O (MB/s) in/out (march 08)	nb files on disk
Alice	47	238	85/257	1.2 M
Atlas	126	681	126/337	3.5 M
CMS	213	1093	1200/1100	570 K
LHCb	51	249	16/35	790 K
public	49	232	72/217	1.9 M
Total	486	2493	1500/2000	8.1M

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

Deployment (Tier 0)

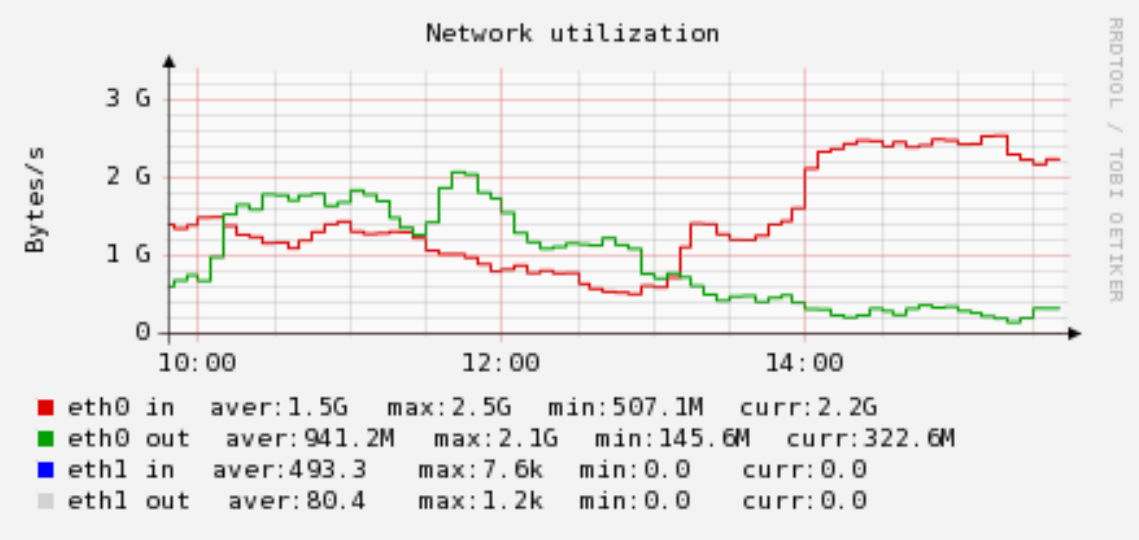
- In the namespace
  - 95 M files
  - recently used the 200 M fileid
- On the tape side :
  - 18.3 PB of tape storage capacity
  - 11.2 PB used on tape
  - 125 tape drives

Deployment (Tier 1s)

- 3 Tier 1s are running CASTOR2 in production
  - RAL (UK)
  - CNAF (Italy)
  - ASGC (Taiwan)
- All running 2.1.6
  - upgrade to 2.1.7 foreseen only after the CCRC, in June

Performance numbers (1)

- handles a large number of hardware
  - 800 diskservers, 125 drives, 5 libraries
- to reach high speed tape migration

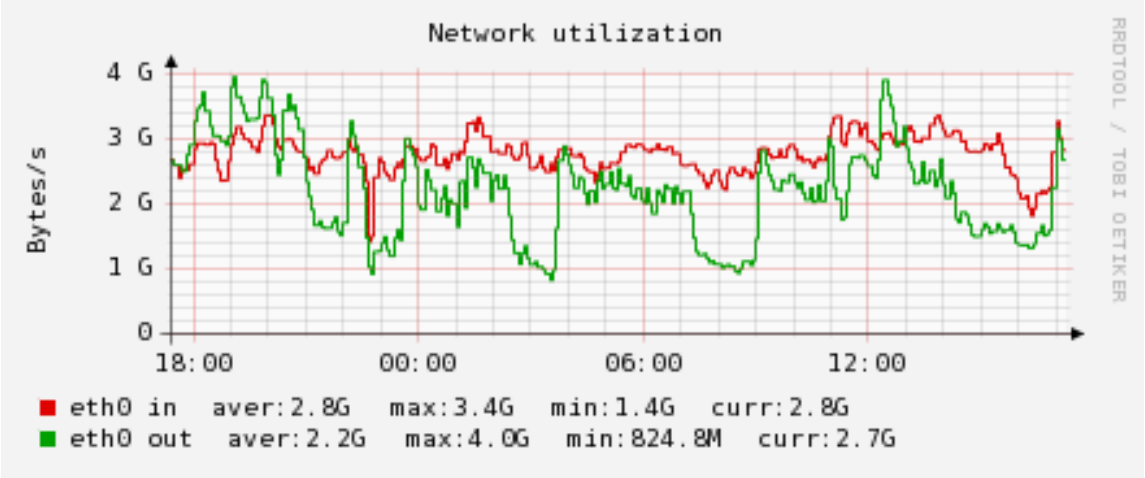


Tape servers network load, February 8th

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

Performance numbers (2)

- The disk cache handles much more
- For a given instance :
  - typical av rate in a busy day : 2.5/1.5 GB/s
  - peak rate on a busy day : 3.4/4.0 GB/s



CMS instance, April30th

Short term plans

- mostly consolidation of the current version of the software
- CASTOR2 will slowly move to a maintenance mode
- main items are
  - security deployment
  - improvements of admin tools
  - further optimization of the tape migrations and recalls, using improved policies

Longer term plans

- topic of the talk by Dirk Duellmann
- main lines
  - an architecture force is trying to define the architecture of the next step
  - includes data management experts from CASTOR and DPM development teams
  - first conclusions by summer time