Storage Resource Sharing with CASTOR

Olof Barring, Benjamin Couturier, Jean-Damien Durand,
Emil Knezo, **Sebastien Ponce**
(CERN)

Vitali Motyakov (IHEP)

ben.couturier@cern.ch

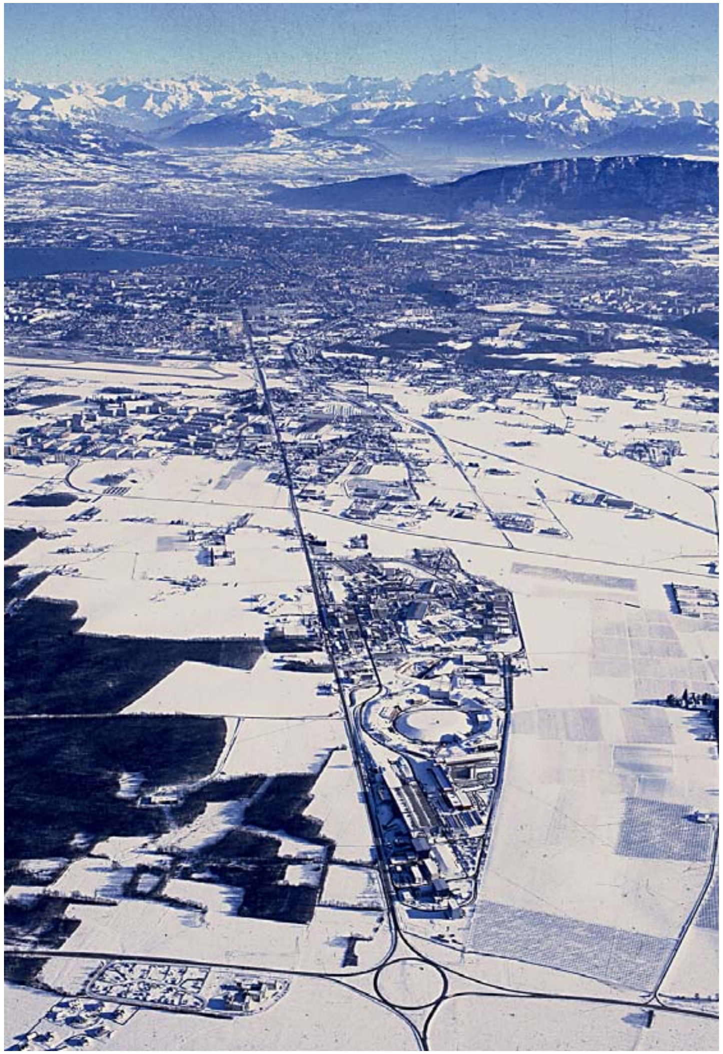
NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

IEEE COMPUTER SOCIETY

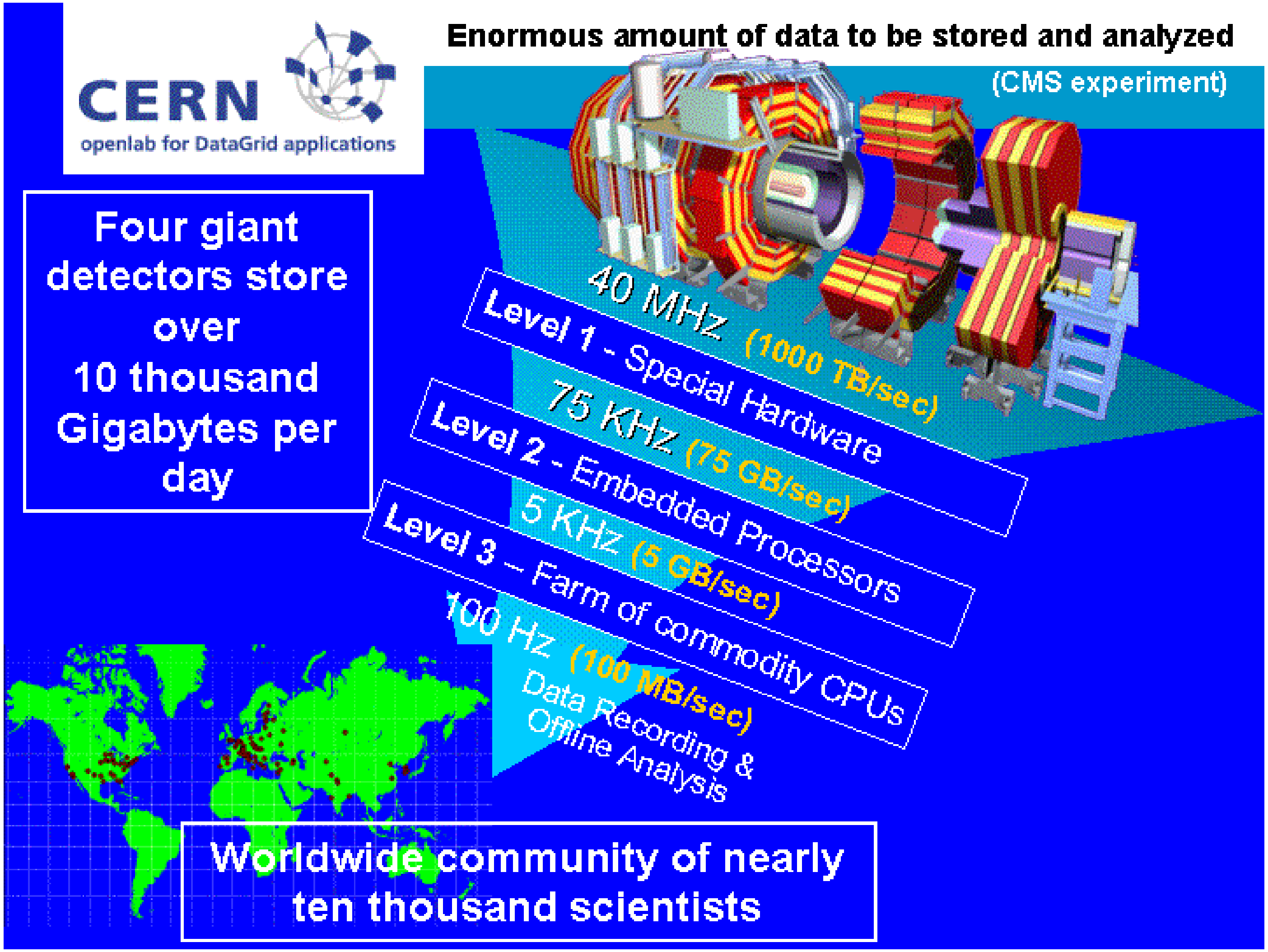
Introduction

- **CERN**
  European Laboratory for particle physics (Geneva)
  (Celebrating its 50th anniversary)

- 2007: The Large Hadron Collider (**LHC**)
  - 4 International Experiments: Alice, Atlas, CMS, LHCb
  - High energy proton or heavy ions collisions (energy up to 14 TeV for proton beams and 1150 TeV for lead ion beams).
  - Bigger amounts of data to store/analyze compared to previous experiments (Up to 4 GB/s, 10 Petabytes per year in 2008)



A LHC Experiment



CERN openlab for DataGrid applications

Enormous amount of data to be stored and analyzed

(CMS experiment)

Four giant detectors store over 10 thousand Gigabytes per day

40 MHz (1000 TB/sec)

Level 1 - Special Hardware

75 KHz (75 GB/sec)

Level 2 - Embedded Processors

5 KHz (5 GB/sec)

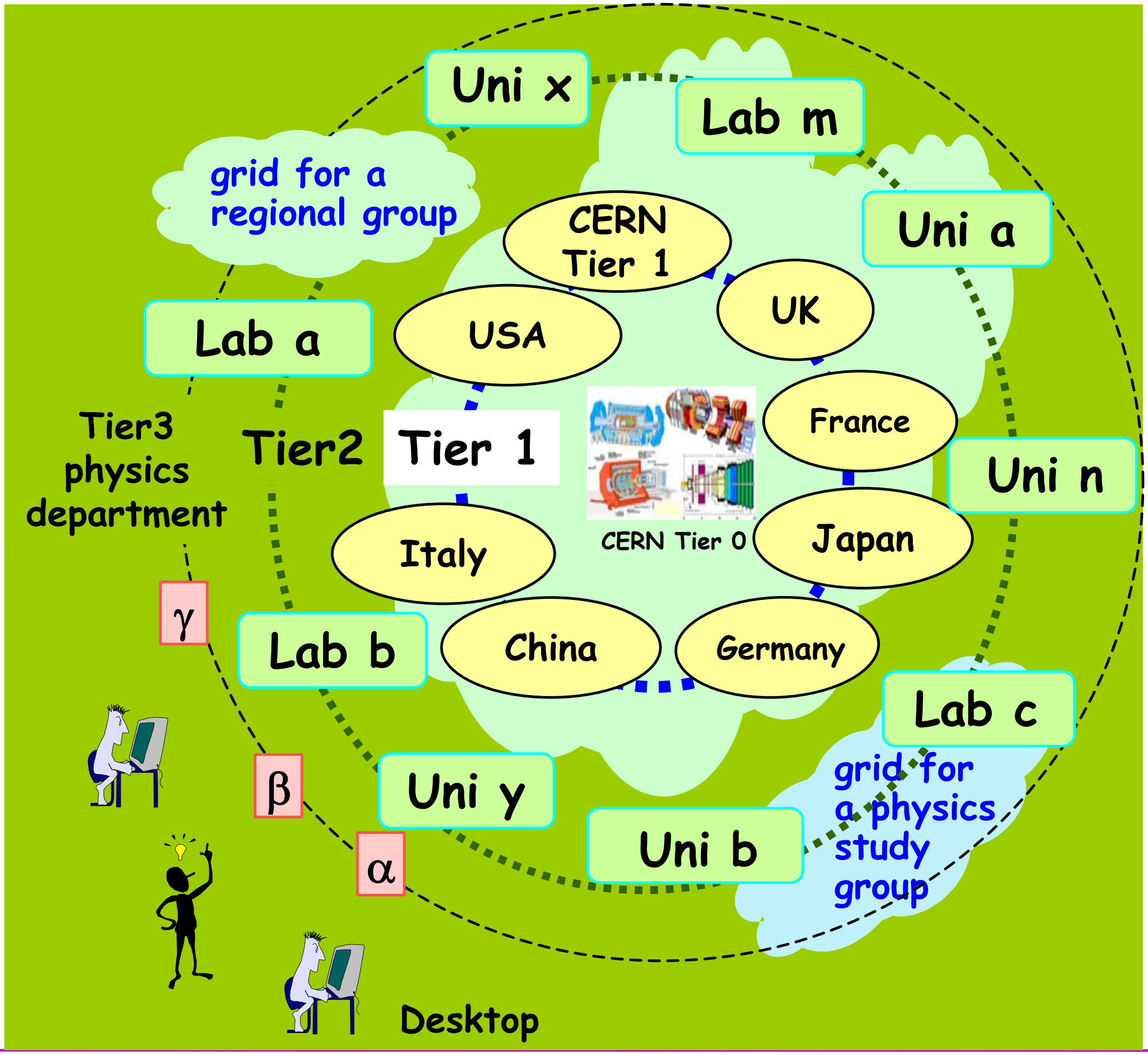
Level 3 – Farm of commodity CPUs

100 Hz (100 MB/sec)

Data Recording & Offline Analysis

Worldwide community of nearly ten thousand scientists

The LHC Computing Grid



CASTOR

- CERN Advanced STORage Manager
  - Hierarchical Storage Manager (HSM) used to store user and physics files
  - Manages the secondary and tertiary storage
- History
  - Development started in 1999 based on SHIFT, CERN's tape and disk management system since beginning of 1990s (SHIFT was awarded a 21st Century Achievement Award by Computerworld in 2001)
  - In production since the beginning of 2001
- http://cern.ch/castor/

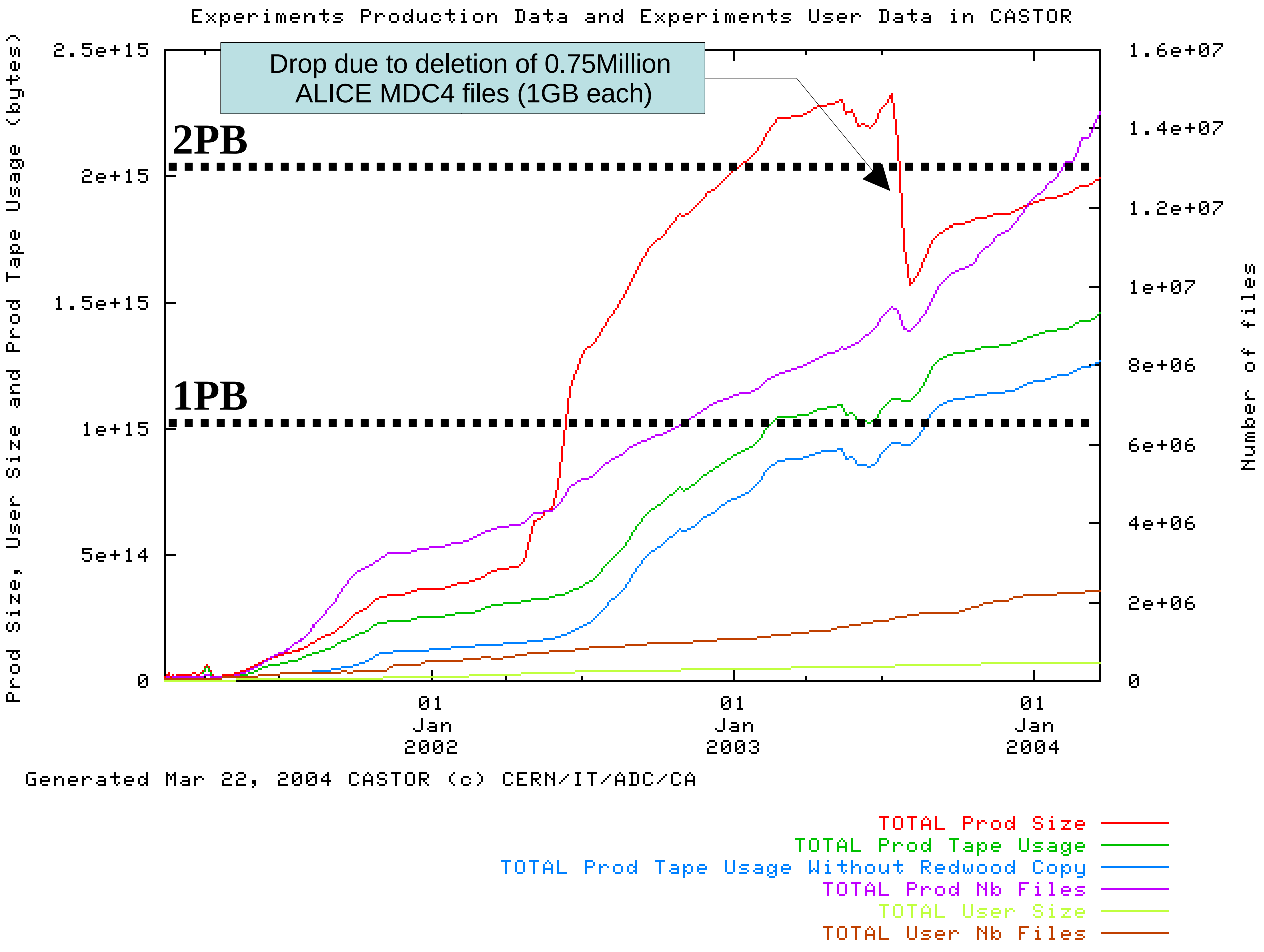
CASTOR deployment

- CASTOR teams at CERN
  - Dev team (5)
  - Operations team (4)

- HW Setup at CERN
  - Disk servers
    - ~ 370 disk servers
    - ~ 300 TB of staging pools
    - ~ 35 stagers (disk pool managers)
  - Tapes and Libraries
    - ~ 90 tapes drives (50 9940B)
    - 2 sets of 5 Powderhorn silos (2 x 27500 cartridges)
    - 1 Timberwolf (1 x 600 cartridges)
    - 1 L700 (1 x 288 cartridges)

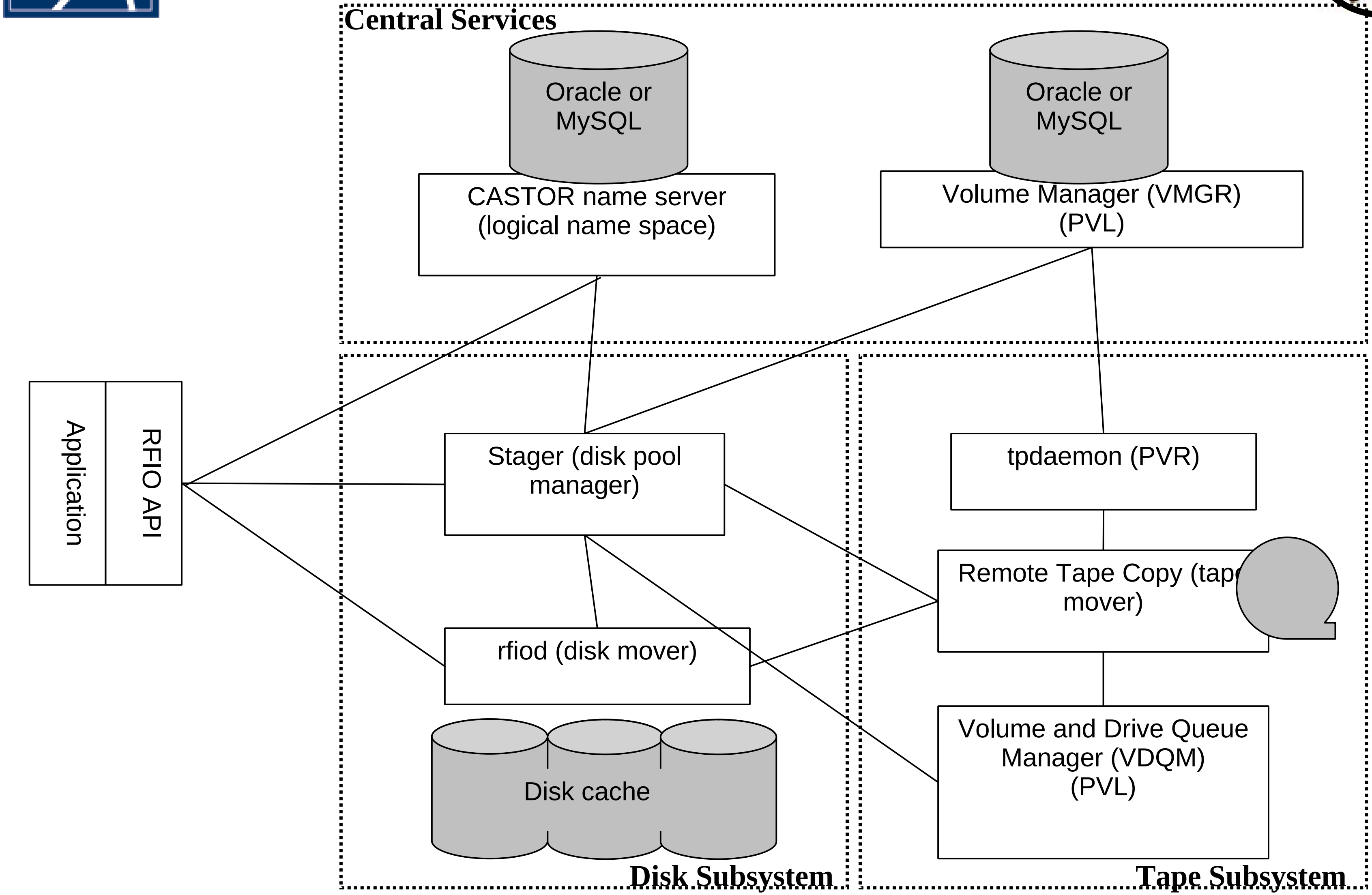
- Deployed in other HEP institutes
  - PIC Barcelona
  - CNAF Bologna
  - ...

CASTOR Data Evolution



Drop due to deletion of 0.75Million ALICE MDC4 files (1GB each)

CASTOR Architecture



Central Services
Oracle or MySQL
CASTOR name server (logical name space)
Oracle or MySQL
Volume Manager (VMGR) (PVL)
Application
RFIO API
Stager (disk pool manager)
rfiod (disk mover)
Disk cache
Disk Subsystem
tpdaemon (PVR)
Remote Tape Copy (tape mover)
Volume and Drive Queue Manager (VDQM) (PVL)
Tape Subsystem

Main Characteristics

- POSIX-like client interface
  - Use of RFIO in the HEP community
- Modular / Highly Distributed
  - A set of central servers
  - Disk subsystem
  - Tape subsystem
- Allows for tape resource sharing
- Grid Interfaces
  - GridFTP
  - Storage Resource Manager (V1.1) (Cooperation between CERN/FNAL/JLAB/LBNL; c.f. http://sdm.lbl.gov/srm-wg/)

Platform/Hardware Support

- Multiplatform support
  - Linux, Solaris, AIX, HP-UX, Digital UNIX, IRIX
  - The clients and some of the servers run on Windows NT/2K
- Supported drives
  - DLT/SDLT, LTO, IBM 3590, STK 9840, STK 9940A/B (and old drives already supported by SHIFT)
- Libraries
  - SCSI Libraries
  - ADIC Scalar, IBM 3494, IBM 3584, Odetics, Sony DMS24, STK Powderhorn (with ACSLS), STK L700 (with SCSI or ACSLS)

Requirements for LHC (1)

- CASTOR currently performs satisfactorily:
  - Alice Data Challenge: 300 MB/s for a week. 600 MB/s maintained for a half day.
  - High request load: 10s of thousands of requests per day per TB of disk cache.

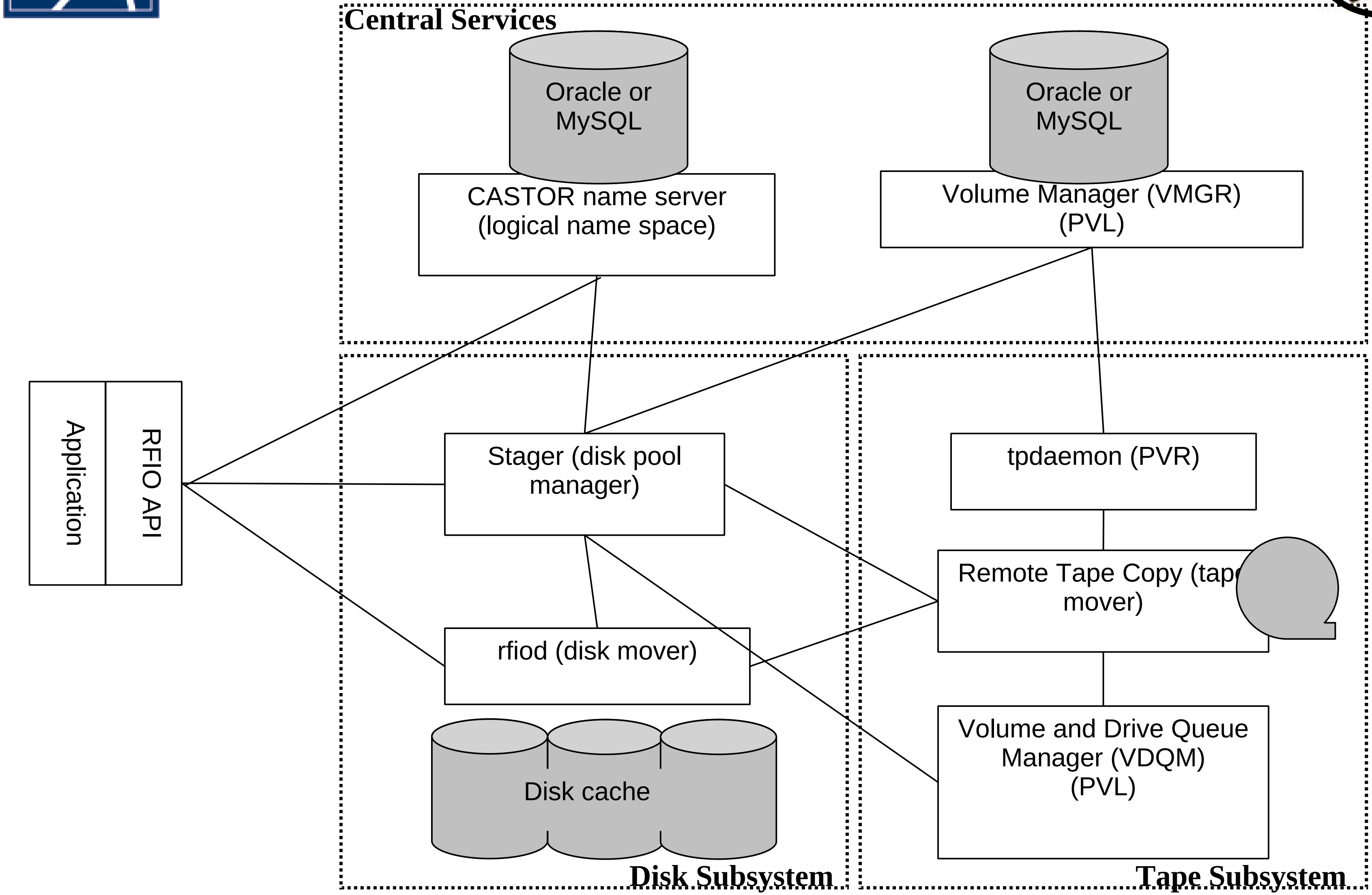
- However, when LHC starts in 2007
  - A single stager should scale up to 500/1000 requests per second
  - Expected system configuration
    - ~ 4PB of disk cache
    - 10 PB stored on tapes per year (peak rate of 4GB/s)
    - ~ 10000 disks, 150 tape drives

Requirements for LHC (2)

- Various types of Workload

	Access Pattern	Amount of Data Involved	Quality of Service ?	Frequency
TIER 0				
Data Recording	Sequential Write	Up to 1 GB/s for Alice... (2 GB/s total)	Necessary to avoid loosing data	Once
Reconstruction	Sequential RW	Complete data-set (up to 2 GB/s as well)	-	A few times per year
Data to Tier 1 centers	Sequential Read	Part of data set	-	Once per tier
TIER 1				
Analysis	Random	Random	-	Always runnning

Stager Central Role


Central Services
Oracle or MySQL
CASTOR name server (logical name space)
Oracle or MySQL
Volume Manager (VMGR) (PVL)
Application
RFIO API
Stager (disk pool manager)
rfiod (disk mover)
Disk cache
Disk Subsystem
tpdaemon (PVR)
Remote Tape Copy (tape mover)
Volume and Drive Queue Manager (VDQM) (PVL)
Tape Subsystem

Limitations of the current system

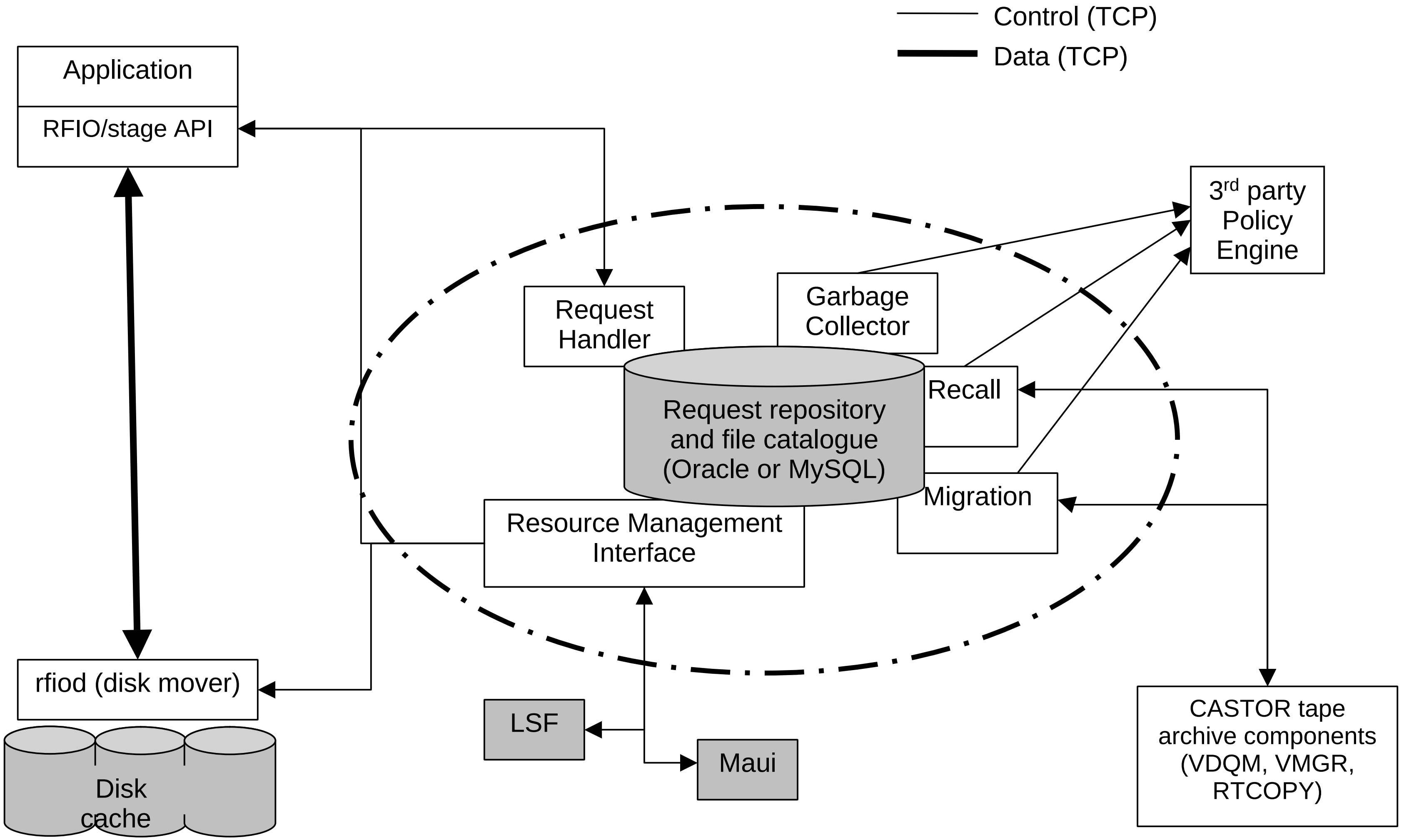
- The CASTOR stager has a crucial role: but the current version limits CASTOR performance
  - CASTOR stager catalogue limited in size (~100 000 files in cache)
  - The Disks resources have to be dedicated to each stager which leads to:
    - Sub-optimal use of disk resources
    - Difficult configuration management. It is difficult to switch disk resources quickly when the workload changes
  - Not very resilient to disk server errors
  - Current stager design does not scale
- Tape mover API not flexible enough to allow dynamic scheduling of disk access

Vision…

- With clusters of 100s of disk and tape servers, the automated storage management faces more and more the same problems as CPU clusters management
  - (Storage) Resource management
  - (Storage) Resource sharing
  - (Storage) Request scheduling
  - Configuration
  - Monitoring
- The stager is the main gateway to all resources managed by CASTOR

Vision: **Storage Resource Sharing Facility**

New CASTOR Stager Architecture



Database Centric Architecture

- Set of stateless multithreaded daemons accessing a DB
  - Request state stored in RDMS
  - Support for Oracle and MySQL
  - Allows for big stager catalogs with reasonable performance
  - Locking/transactions handled by the RDBMS
- Stager as scalable as the database it uses
  - Use of DB clustering solutions if necessary
- Easier administration
  - Standard backup procedures
  - less problems to restart in case of problems

Resource Scheduling

- Externalized Scheduling Interface
  - Currently developped for
    - MAUI (version 3.2.4) (http://supercluster.org/maui)
    - LSF 5.1 (Platform Computing – http://www.platform.com)
  - Leverage the advanced features from CPU scheduling:
    - Fair share
    - Advanced reservations
    - Load balancing
    - Accounting…

➔This will allow to:

- Share of all disk resources, with a fair share between the LHC experiments
- Exploit all resources at the maximum of their performance (avoid hotspots on disk servers…)
- follow the evolution of scheduling systems…

Other improvements

- Improved Request Handling
  - Request throttling possible
- Improved security
  - Strong authentication using GSI or Kerberos
- Modified tape mover interface
  - Allows for just-in-time scheduling of the disk resources when copying to/from tape
- Controlled Access to disk resources
  - All user access to disk resources will have to be scheduled though the stager, so as to control the IO on disk servers

Conclusion

- Current stager at CERN has shown its limitations
- New CASTOR stager proposal aims for
  - A pluggable framework for intelligent and policy controlled file access scheduling
  - Evolvable storage resource sharing facility framework rather than a total solution
  - File access request running/control and local resource allocation delegated to disk servers
- Currently In development…
  - Proof of concept/prototype implemented