

Response to the RTAG on Mass Storage (RTAG 5)

Rtag Description

Tier 0 Mass Storage System RTAG (RTAG 5)

Members and experts: Wisla Carena (chair and ALICE), Joel Closier (LHCb), Steve O'Neal (ATLAS), Harry Renshall IT/DS, Tony Wildish (CMS)

**Focusing on the current choice of an HSM : CASTOR
as a valid prototype/candidate for the LHC era**

**Focusing on the ALICE requirements as the other experiments have not yet
defined in detail their mass storage system requirements
(still true in the LCG High-Level planning document)**

<http://lhcgird.web.cern.ch/LHCgrid/sc2/RTAG5/>

CASTOR Overview

- **CERN development since January 1999**
- **A hierarchical storage manager handles the secondary storage (disk cache), the unique namespace and the migration/recall to tertiary storage (magnetic tapes)**
- **Used now by all experiments**
- **Contains currently 6.2 million files which corresponds to about one PB of data**

<http://it-div-ds.web.cern.ch/it-div-ds/HSM/CASTOR/Welcome.html>

CASTOR software

Free software package

<http://it-div-ds.web.cern.ch/it-div-ds/HSM/CASTOR/DIST/CERN/>

Used by about 10 outside Institutes

Requires adequate local system admin support and expertise

CERN support is on 'Best Effort' basis

Distribution conditions :

<http://it-div-ds.web.cern.ch/it-div-ds/HSM/CASTOR/DIST/CONDITIONS/>

RTAG details (1)

Questions and observations :

“The following two items have emerged as top priority for LHC experiments:

à Performance especially for Data Challenges, where the aggregate performance did not scale with the number of disk servers,

and for production

à Reliability improvement and adequate monitoring tools “

“During the ADC III, in 2001, the planned milestone of 100 MB/s sustained to tape has not been achieved. The two main causes were the limited performance of the then available generation of IDE-based disk servers and the inefficient load balancing performed by CASTOR amongst the disk servers. “

RTAG details (2)

Answers :

This years ALICE MDC IV has started in April and shown already some interesting results.

The full MDC is scheduled for October as only then the new Generation of Tape drives (9940B) will be available.

- **Load balancing improved from last year**
- **Mixing hardware and software (CASTOR + Farm), Disk server scaling works with and without CASTOR (no migration to tape yet)**
- **CASTOR team is working heavily on reliability issues, much more stable now**
- **Light weight file creation protocol implementation**
- **Monitoring of some CASTOR aspects à**

<http://it-div-ds.web.cern.ch/it-div-ds/HSM/CASTORMONITOR/>

Some benchmark figures

- DATE + Aliroot + CASTOR using 20 disk servers and 50 producers

1 GB file ==> **350 MB/s into the disk servers (CPU load 40 %)**

Extra streams added from outside (Bernd) did not change the performance. The total data rate for a disk server was 35 MB/s (write only) or 20 MB/s write + 20 MB/s read.

==> disk servers are not a bottleneck.

- Changing file size from 1 GB to 1.8 GB did not change the performance.

- Test of file creation

140 files per second on local EXT3

40 files per second using RFIO

5.5 files per second for CASTOR files

- Using 10 disk servers instead of 20

The performance dropped from 350 MB/s to 220 MB/s (the performance per disk server is a bit higher).

- Test reading + writing with RFIO

- 10 disk servers + 30 CPU servers (on Gigabit)

1 write + 1 read per file system

==> 260 MB/s write + 130 MB/s read

- 20 disk servers + 60 CPU servers (on Gigabit)

==> 350 MB/s write + 300 MB/s read or 460 MB/s write only

- **33 disk servers + 66 CPU servers (on Gigabit)**

==> 500 MB/s write + 500 MB/s read

RTAG details (3)

Questions and observations :

“We are concerned by the long-term evolution and the manpower devoted to the development and the support of CASTOR. “

Castor Manpower

Has increased from 2 to 7 people (2001 à 2002)

- **7 people working in the CASTOR team (4 staff, 1 fellow, 1 technical students, 1 coorporant)**
- **Distribution of work à 3.3 FTE deployment
1.7 FTE development
(development == focus on stability, scalability and performance)**

(counting tech. Students as 0.5)

including now one LCG person

**à Deployment of GRID tools and interfaces related to the HSM system
(e.g. GridFTP, scp, WP5, WACDR server)**

Castor Workplan

Some extract , short term for the next 12 month:

1. **Implementation of 'fair-share' tape drive scheduling (end 2002)**
2. **Improved statistics**
3. **Redesign of the CASTOR stager (Q2 2003)**
4. **Improvement of robustness**
5. **Support for files > 2GB (Q4 2002)**

One FTE dedicated for the month of October for the ALICE MDC IV full HSM test aimed at 200 MB/s performance (300 MB/s peak)

http://it-div-ds.web.cern.ch/it-div-ds/HSM/CASTOR/DOCUMENTATION/USERGUIDE/castor_plans_2002.html

Castor Workplan (2)

à **Optimisation of hardware and software components for the scheduled mock data challenges**

ALICE	2003	300 MB/s into CASTOR
	2004	450 MB/s
	2005	750 MB/s

(no requirements from the other experiments yet)

à **The redesign of the CASTOR stager module will include the necessary feature for the long term Grid SRM architecture.
Full production version in 2004**

CASTOR long term plans

Heavily depends on :

à Computing models from the experiments

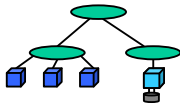
à Technology and market developments (e.g. Disks versus tapes, pricing)

à Data Challenges in the different domains :

CDR, reprocessing, MC production, Analysis

Computing model of the Experiments

Benchmark and performance cluster
(current architecture and hardware)



Data Challenges
Experiment specific IT base figures

Benchmark and analysis framework

Components
LINUX, CASTOR, AFS, LSF,
EIDE disk servers, Ethernet, etc.

Architecture validation

PASTA investigation

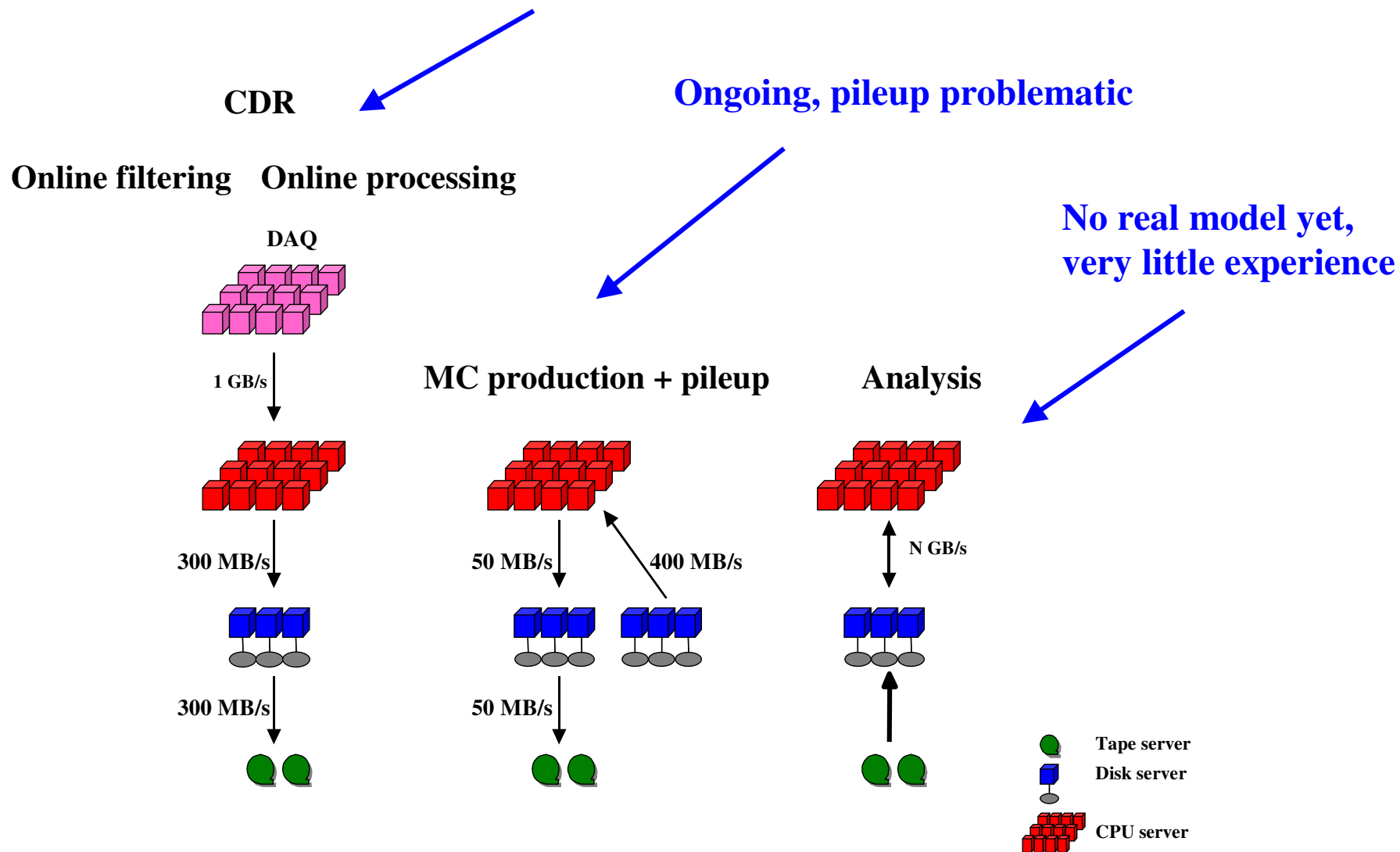
R&D activities (background)
à iSCSI, SAN, Infiniband
à Cluster technologies

**Data challenge views concerning the HSM system
are currently dominated by the ALICE requirements for online filtering and CDR**

CMS, ATLAS and LHCb data challenges are focused on MC production.

CMS has moved its focus lately also on the analysis framework and requirements

Current Focus



Mixture of hardware (disk server) and software (CASTOR,OBJ,POOL) optimization

“We recommend to get at least a preliminary table similar to Table 3 for all the experiments, as soon as their Computing Model is available, in order to be able to calculate the final requirements for T0 at CERN. “

Parameter	Unit	Conversion factor	ALICE	
			p-p	Pb-Pb
Nb of assumed Tier1 nodes at CERN			4	
Event recording rate	Hz	1	100	50
RAW Events size	MB	1.00E+06	0.2	2.5
REC/ESD Events size	MB	1.00E+06	0.02	2.5
AOD Events size	kB	1.00E+03	5	250
TAG Events size	kB	1.00E+03	1	10
Running time per year	Mseconds	1.00E+06	10	1
Computing time per year	Mseconds	1.00E+06	20.3	20.3
Events/year	Giga	1.00E+09	1.00	0.05
Storage for raw data	PB	1.00E+15	0.2	1.3
Storage for real (raw+rec+AOD+TAG) data	PB	1.00E+15	0.3	1.5
RAW SIM Events size	MB	1.00E+06	0.4	600
REC/ESD SIM Events size	MB	1.00E+06	0.02	5
Events SIM/year	Giga	1.00E+09	0.1	0.0001
Number of rec passes	Nb		2	2
Storage for simul. data	TB	1.00E+12	42.0	60.5
Storage for calibration	TB	1.00E+12	0.0	0.0
Tape storage at CERN T0+T1				
Tape storage at each Tier1 (Avg)	PB	1.00E+15	0.2	
Σ Tape storage /year		1.00E+15	2.8	
Tape bandwidth at CERN T0				
Tape bw raw data at CERN T0		1.00E+06	20	1250
Tape bw induced by bn data at T0		1.00E+06	191	
Tape bw induced by pp data at T0		1.00E+06	33	
Total bw at CERN T0		1.00E+06	244	1250
Tape bandwidth at each Tier1 (Avg)	MB/s	1.00E+06	3	14
Disk storage at CERN T0+T1				
Disk storage at each Tier1 (Avg)	PB	1.00E+15	0.2	
Σ Disk storage		1.00E+15	1.1	

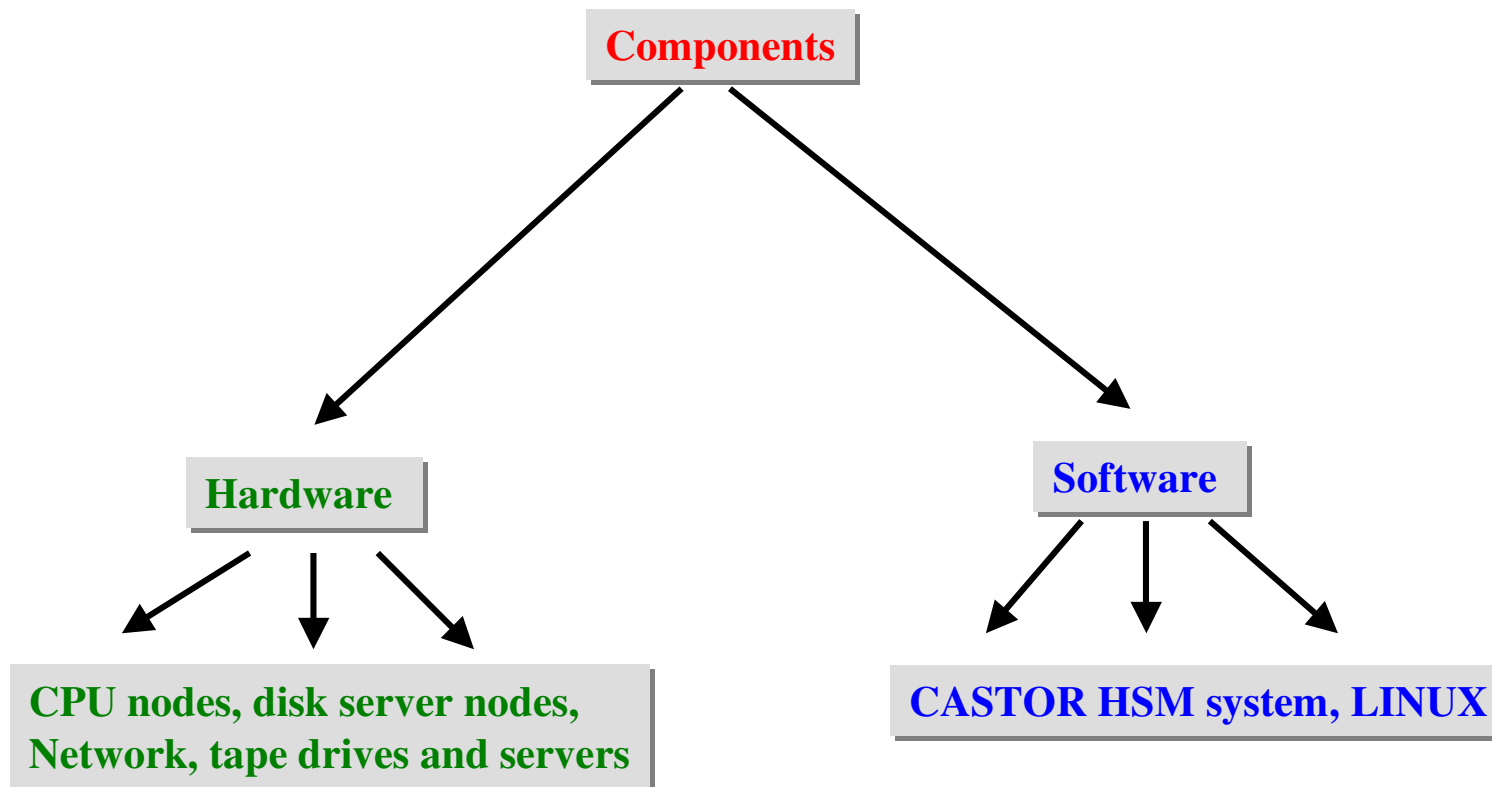
Table 3 – Update of LHC Computing Review Table

à Still needs to be provided by CMS, ATLAS and LHCb

“Experiments acknowledge the request from the PEB to have the revised planning of the Data Challenges available in May 2002, to be mapped on short-term LCG activity, CERN services, experiment-organised resources and Grid testbeds. Up to now, only ALICE has derived mass storage requirements from the plans for LCG Phase 1. “

à No changes yet

CASTOR Tests



Necessity :

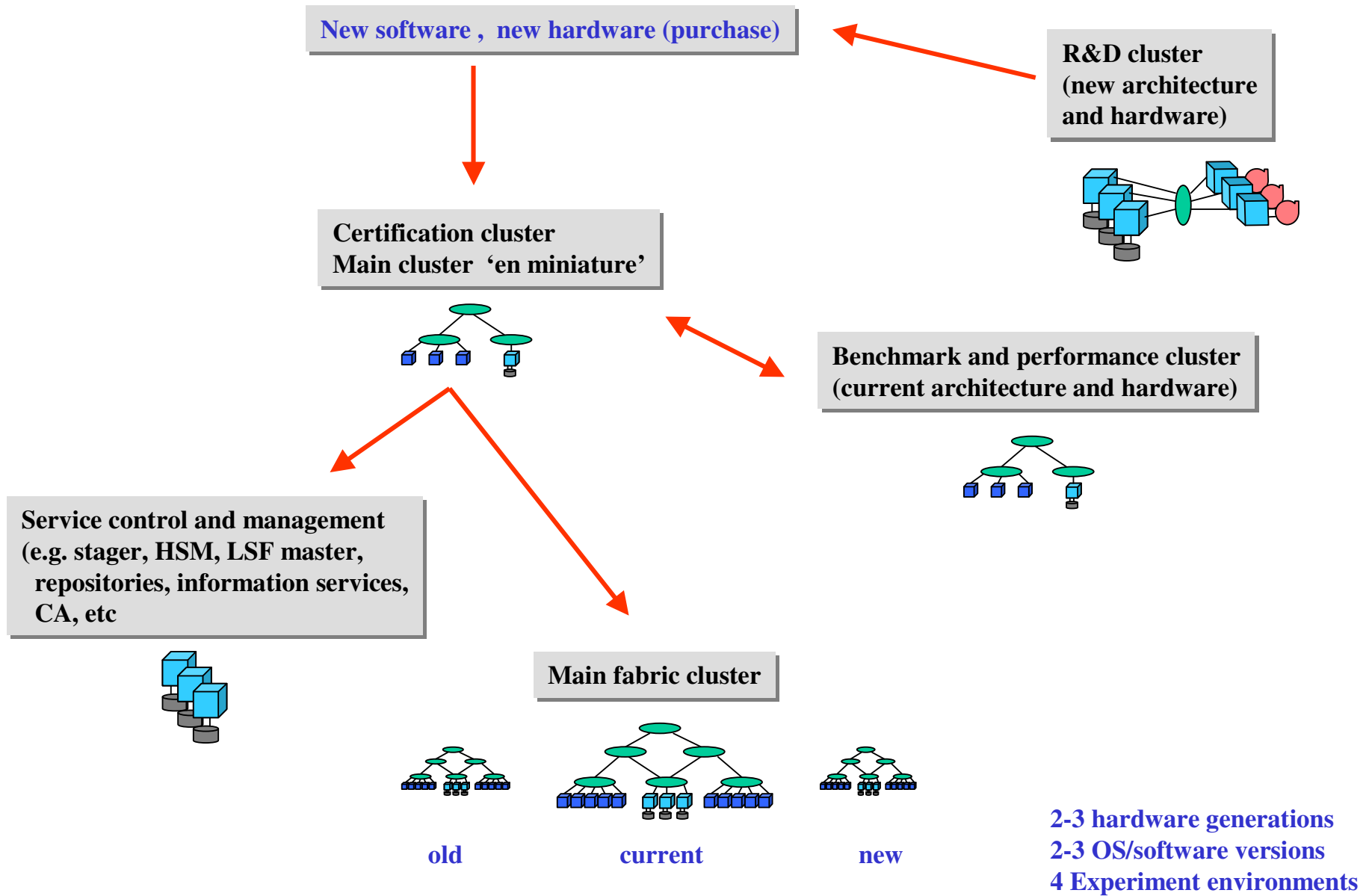
Dedicate 100 cpu nodes and 36 disk servers and 20 tape drives

Special computing data challenges :

à online test, filtering level x for all four experiments

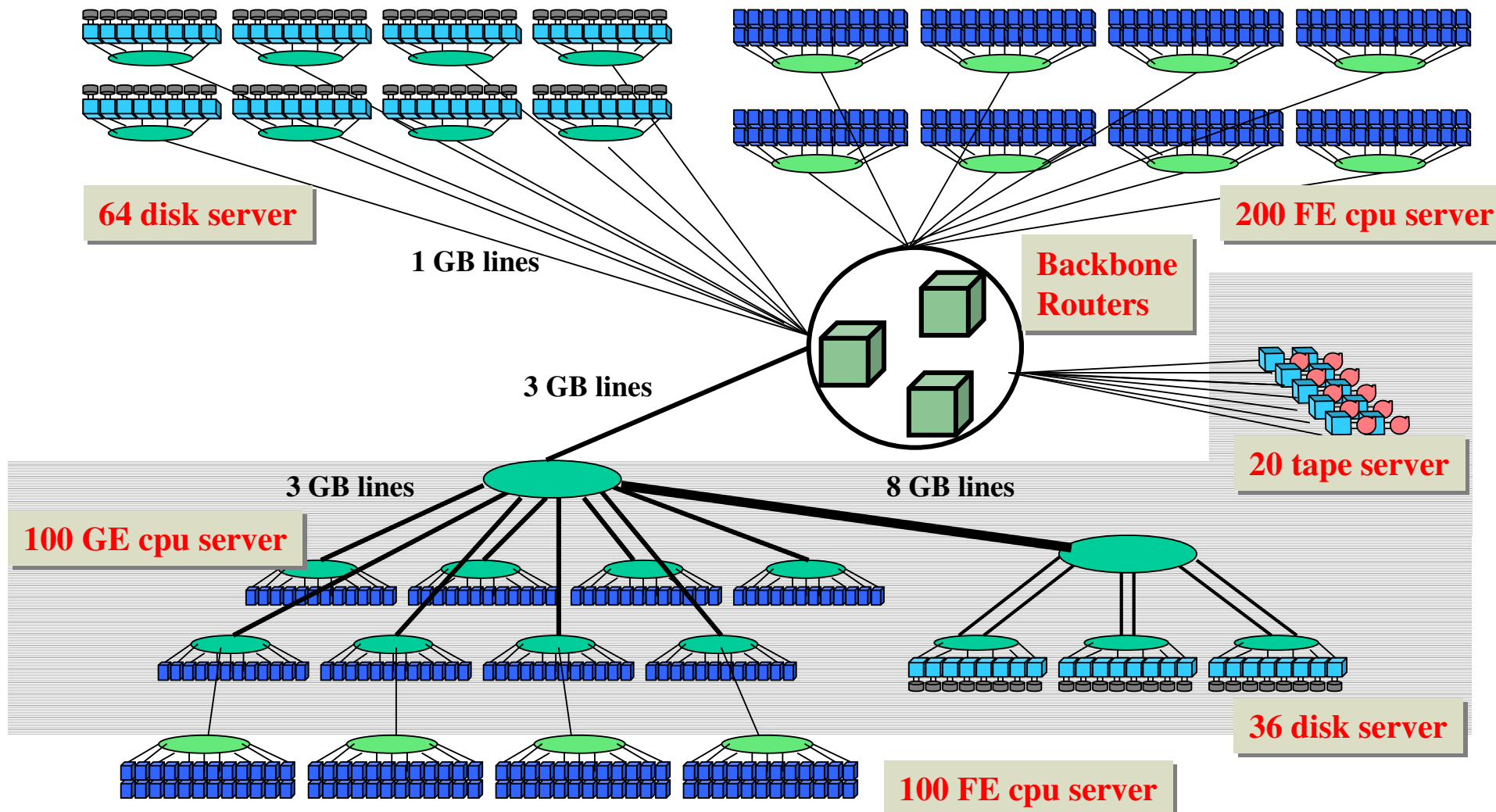
**à IT architecture verification
hardware configuration , stability and performance, HSM**

**à network tests
10 Gb equipment (openlab, Enterasys)**



LCG Testbed Structure

100 cpu servers on GE, 300 on FE, 100 disk servers on GE (~50TB), 10 tape server on GE



Summary

- **CASTOR personnel so far sufficient**
- **Work plan exists, conforms with the requests from the RTAG**
- **Long term planning depends heavily on the computing models of the experiments and their data challenge evaluation**
- **More details after the ALICE full MDC IV milestone in October**
- **Analysis is a key area important for CASTOR but yet really understood**
- **Need dedicated hardware permanently for special DCs and IT architecture verification**
- **Influence of the persistency model needs to be understood**