


The ALICE storage system



Adriana Telesca

PH/AID CERN

ACEOLE 6 months meeting

April 2-3 2009

Outline

- 1 Introduction
 - The ALICE experiment
- 2 Hardware and system software performance
 - Hardware performance test
 - Performance tests with single stream
 - Performance tests with multiple streams
- 3 Hardware and application software performance
 - Performance with DATE
- 4 Conclusions
- 5 Marie Curie ITN ACEOLE project
 - Training programme

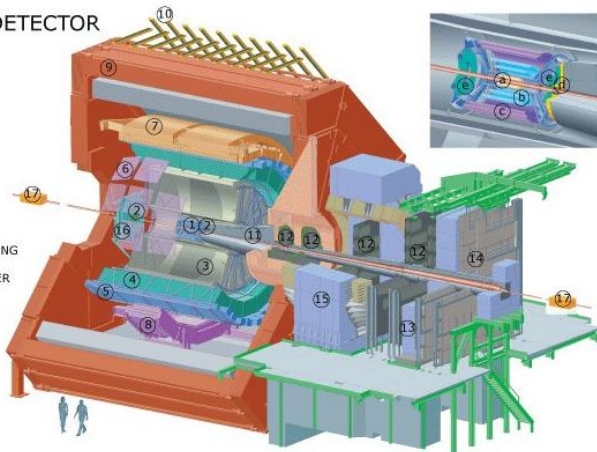
Outline

- 1 **Introduction**
 - The ALICE experiment
- 2 Hardware and system software performance
 - Hardware performance test
 - Performance tests with single stream
 - Performance tests with multiple streams
- 3 Hardware and application software performance
 - Performance with DATE
- 4 Conclusions
- 5 Marie Curie ITN ACEOLE project
 - Training programme

ALICE detectors

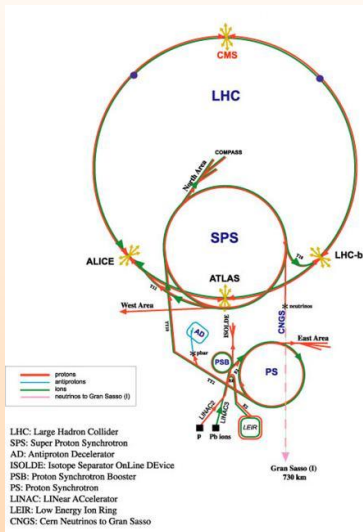
THE ALICE DETECTOR

1. ITS
2. FMD , T0, V0
3. TPC
4. TRD
5. TOF
6. HMPID
7. EMCAL
8. PHOS CPV
9. MAGNET
10. ACORDE
11. ABSORBER
12. MUON TRACKING
13. MUON WALL
14. MUON TRIGGER
15. DIPOLE
16. PMD
17. ZDC

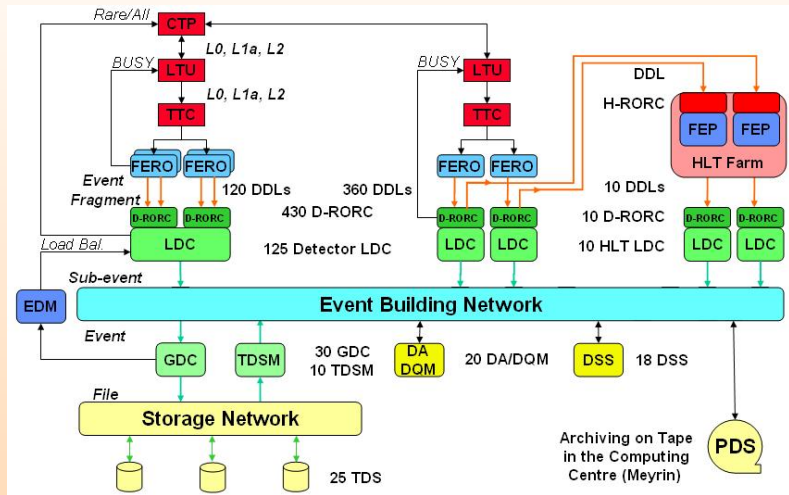


- a. ITS SPD Pixel
- b. ITS SDD Drift
- c. ITS SSD Strip
- d. V0 and T0
- e. FMD

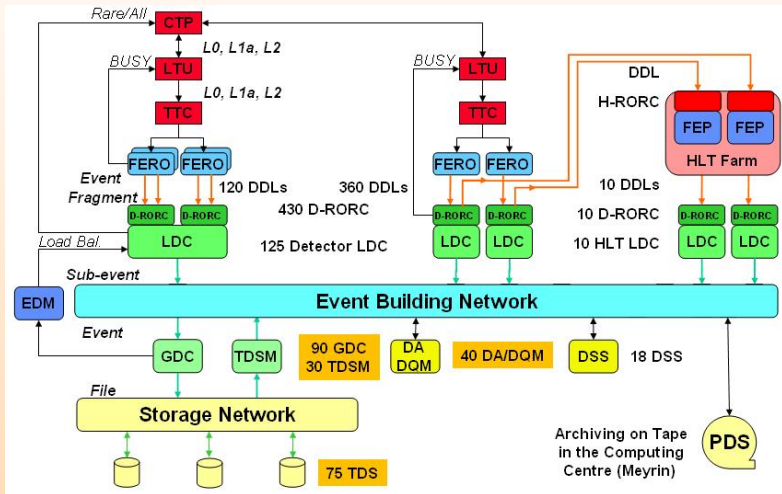
ALICE location



Trigger - DAQ - HLT '08



Trigger - DAQ - HLT '09



Performance requirements

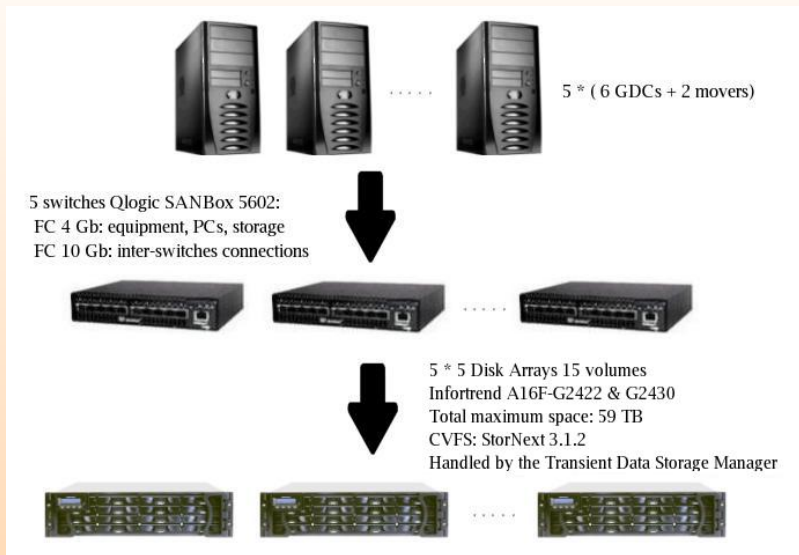
The DAQ system has the following requirements :

- an aggregate event building bandwidth of up to 2,5 GBytes/s
- a storage capability of up to 1,25 GBytes/s

which result in a total of more than 1 PBytes of data every year.

This makes the performance of the mass storage devices a dominant factor for the overall system behavior and throughput.

Current Storage System



Deployment of the ALICE Storage Area Network in '09

Two QLogic SANbox 9000 Stackable
FC Switches.

Each with a maximum of 8 I/O blades.

Each blade with:

- 16 * FC 8/4/2 Gb ports
- 8 * FC 10 Gb ports



Outline

- 1 Introduction
 - The ALICE experiment
- 2 Hardware and system software performance**
 - Hardware performance test
 - Performance tests with single stream
 - Performance tests with multiple streams
- 3 Hardware and application software performance
 - Performance with DATE
- 4 Conclusions
- 5 Marie Curie ITN ACEOLE project
 - Training programme

Test Storage parameters

Storage configuration parameters which can impact the system performance are:

- Block size
- File size
- RAID configuration
- Storage array configuration

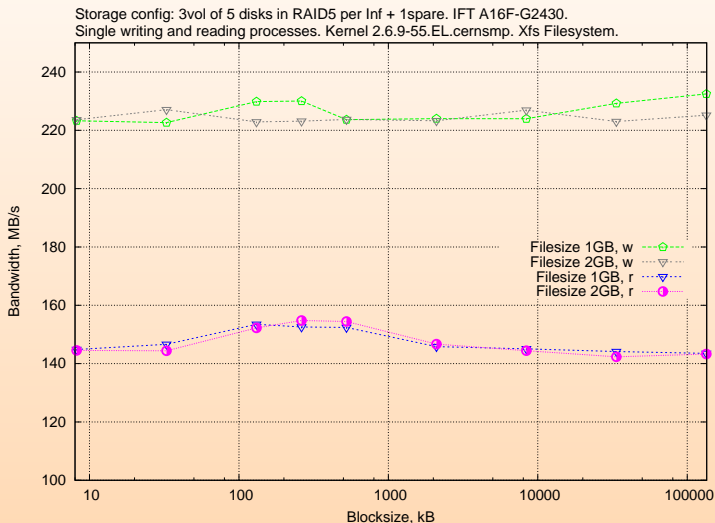
Hardware Test procedure

The test software is a standalone client called `lmdd` which:

- copies a specified input file filled by random data to a specified output;
- can be run simultaneously with other `lmdds` to perform parallel writing/reading streams;
- prints out the timing statistics.

Block size and File size

Rate performance according to different file and block sizes. Xfs Unix file system.



Storage setup parameters

Storage configuration parameters chosen are:

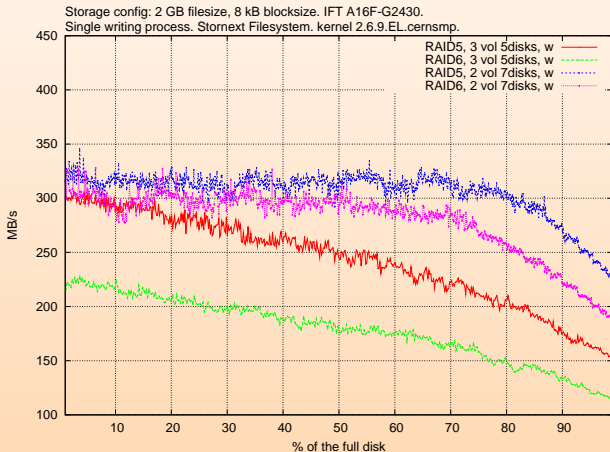
- Block size 8 KB
- File size 2 GB
- StorNext cluster file system

Storage configuration parameters tested are:

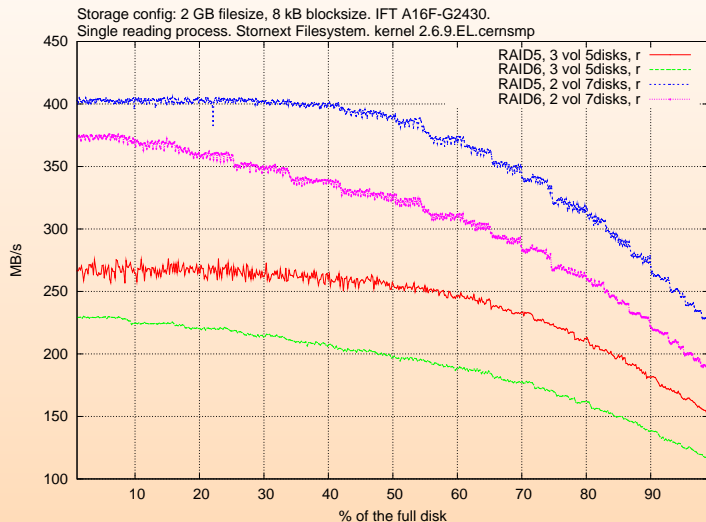
- RAID configuration :
 - RAID 5
 - RAID 6
- Storage array configuration (16 disks):
 - 3 volumes of 5 disks + 1 spare
 - 2 volumes of 7 disks + 2 spares

Single writing

Storage performance tested according to the volumes/RAID configuration by performing single writing and reading operations from one GDC to one disk volume.



Single reading



Relative performance

Considering RAID 5 with 5 disks as a reference, the relative performance is:

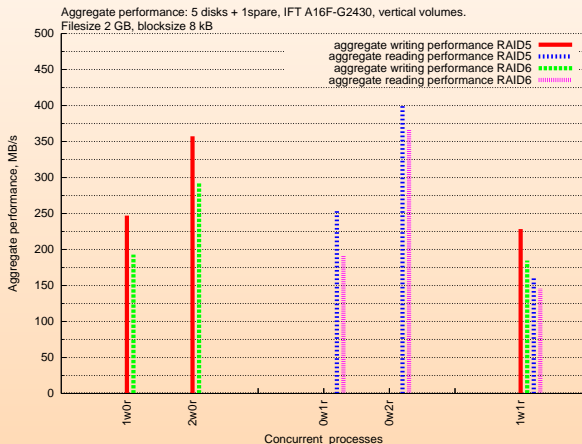
	5 disks	RAID set	7 disks	RAID set
	R5	R6	R5	R6
Writing	100	75	125	115
Reading	100	80	150	130

Multiple concurrent operations

- Concurrent activities on the same volume produce disastrous results
- The StorNext cluster file system allows to define an "affinity" which associates a file system folder to a logical unit. In this way we can address concurrent streams to different volumes
- An investigation on the coexistence of more streams on the same disk array is needed

Multiple concurrent writings and readings

Storage performance tested according to RAID configuration by performing concurrent writing and reading operations from two GDCs to two disk volumes.



Outline

- 1 Introduction
 - The ALICE experiment
- 2 Hardware and system software performance
 - Hardware performance test
 - Performance tests with single stream
 - Performance tests with multiple streams
- 3 Hardware and application software performance**
 - Performance with DATE
- 4 Conclusions
- 5 Marie Curie ITN ACEOLE project
 - Training programme

DATE

DATE (ALICE Data Acquisition and Test Environment) is the software framework of the ALICE DAQ.

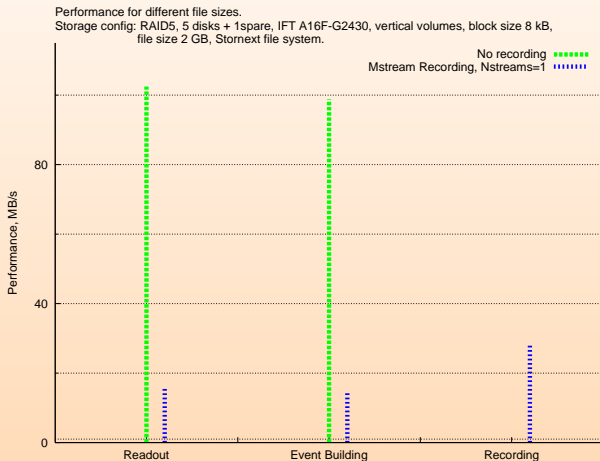
The DATE system performs different functions:

- Readout
- Event building
- Data recording

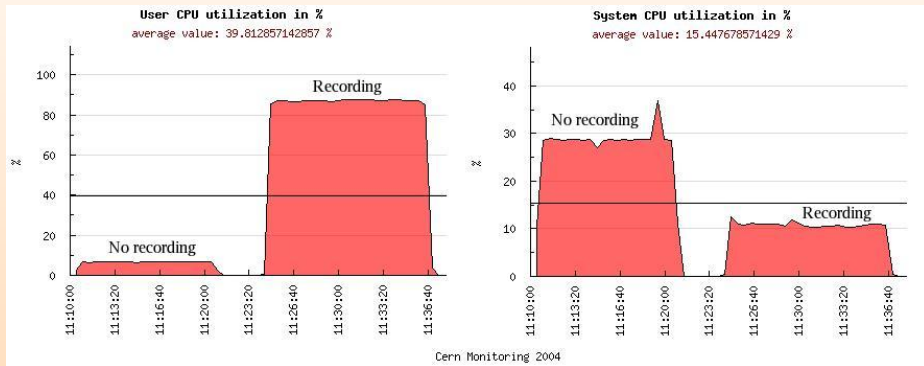
It has been decided relatively late to format the information in the reconstruction-ready format ROOT.

Readout, event building and recording performance

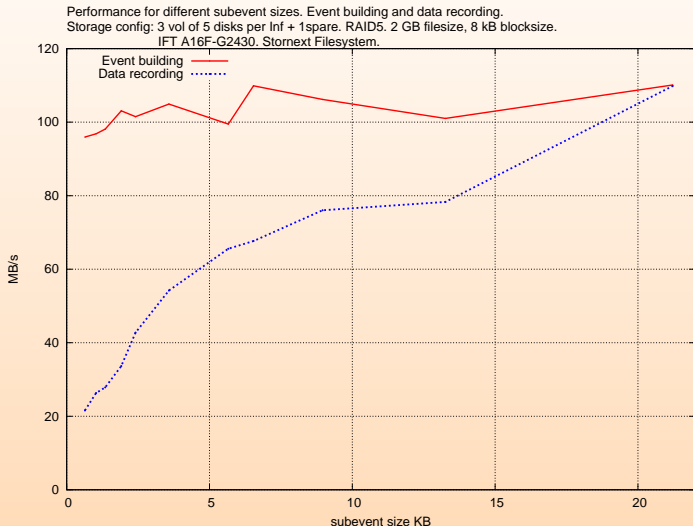
For some runs we experienced a drop in performance between "No recording" and "Data recording" run modalities.



User and system CPU utilization



Performance for different subevent sizes



Outline

- 1 Introduction
 - The ALICE experiment
- 2 Hardware and system software performance
 - Hardware performance test
 - Performance tests with single stream
 - Performance tests with multiple streams
- 3 Hardware and application software performance
 - Performance with DATE
- 4 **Conclusions**
- 5 Marie Curie ITN ACEOLE project
 - Training programme

Conclusions

- **Storage hardware and system software** provide adequate performance for ALICE;
- **The storage hardware** provides different performance for different configurations. RAID 5 maximizes the performance but, if we choose RAID 6, we have a gain in reliability with a small loss in performance;
- The user CPU utilization impacts **the software and hardware** performance. We can obtain performance compatible to the ALICE needs for subevent sizes bigger than 20 kB.
- More investigation is needed to understand the loss in performance below 20 kB.

Outline

- 1 Introduction
 - The ALICE experiment
- 2 Hardware and system software performance
 - Hardware performance test
 - Performance tests with single stream
 - Performance tests with multiple streams
- 3 Hardware and application software performance
 - Performance with DATE
- 4 Conclusions
- 5 Marie Curie ITN ACEOLE project
 - Training programme

The ACEOLE program gives me the possibility to invest on my education.
The trainings that I attended up to now are:

- StorNext File System training, Munich, 13-15 January 2009
- Qlogic Fibre Channel Specialist training and certification, Wokingham (UK), 17-19 February 2009
- Advanced StorNext File System training, Munich, 10-11 March 2009
- General and Professional English Course, CERN, from 27 February 2009 to end June 2009
- IEEE NPSS Real Time Conference 2009, Beijing, 10-15 May 2009.
Accepted abstract for oral presentation. Abstract title: "THE ALICE STORAGE SYSTEM: an Analysis of the Impact on the Performance of the Configuration Parameters and of the Load of Concurrent Streams"