LHCC Review
Nov. 22nd-23rd, 2004

LCG Fabric Area
Wide area networking

Rapporteur:
Volker Guelzow
DESY

Nov. 25th, 2004

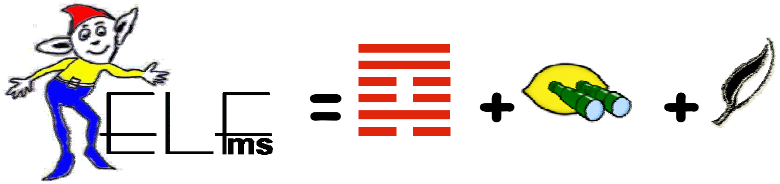
Reviewers Comment:
very well on track

Refurbishment of the left side of the Computer Center has started

New structure on the already refurbished right side

During the period May-August more than 800 nodes were moved with minor service interruptions (AFS, NICE, MAIL, WEB, CASTOR server, etc.)
Very man-power intensive work , tedious and complicated scheduling.

Tools



- ◆ ELFms is deployed in production at CERN
  - Stabilized results from 3-year developments within EDG and LCG
  - Consistent full-lifecycle management and high automation level
  - Providing real added-on value for day-to-day operations
- ◆ Quattor and LEMON are generic software
  - Other projects and sites getting involved
- ◆ Site-specific workflows and “glue scripts” can be put on top for smooth integration with existing fabric environment
- ◆Reviewers comment: A well defined set of important tools is choosen for fabric management. The tools will improve in a evolutionary way, deployment outside LHCtakes place.

CPU server

- **price penalties for 1U, 2U and blade servers (between 10% and 100%)**

- **The technology trend moves away from GHz to multi-core processors**
- analysts see problem with re-programming applications (multithreading)

Reviewers Comment: This, together with 64 bit processors will cause significant effort for the experiments

- **Will start at the end of November an activity (Exp. + IT)in the area application performance**
- **To evaluate the effects on the performance of**
  1. **different architectures (INTEL, AMD, PowerPC)**
  2. **different compilers (gcc, INTEL, IBM)**
  3. **compiler options**

- **Total Cost of Ownership and farm architecture in mind**

Reviewers Comment: Considered as a small but very useful effort

Disk server

we will try to buy ~ 500 TB disk space in 2005

- **need more experience with much more disk space**
- **tuning of the new Castor system**
- **getting the load off the tape system**
- **test the new purchasing procedures**

- **Reviewers Comment:**
  agree, capacity is not a problem,
  the random access speed is an issue.
  Large test setups are mandatory

Tape servers, drives, robots

Boundary conditions for the choices of the next tape system for LHC running:

- **Only three choices (linear technology) : IBM, STK, LTO Consortium**
- **The technology changes about every 5 years, with 2 generations within the 5 years (double density and double speed for version B, same cartridge)**
- **The expected lifetime of a drive type is about 4 years, thus copying of data has to start at the beginning of the 4th year**
- **IBM and STK or not supporting each others drives in their silos**

- **Reviewers Comment:**

Those technical issues can only be sorted out by test setups, which are planned. The effort moving from one media to the other is very large.

Tape storage

combination of problems example : small files + randomness of access
possible solutions :

- **concatenation of files on application or MSS level**
- **extra layer of disk cache, Vendor or ‘home-made’**
- **hierarchy of fast and slow access tape drives**
- **very large amounts of disk space**
- **……….**

- **Analysis of parameters determine datamanagement**

Reviewers Comment:
The design of the data management system has to be developed in strong interaction with the experiments. Therefore the Computing models of the Experiments are urgently needed. The system should be flexible, so that in case of eg Analysis comes through the backdoor, this can be handled.

Planned data challenges

- **Dec04 - Service Challenge 1 complete**
  mass store-mass store, CERN+3 sites, 500 MB/sec between sites, 2 weeks sustained
- **Mar05 - Service Challenge 2 complete**
  reliable file transfer service, mass store-mass store, CERN+5 sites, 500 MB/sec between sites, 1 month sustained
- **Jul05 - Service Challenge 3 complete**
  mock acquisition - reconstruction - recording – distribution, CERN + 5 sites, 300 MB/sec., sustained 1 month
- **Nov05 – ATLAS or CMS Tier-0/1 50% storage & distribution challenge complete**
  300 MB/sec, 5 Tier-1s (This is the experiment validation of Service Challenge 3)

  Tier-0 data recording at 750 MB/sec
  → ALICE data storage challenge VII completed

- continuous data challenge mode in 2005
- use the high-throughput cluster for continues tests, expand the disk space
- start to use the new network backbone as soon as possible

Reviewers comment: Data-Service-Challenges are considered as extremely important test scenarios

Linux

Strategy :

1. Use Scientific Linux for the bulk installations, Farms and desktops
2. Buy licenses for the RedHat Enterprise version for special nodes (Oracle) ~100
3. Support contract with Redhat for 3rd level problems
   contract is in place since July 2004, ~50 calls opened, mixed experience
   review the status in Jan/Feb whether it is worthwhile the costs
4. We have regular contacts with RH to discuss further license and support issues.

The next RH version REL4 is in beta testing and needs some ‘attention’ during next year

Reviewers Comment: Well accepted strategy in HEP, well communicated via HEPiX

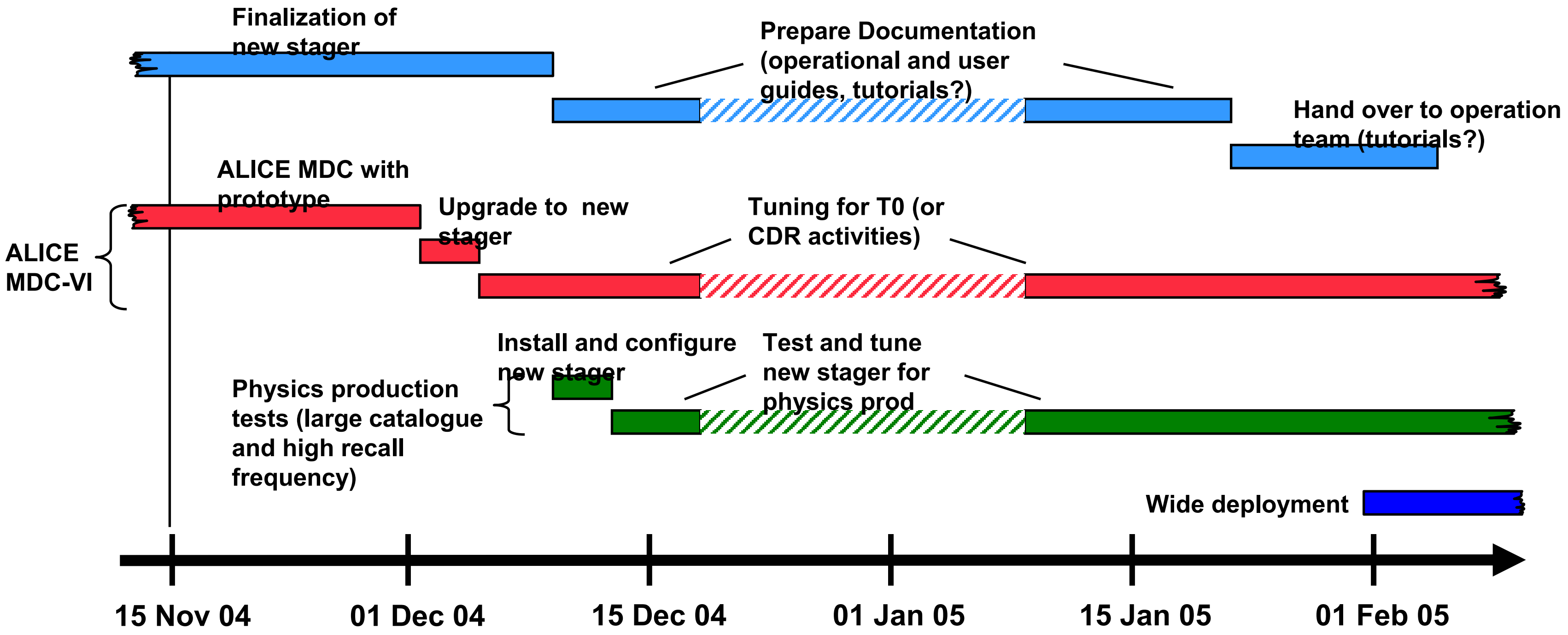
File systems

Reviewers Comment: today an open game

Batch system:

Reviewers Comment: up to the site

Severe delay of Castor/disk pool manager



New system is deployed in the high-throughput cluster and heavily tested. One additional person has been added specifically for testing from IT using the ALICE MDC programs. Good performance but yet too many instabilities, debugging phase

New stager developments delay

Several not foreseen but important extra activities :

The CASTOR development team has also the best knowledge of the internals of the current CASTOR system, thus are often involved in operational aspects as these have higher priority than developments.

To ease the heavy load on the CASTOR developers we were able to use man-power from our collaboration with PIC (Spain) and IHEP (Russia). These persons had already experience with CASTOR and were able to very quickly pick up some of the development tasks (there was no free time for any training of personal).

Because of the delays there was a risk to miss the ALICE MDC-VI milestone
Run with 80% prototype

Reviewers Comment:
Castor is critical for some groups. Beside the competition between daily business and development, nine month is a long period. Despite the management has looked for other resources it still seems to be understaffed. It's recommended to LHCC to check on shorter timescales on the progress

Complexity

Hardware components

	end 2004	2008
CPU capacity [SI2000]	2 Million	20 million
Disk space [TB]	450	4000
# CPU server	2000	4000
# disks	6000	8000
# disk server	400	800
# tape drives	50	200?
# tape cartridges	50000	50000

→ today we are less than a factor 2 in hardware complexity away from the system in 2008

Reviewers Comment: This gives good confidence

Network WAN

Observations, addressed by David FOSTER

- New: WAN is now part of computing
- Not all T1s are clear
- Traffic T1 -T1 and T1-T2 is unclear
- Connectivity of T1s in their responsibility

Reviewers Comment:

- Technically no difficulties foreseen,
- More a matter of money
- The computing models are needed, for Traffic simulation
- Coordination of T1 links with CERN is needed, maybe through GDB or GDA,
- proposed workshop in January is a good start.
- The roadmap is reasonable

Material purchase procedures

new proposal to be submitted to finance committee in December:

- for CPU and disk components
- covers offline computing and physics data acquisition (online)
- no 750 KCHF ceiling per tender
- speed up of the process (e.g. no need to wait for a finance committee meeting)
- effective already for 2005

Recommendation:
Strong Support for the new proposed purchase procedures, will ease things a lot, otherwise delays can be expected

Staff

About 10 FTE (out of 45) are missing in 2005.

Reviewers comment:
Losing a quarter of manpower would cause severe damage to the project, in particular because in 2005 and 2006 the data management has to be set up.

Summary

- Well managed set of projects
- Urgently needs more input from the experiments for data management
- Problem of Castor and Disk pool mgr due to competition between daily business and development (understaffed?)
- Networking needs more coordination from Tier1 sites
- Losing 10 staff out of 45 end phase I would cause severe problems, the amount seems to be about adequate for the tasks
- The proposed way to purchase equipment is strongly supported