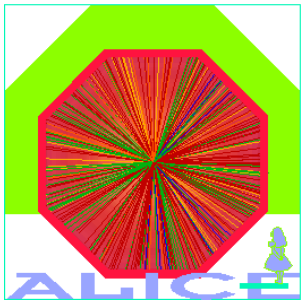


MC data production, reconstruction and analysis - lessons from PDC'04

Latchezar Betev

Geneva, December 9, 2004

Outline

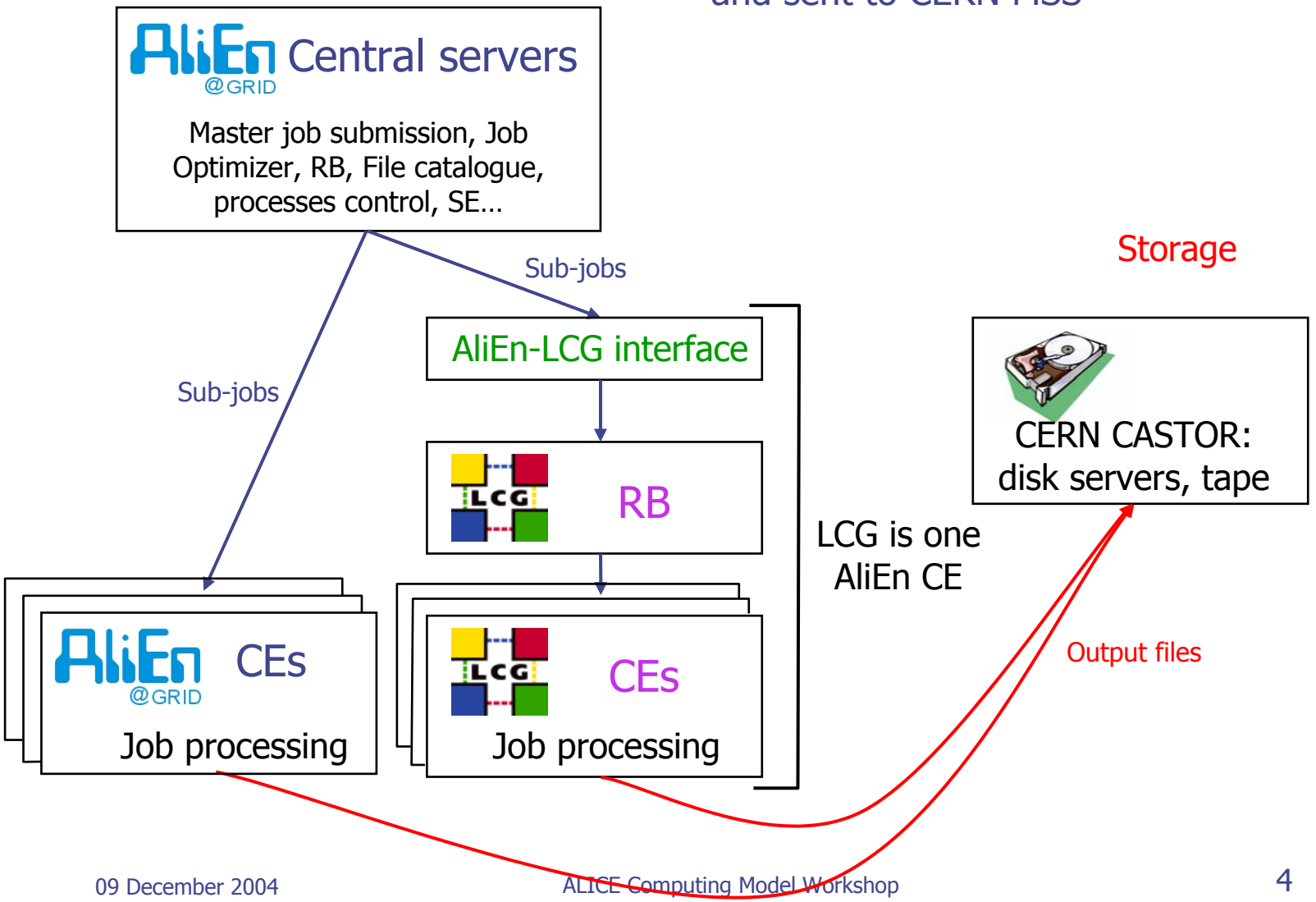
- Goals, structure and tasks
- Principles and platforms
- Statistics
- Operation Methods
- Monitoring
- Site participation and operation
- Summary

Goals, structure and tasks

- Test and validate the ALICE Offline computing model:
  - Produce and analyse ~10% of the data sample collected in a standard data-taking year
  - Use the complete set of off-line software: AliEn, AliROOT, LCG, Proof and in Phase 3 - the gLite and the ARDA analysis prototype
  - ***Test*** of the software and ***physics analysis*** of the produced data for the Alice PPR
- Structure – logically divided in three phases:
  - Phase 1 - Production of underlying Pb+Pb events with different centralities (impact parameters) + production of p+p events
  - Phase 2 - Mixing of signal events with different physics content into the underlying Pb+Pb events
  - Phase 3 – Distributed analysis

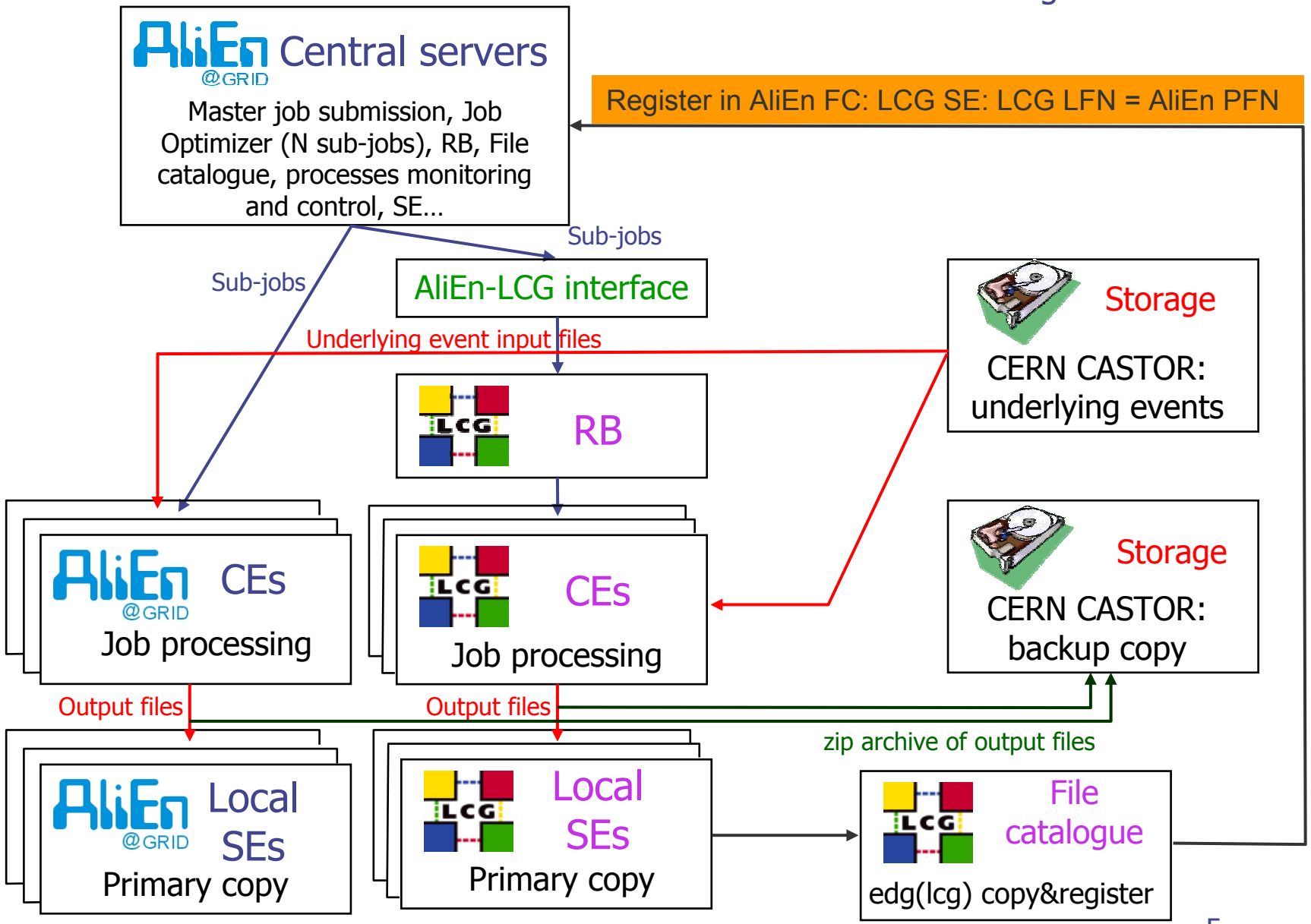
Phase 1 job structure

- Task - simulate the data flow in reverse: events are produced at remote centres and sent to CERN MSS

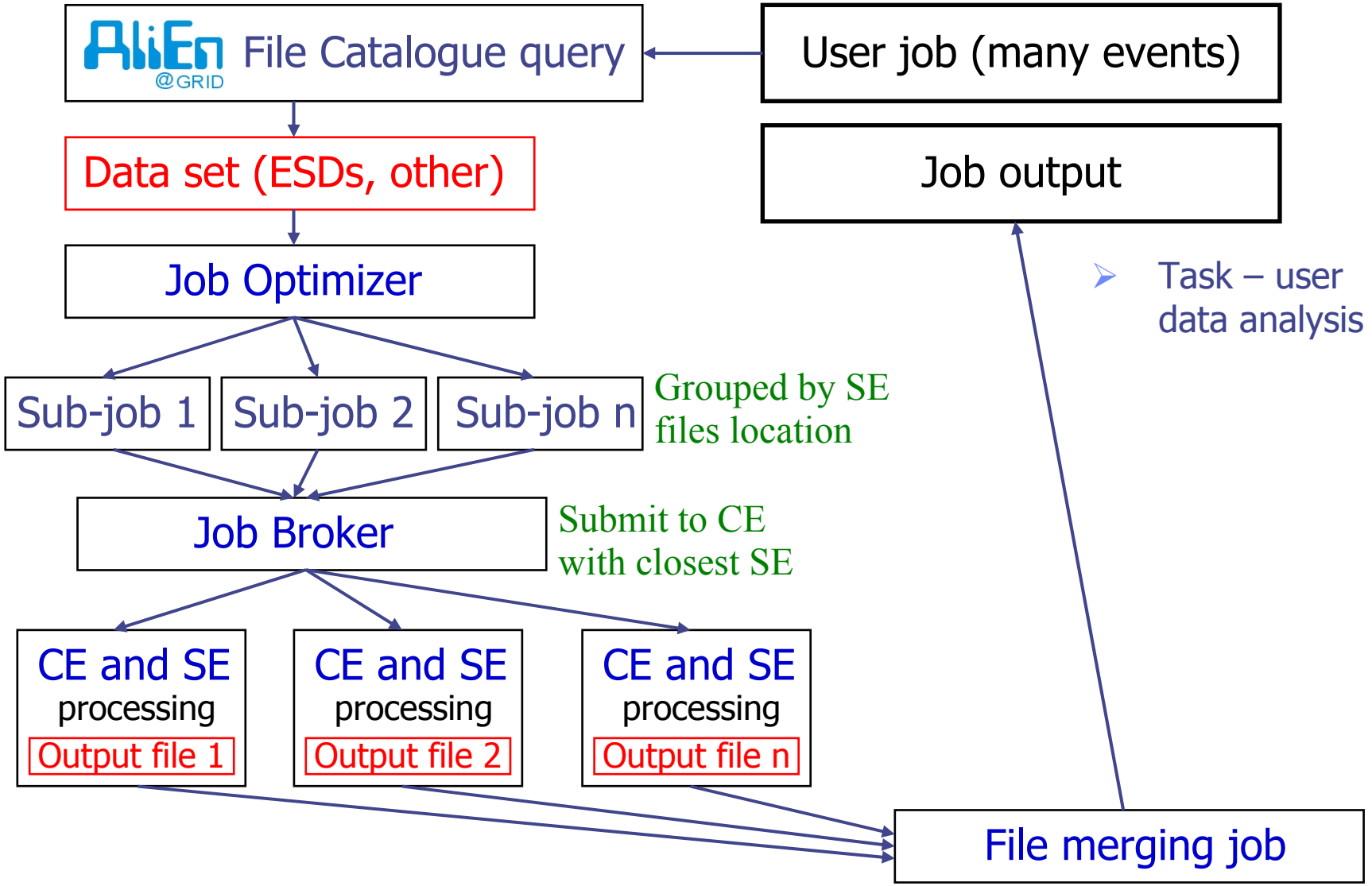


Phase 2 job structure

- Task - simulate the event reconstruction and remote event storage



Phase 3 job structure



Event statistics

- Phase 1 conditions and events:

Centrality name	Impact parameter value [fm]	Produced events
Cent1	0 - 5	20K
Per1	5 - 8.6	"
Per2	8.6 - 11.2	"
Per3	11.2 - 13.2	"
Per4	13.2 - 15	"
Per5	> 15	"

The distributed analysis

- Simplified view of the E2E ALICE analysis prototype:
  - ALICE experiment provides the UI (ROOT) and the analysis application
  - GRID middleware provides all the rest

- Analysis model:
  - execution: analysis tasks are produced from
    - GRID shell commands for batch analysis
    - ROOT prompt
      - interactive analysis mode: PROOF
      - batch analysis mode: job splitting

➢ Phase 2 physics signals:

Signal	No.of signal events per underlying	Number of jobs
Jets (un- and quenched) cent 1		
Jets PT 20-24 GeV/c	5	1666
Jets PT 24-29 GeV/c	5	1666
Jets PT 29-35 GeV/c	5	1666
Jets PT 35-42 GeV/c	5	1666
Jets PT 42-50 GeV/c	5	1666
Jets PT 50-60 GeV/c	5	1666
Jets PT 60-72 GeV/c	5	1666
Jets PT 72-86 GeV/c	5	1666
Jets PT 86-104 Gev/c	5	1666
Jets PT 104-125 GeV/c	5	1666
Jets PT 125-150 GeV/c	5	1666
Jets PT 150-180 GeV/c	5	1666
Total signal	399840	39984
Jets (un- and quenched) per 1		
Jets PT 20-24 GeV/c	5	1666
Jets PT 24-29 GeV/c	5	1666
Jets PT 29-35 GeV/c	5	1666
Jets PT 35-42 GeV/c	5	1666
Jets PT 42-50 GeV/c	5	1666
Jets PT 50-60 GeV/c	5	1666
Jets PT 60-72 GeV/c	5	1666
Jets PT 72-86 GeV/c	5	1666
Jets PT 86-104 Gev/c	5	1666
Jets PT 104-125 GeV/c	5	1666
Jets PT 125-150 GeV/c	5	1666
Jets PT 150-180 GeV/c	5	1666
Total signal	399840	39984

PHOS	**cent 1**	
Jet-Jet PHOS	1	20000
Gamma-jet PHOS	1	20000
Total signal	40000	40000
D0	**cent 1**	
D0	5	20000
Total signal	100000	20000
Charm & Beauty	**cent 1**	
Charm (semi-e) + J/psi	5	20000
Beauty (semi-e) + Y	5	20000
Total signal	200000	40000
MUON	**cent 1**	
Muon coctail cent1	100	20000
Muon coctail HighPT	100	20000
Muon coctail single	100	20000
Total signal	6000000	60000
MUON	**per 1**	
Muon coctail per1	100	20000
Muon coctail HighPT	100	20000
Muon coctail single	100	20000
Total signal	6000000	60000
MUON	**per 4**	
Muon coctail per4	5	20000
Muon coctail single	100	20000
Total signal	2100000	40000
Grand total	15239680	339968

Resources statistics

- Job, storage, data volumes and CPU work:
  - Number and duration:
    - 400 000 jobs
    - 6 hours/job
  - Number of files:
    - AliEn file catalogue: 9 million entries
    - 4.5 milion files distributes at 20 computing centres world-wide
  - Data volume:
    - 30 TB stored at CERN CASTOR
    - 10 TB stored at remote SEs
    - 200 TB network transfer CERN –> remote computing centres
  - CPU work:
    - 750 MSi2K hours

Principles and platforms

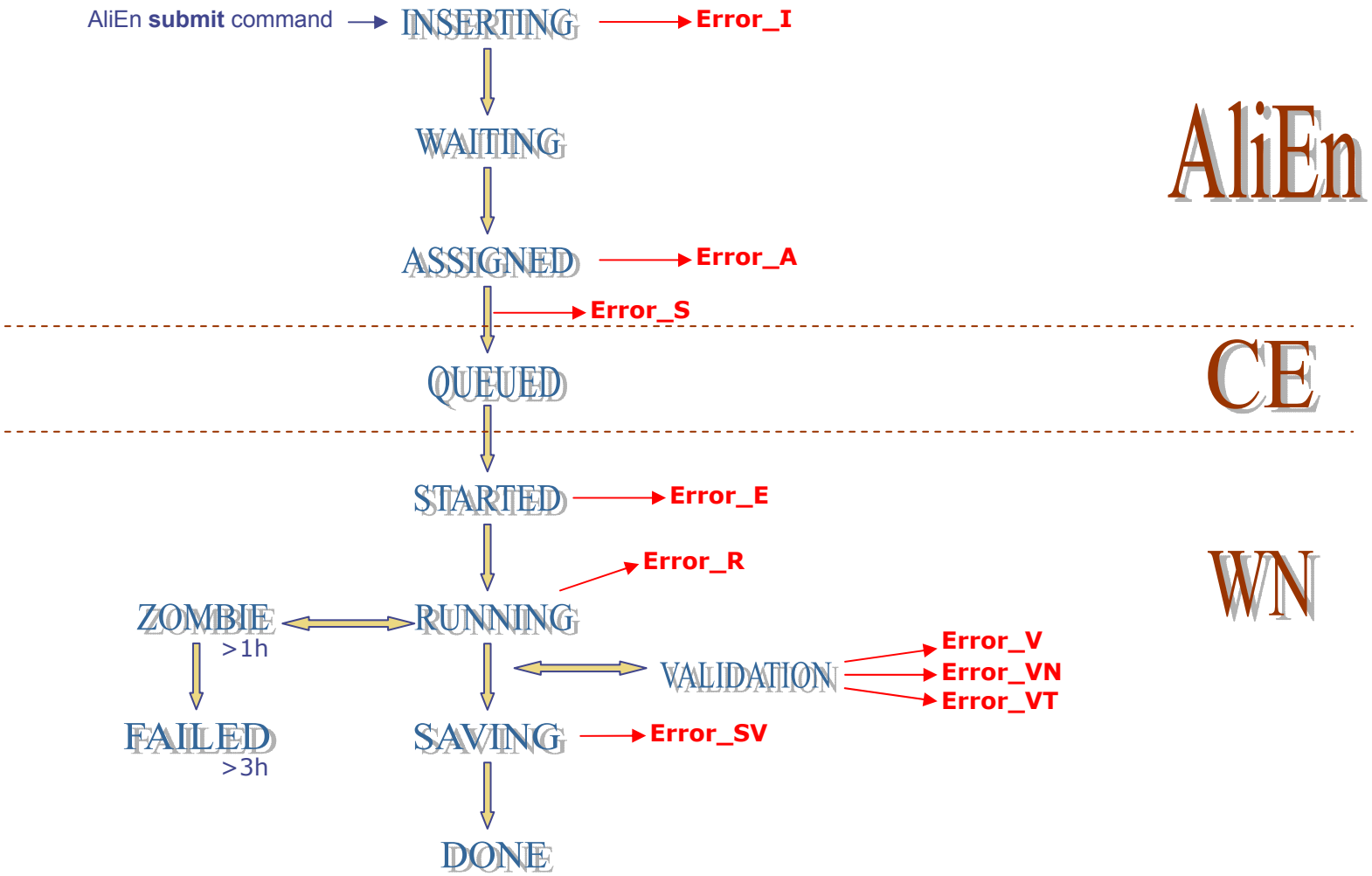
- True GRID data production and analysis: all jobs are run on the GRID, using only ***AliEn*** for access and control of native computing resources
- LCG GRID resources: access through AliEn-LCG interface
- In phase 3: gLite***+PROOF with ARDA E2E Prototype for ALICE***
- Software: AliRoot/GEANT3/ROOT/gcc3.2 libraries - distributed by AliEn:
  - The AliROOT code was kept backward compatible throughout the excercise
- Used platforms:
  - GCC 3.2 + i686 32-bit Cluster
  - GCC 3.2 + ia64 Itanium Cluster

Operation methods and groups

- Phase 1 and 2:
  - Central job submission – one person in charge of everything
- Phase 3:
  - Many users with centralized user support
- ALICE control and responsibility for the central AliEn services
- CERN storage and networking: IT/FIO, IT/ADC
- LCG operation: IT Grid Deployment Team
- Local CE: one local expert (typically the site administrator)

- The above structure was/is functioning very well:
  - Regular task-oriented group meetings
  - Remote (e-mail) consultations and error reporting to the experts at the CEs
  - More sophisticated tools: LCG Savannah, Global Grid User Support at FZK

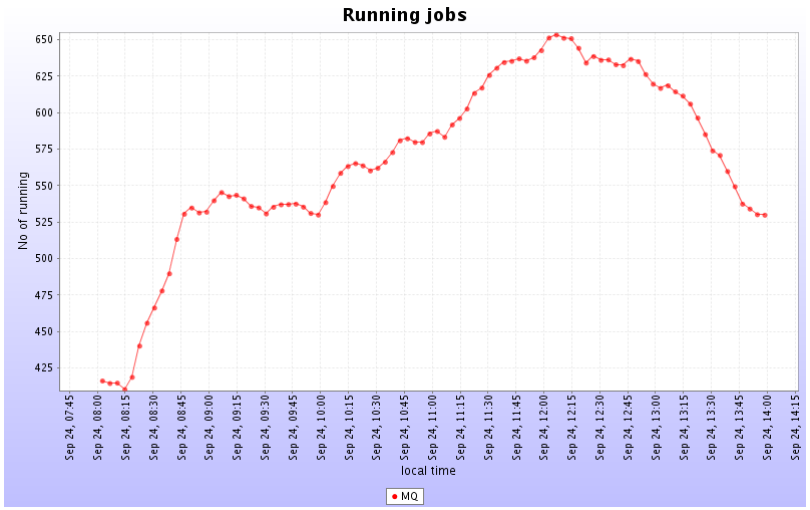
Monitoring – job wrapper

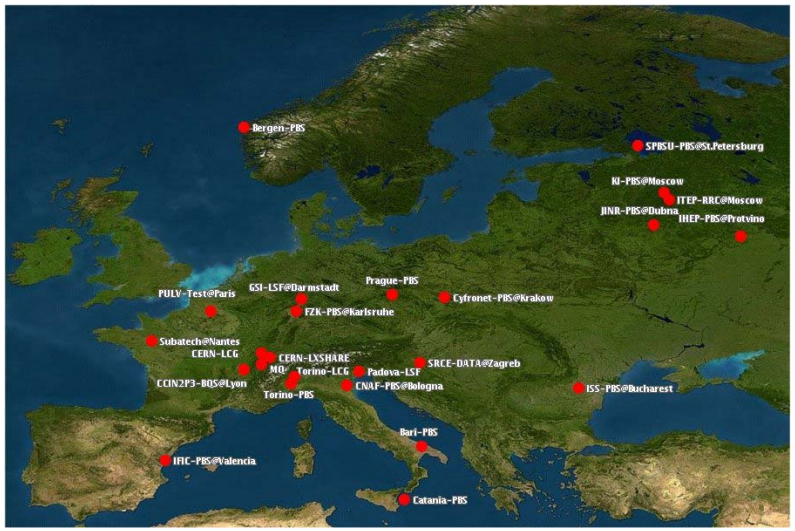


History Monitoring

MonALISA
MONitoring Agents using a Large Integrated Services Architecture

- ALICE repository – history of the entire DC
- ~ 1 000 monitored parameters:
  - Running, completed processes
  - Job status and error conditions
  - Network traffic
  - Site status, central services monitoring
  - ….
- 7 GB data
- 24 million records with 1 minute granularity – these are being analysed with the goal of improving the GRID performance



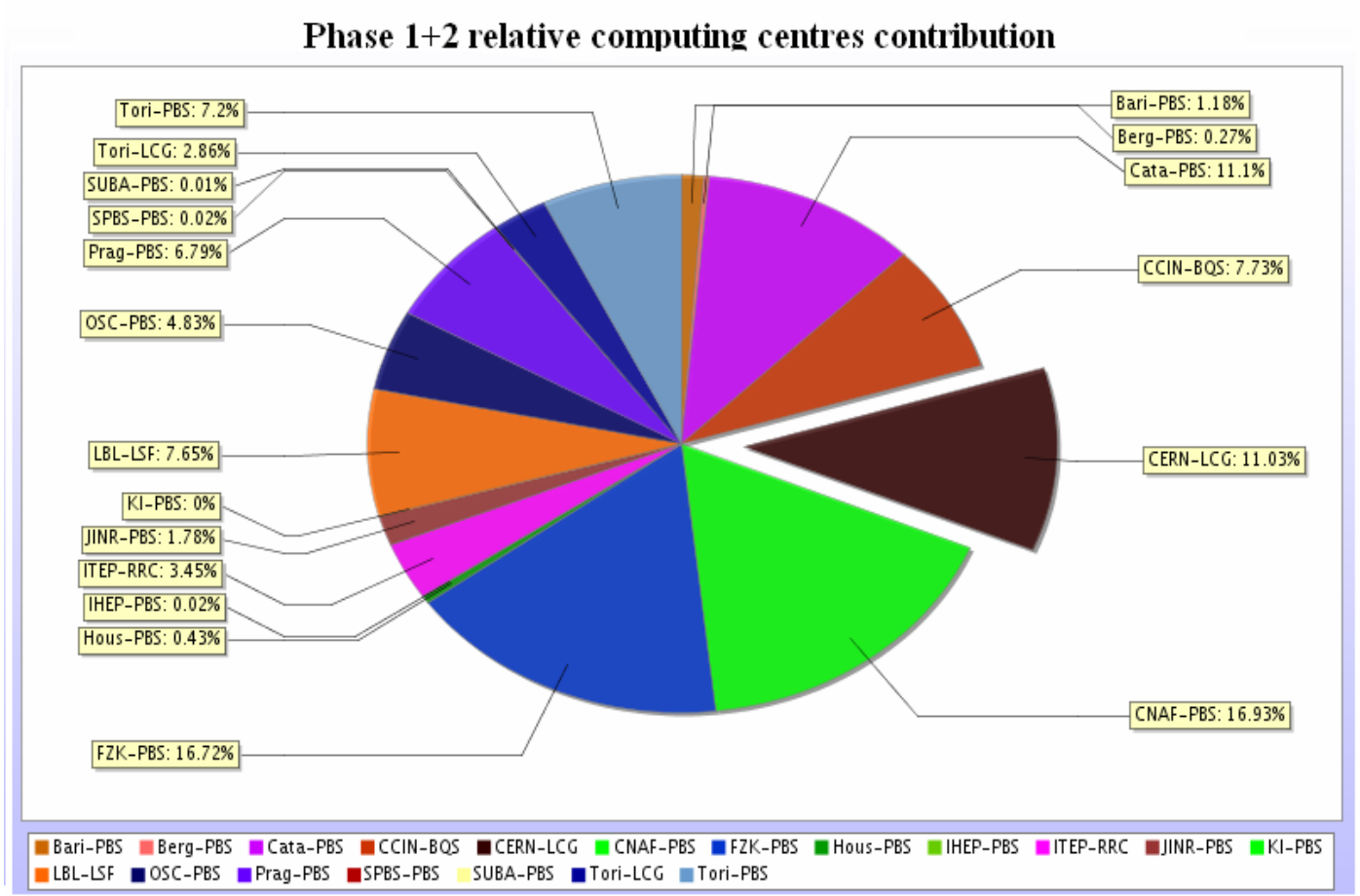


Job failure rates (AliEn)

- 1% - error submission:
  - CE scheduler not responding
- 3% - error while loading input data:
  - Remote SE not responding
- 10% - error during execution:
  - Job aborted – insufficient WN memory, AliRoot problem
  - Job cannot start – missing application directory
  - Job killed by CE scheduler
  - WN or global CE malfunction (all jobs on a give site die)
- 2% - error while saving output data:
  - Local SE not responding

- The above conditions are reported by the ClusterMonitor:
  - WN identification and type of problem:
    - "Out of memory", "Cannot find /home/aliprod/…."
- Are used for reporting the detailed problem to the CE administrator
- And may result in:
  - Automatic blocking of the site from the ALICE VO until the problem is fixed
  - Resubmission of the failed job
- **In cooperation with the CE administrators, the sites have been tuned during the DC and the failure rates are now very low**

Site participation



- 17 sites under AliEn direct control and additional resources through GRID federation (LCG)

Summary

- Running since 9 months with ***AliEn*** and currently in Phase 3 using the ***ARDA E2E*** analysis prototype and (soon) ***gLite***
- Permanent improvement of the AliEn, following requirements with increasing complexity:
  - More functionality, control and monitoring tools: job handling, job resubmission
  - The PDC'04 demonstrated the AliEn design scalability
- Shown successfully GRID interoperability: ***AliEn*** and ***LCG***
- The offline computing model has been extensively tested and validated during the PDC'04

Summary (2)

- The framework and duration of the PDC illustrated:
  - Many of the challenges we encountered would not have shown in a short DC:
    - Operational problems of the GRID and CE machinery for extended periods of time
    - Keeping a stable and backward compatible software, which is constantly being developed
    - Need for a stable personnel, especially at the T2 type computing centres
  - Keeping the pledged amount of computing resources throughout the exercise at the CEs:
    - Local priorities necessitate to be flexible
  - Flip side – using the provided resources to the maximum capacity:
    - Not always possible – breaks were needed to do software development and fixes, sometimes with very little advance warning

Summary (3)

- As expected – the most challenging part is the multi-user operation during phase 3:
  - The middleware needs to be "protected" and fortified in several areas
  - Documentation and "recipes" should be published and kept up-to-date
- Since the AliEn development is frozen:
  - The needed improvements are incorporated in ***gLite***
  - Phase 3 will also (we hope) provide a feedback to the ***gLite*** developers