The ALICE Data Challenge 2004 and the ALICE Distributed Analysis Prototype

Andreas Peters

September 29, 2004

CHEP'04

Interlaken

Outline

- Alice DC'04 Goals and Structure
- Alice DC'04 Phases
  - Phase 1 – Distributed Production
  - Phase 2 – Distributed Mixing/Reconstruction
  - Phase 3 – Distributed Analysis
- Alice DC'04 Performance
- Alice Distributed Analysis Prototype

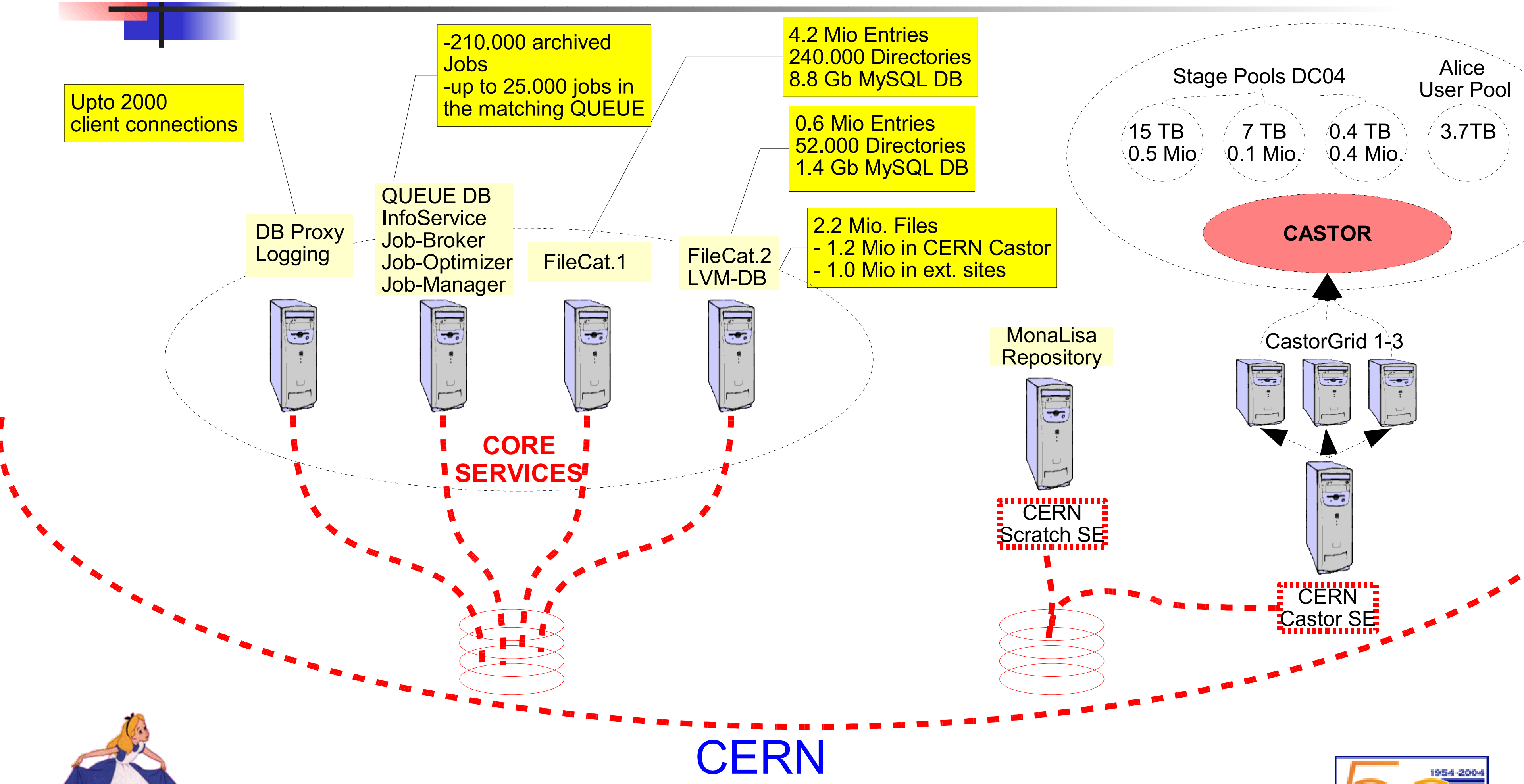
DC'04 Goals and Structure

- 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 entire system: AliEn, AliROOT, LCG, Proof…
  - 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

DC'04 Principles

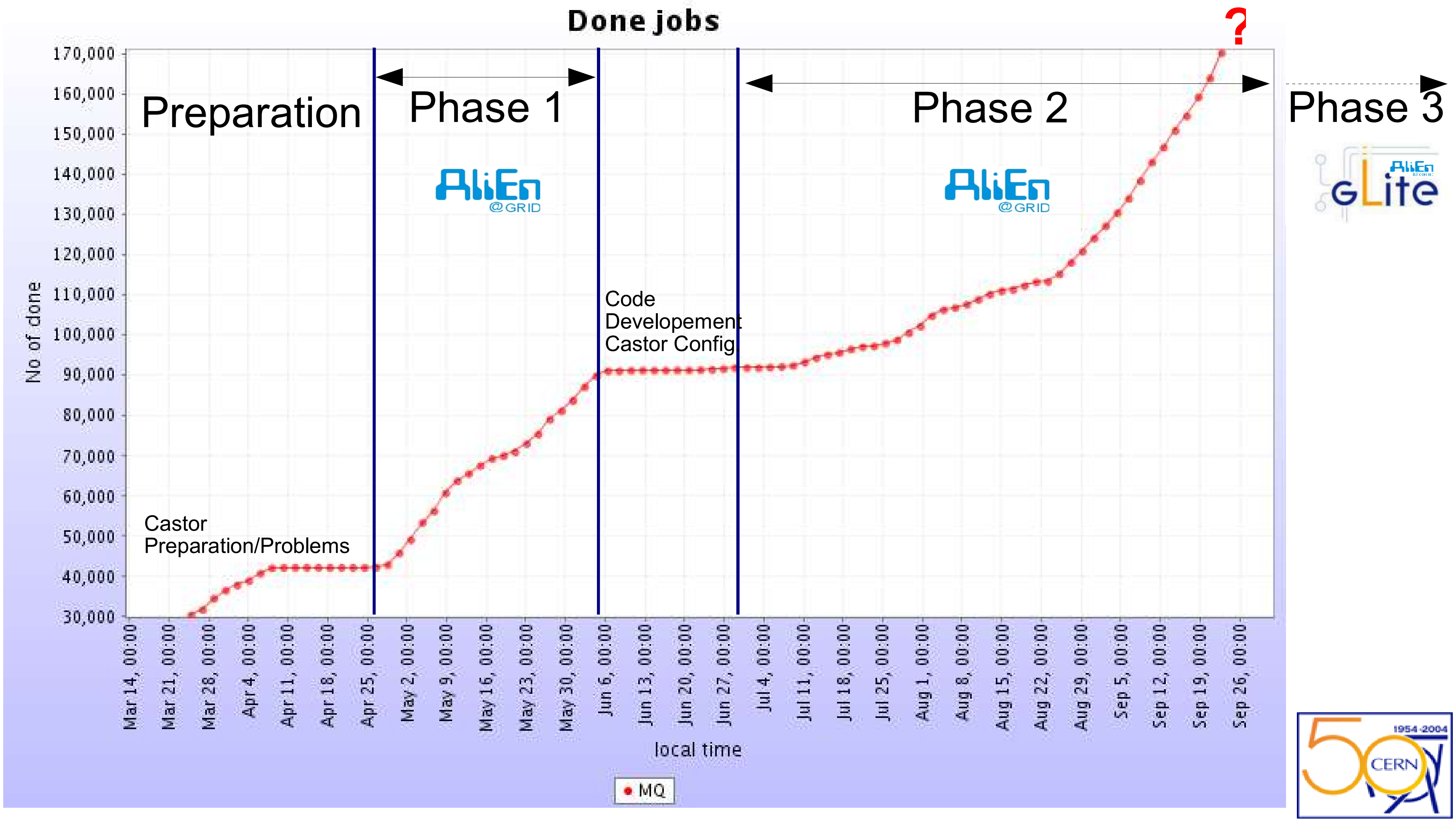
- Principles:
  - True GRID data production and analysis: all jobs are run on the GRID, using only **AliEn** for access and control of native computing resources the LCG resources
  - In phase 3 gLite+**PROOF** **(ARDA E2E Prototype for ALICE)**
  - Software AliRoot/GEANT3/ROOT/gcc3.2 libs distributed by AliEn
    - Used platforms
      - GCC 3.2 + i686 32-bit Cluster
      - GCC 3.2 + ia64 Itanium Cluster

DC'04 Hardware
– central components -



DC'04 Timeline

during the last 6 months



Preparation

Phase 1

Phase 2

Phase 3

Castor
Preparation/Problems

Code
Developement
Castor Config

DC'04 Monitoring

MonALISA
MONitoring Agents using a Large Integrated Services Architecture

Bergen-PBS

ITEP-RRC@Moscow

JINR-PBS@Dubna

Prague-PBS

PULV-Test@Paris

FZK-PBS@Karlsruhe

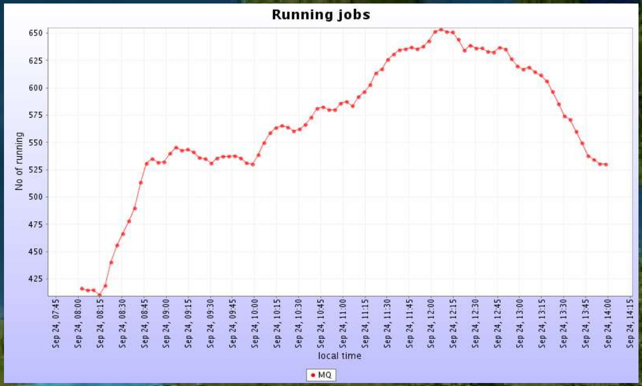
Cyfronet-PBS@Krakow

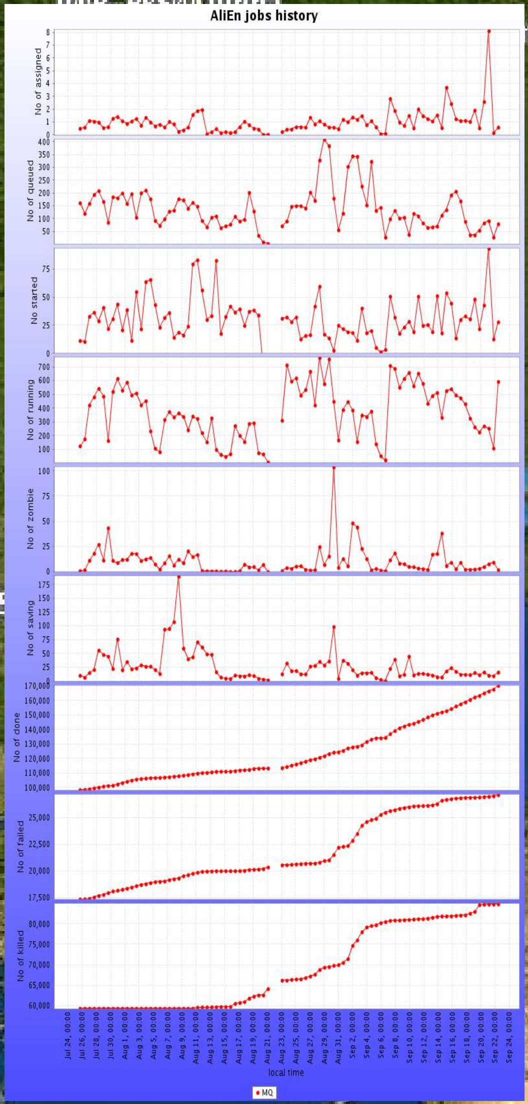
Subatech@Nantes

CERN-LCG

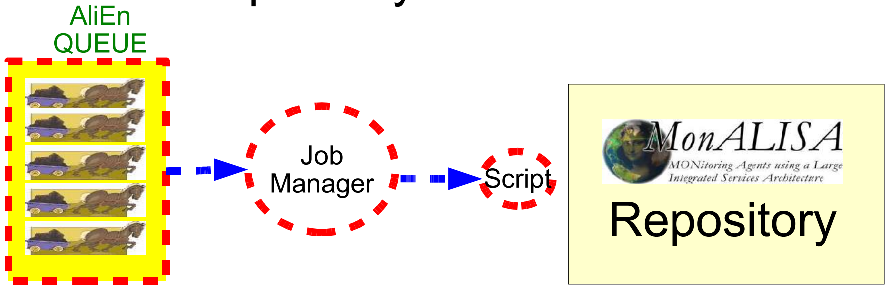
CERN-LXSHARE

ISS





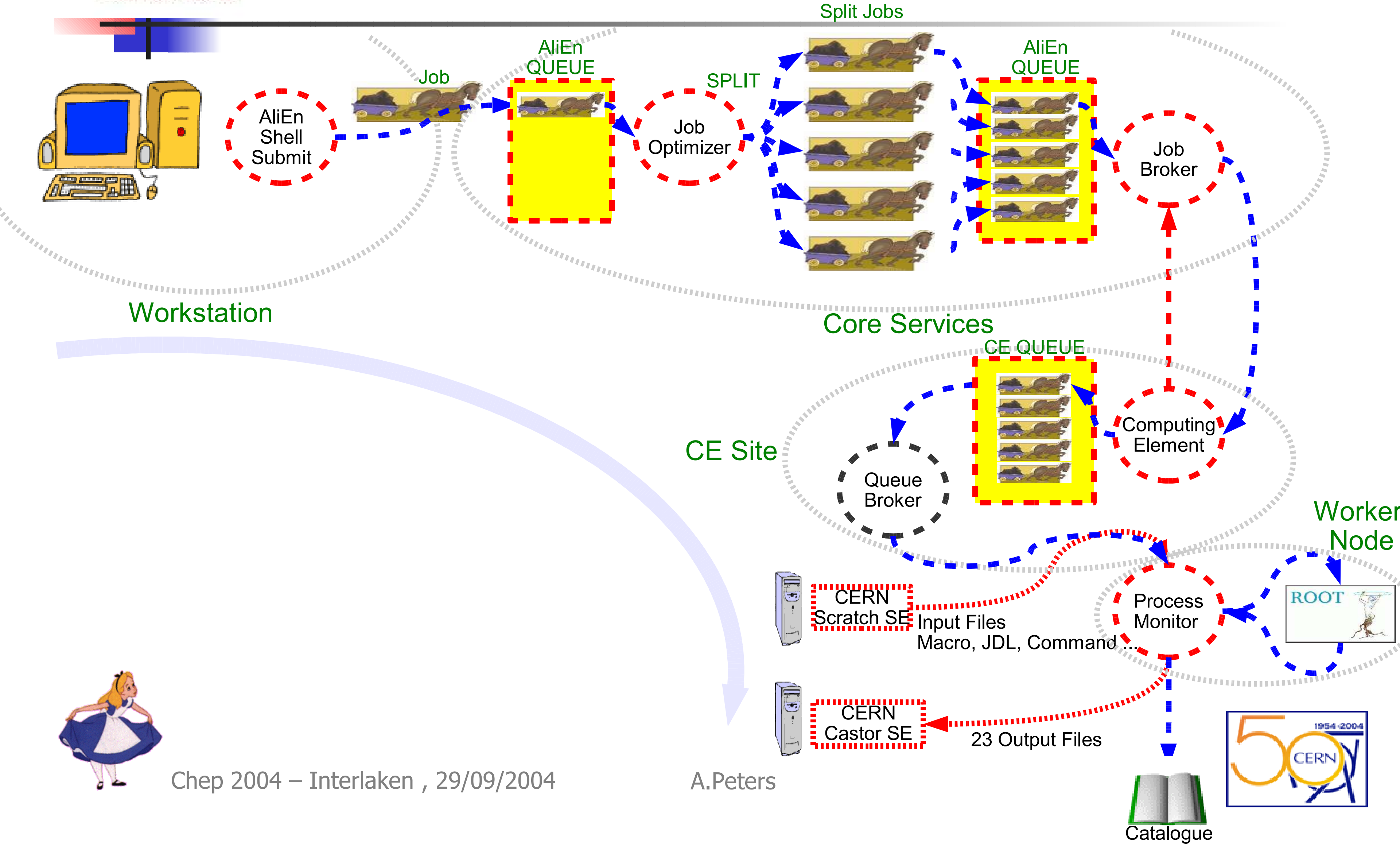
AliEn Queue information exported direct into MonaLisa Repository DB:



AliEn
QUEUE
Job Manager
Script
MonALISA
MONitoring Agents using a Large Integrated Services Architecture
Repository

http://aliens3.cern.ch:8080

Phase 1 Job Creation and Flow



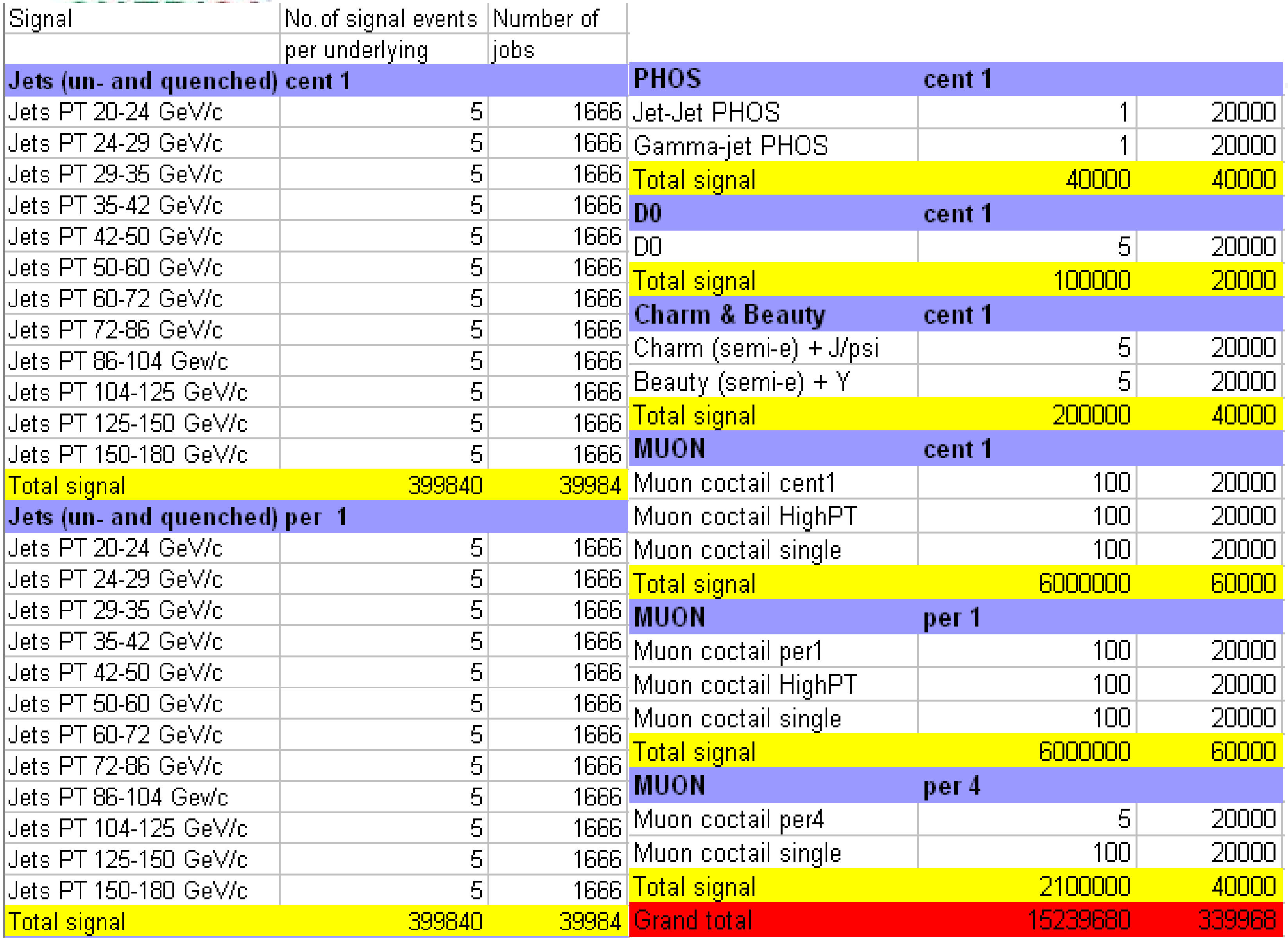
Phase 1 Processing

- Production of underlying Pb+Pb events with different centralities (impact parameters) + production of p+p events
  - Number of Jobs:
    - 6 x 20.000 events (type cent1/per1-5) = 56.000 jobs
      - 22.000 jobs á 8 hours (cent 1)
      - 22.000 jobs á 5 hours (per 1),
      - 12.000 jobs á 2.5 hours (per2-per5)
  - Number of files:
    - ~36 files per job
    - AliEn file catalogue: ~2.0 million files
    - CERN Castor: 1.3 million
  - File size:
    - Total: 26 TB (split on two CASTOR stagers)

Phase 2 - Mixing

- Mixing of signal events with different physics content into the underlying Pb+Pb events (underlying events are reused several times)
- Test of:
  - Standard production of signal events
  - Stress test of network and file transfer tools
  - Storage at remote SEs, stability (crucial for phase 3)
- Conditions, jobs …:
  - 62 different conditions
  - 340K jobs, 15.2M events
  - 10 TB produced data
  - 200 TB data transfer from CERN
  - 500 MSI2K hours CPU

Repartition of tasks (physics signals):

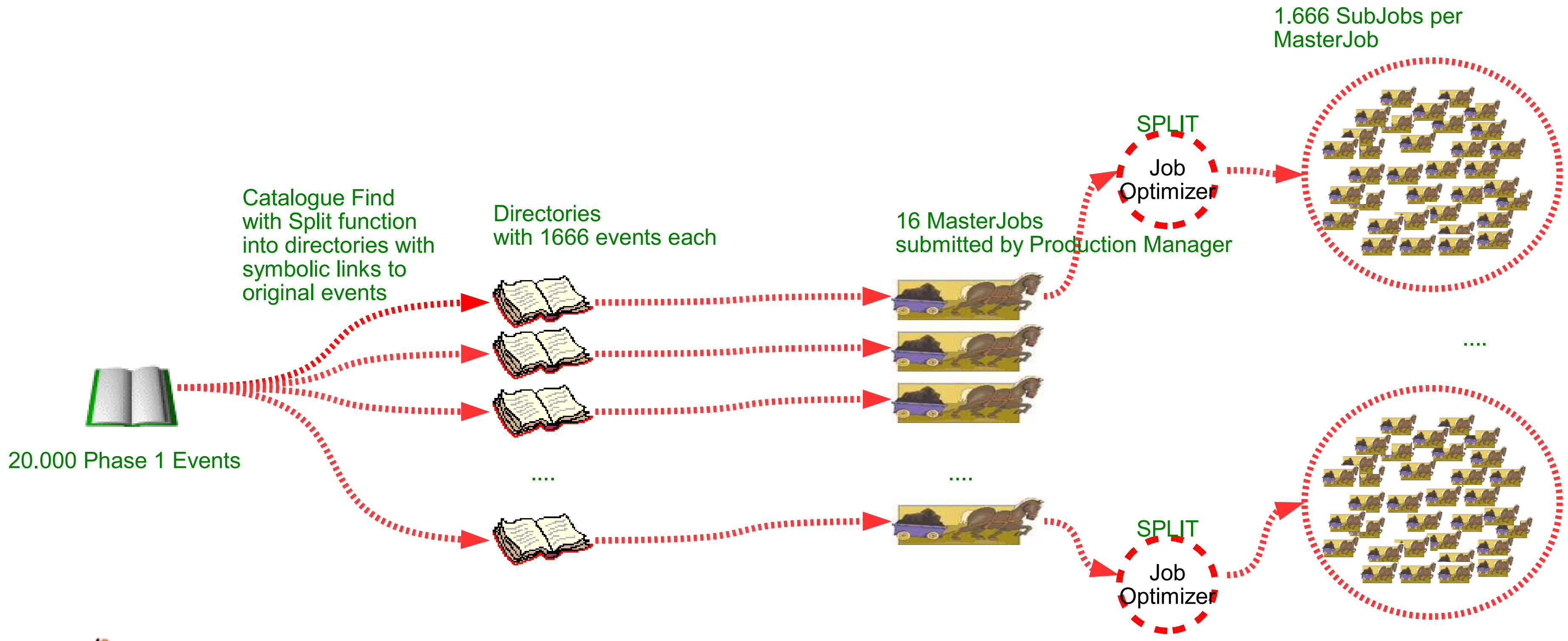


Signal	No.of signal events	Number of
	per underlying	jobs
Jets (un- and quenched) cent 1		
Jets PT 20-24 GeV/c	5	1666
Jets PT 24-29 GeV/c	5	1666
Jets PT 29-35 GeV/c	5	1666
Jets PT 35-42 GeV/c	5	1666
Jets PT 42-50 GeV/c	5	1666
Jets PT 50-60 GeV/c	5	1666
Jets PT 60-72 GeV/c	5	1666
Jets PT 72-86 GeV/c	5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

Phase 2 Job Creation Mixing in Jet Events

12x20.000 Input Files →16 Jobs → 20.000 Jobs → 200.000 OuputFiles



1.666 SubJobs per MasterJob

SPLIT

Job Optimizer

Catalogue Find with Split function into directories with symbolic links to original events

Directories with 1666 events each

16 MasterJobs submitted by Production Manager

....

20.000 Phase 1 Events

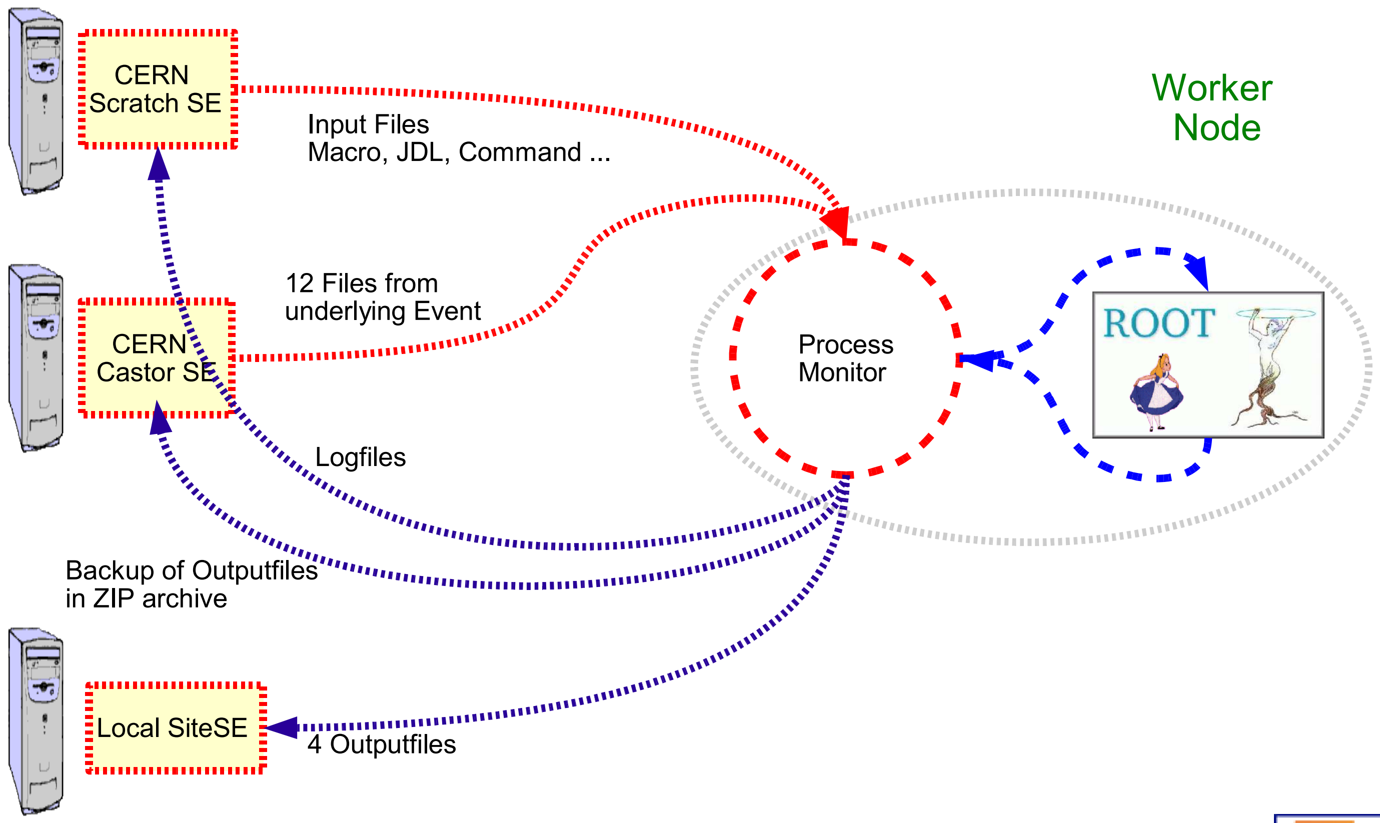
....

....

SPLIT

Job Optimizer

Phase 2 Job Data Flow



CERN Scratch SE

Input Files
Macro, JDL, Command ...

Worker Node

CERN Castor SE

12 Files from underlying Event

Process Monitor

Logfiles

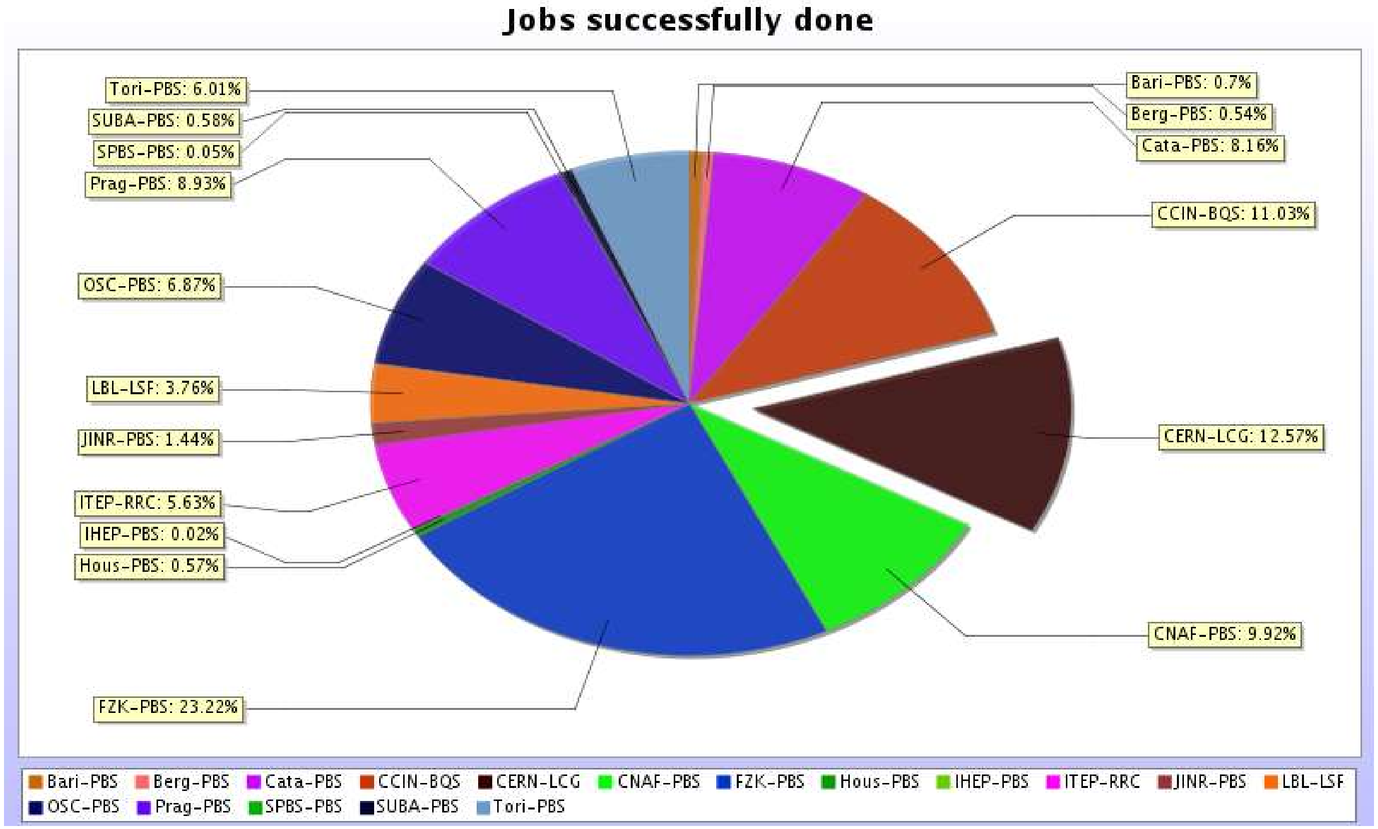
Backup of Outputfiles in ZIP archive

Local SiteSE

4 Outputfiles

Phase 2 Site Participation

16 AliEn sites + LCG



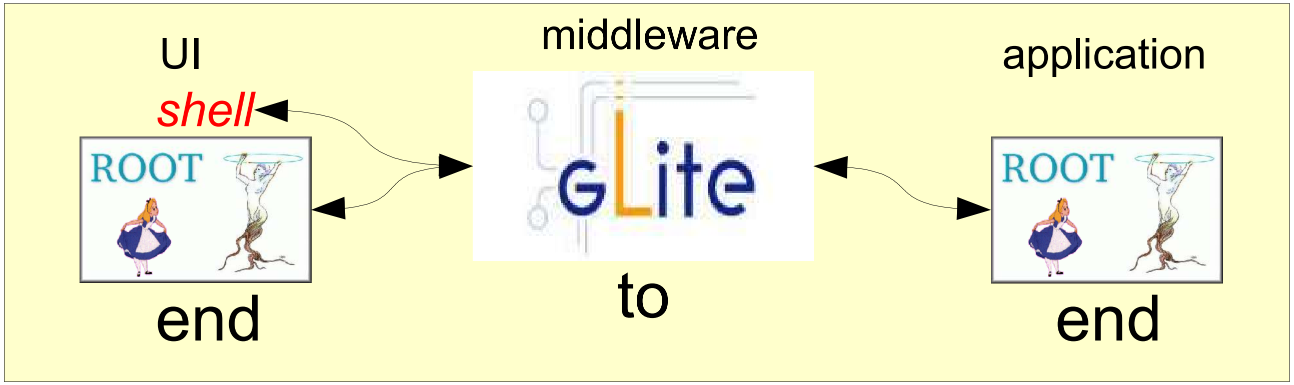
Phase 1 + 2 (3) Lessons learned

- .... „the present CASTOR version cannot keep more than 500.000 files staged on disk (per stager)“
  - Avoid small files and avoid to delete files in CASTOR
- Monitoring tools and control functions are the most essential tool for running large scale productions and identification of problems and bottlenecks
- production usage is a huge simplification of the multi-user usage ⇒ stronger reglementations are needed

Phase 3 – Distributed Analysis

E2E Protoype of Alice

- analysis approach:
  - ALICE experiment provides the UI (ROOT) and the analysis application
  - GRID middleware provides all the rest



UI

Phase 3 – Distributed Analysis
analysis model

- analysis model:
  - creation
    - users produce 'data sets' for analysis on the fly with catalogue and metadata queries
    - users use already produced 'data set' objects stored in a catalogue or locally
  - execution: analysis tasks are produced from
    - GRID shell commands for batch analysis
    - ROOT prompt
      - interactive analysis mode (PROOF)
      - batch analysis mode (w. job splitting)

Phase 3 – Distributed Analysis

Analysis Execution

From a grid-enabled shell: **glite submit analysis.jdl *try-01***

1 job per input data file

analysis.jdl

```
Executable = "aliroot";
Split="file";
Packages = {"AliRoot::4.01.Rev.04",
            "GEANT3::v0-6"};
Arguments = "Alice::Commands::AliRootS -x anarun.C";
InputFile= {"LF:/alice/cern.ch/user/a/aliprod/demo/anabox.tgz",
            "LF:/alice/cern.ch/user/a/aliprod/demo/analyze",
            "LF:/alice/cern.ch/user/a/aliprod/demo/anarun.C"
       };
InputData= {"LF:/alice/cern.ch/user/a/aliprod/demo/AOD/00001/*AOD.root"};
OutputFile= {"AOD.anal.pi+pi+.root@Alice::CERN::Scratch^aioforce",
             "sim.log@Alice::CERN::Scratch^aioforce"};
OutputDir={"demo/AOD/$1"};
```

Input Data Set⇒ Catalogue Query

Parameter from the submit command

Phase 3 – Distributed Analysis
Analysis Model – Batch Analysis

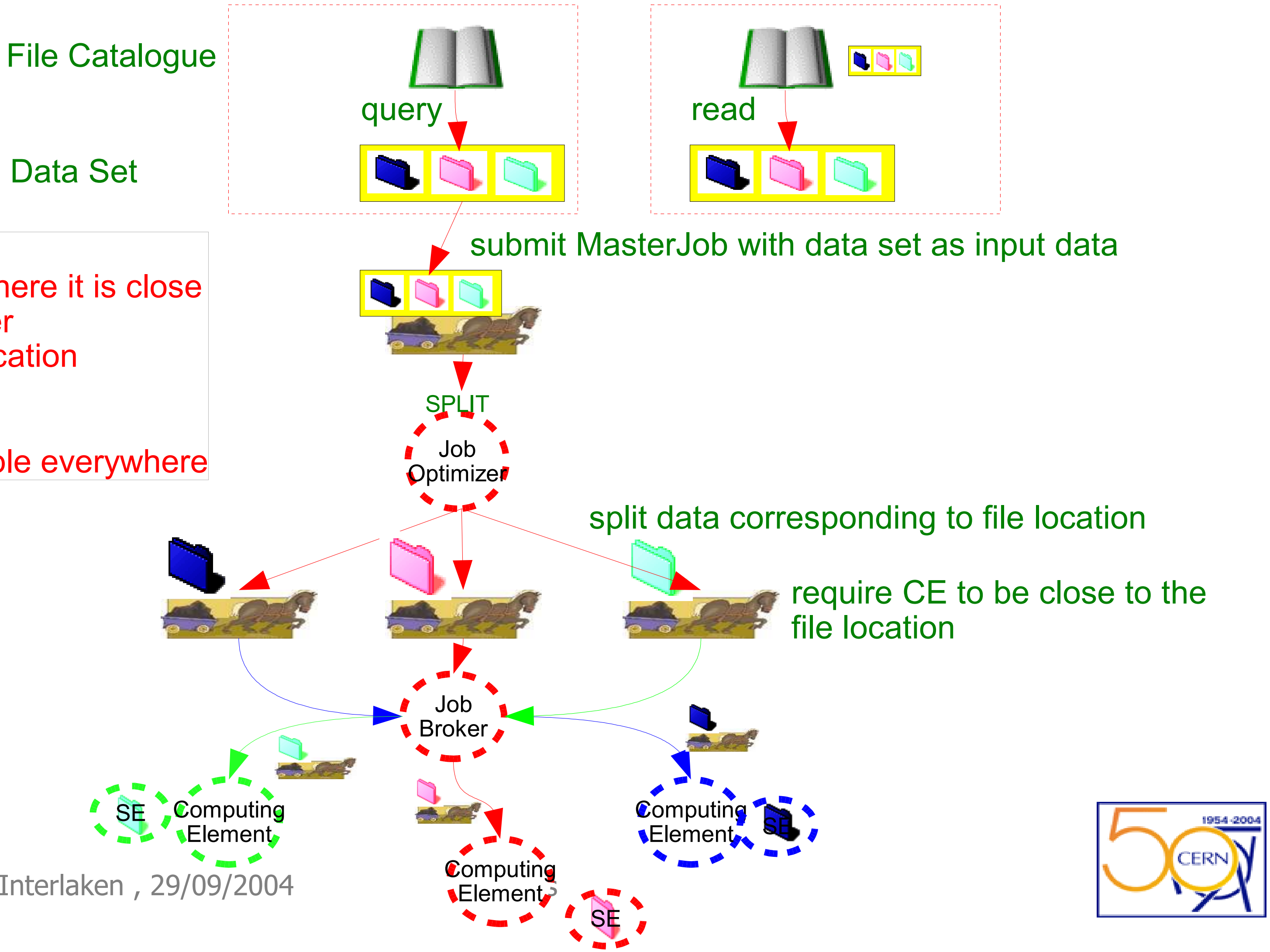
1st possibility:

-process data where it is close

-jobs don't trigger active data replication

2nd possibillity:

-data is accessible everywhere



File Catalogue

query

read

Data Set

submit MasterJob with data set as input data

SPLIT

Job Optimizer

split data corresponding to file location

require CE to be close to the file location

Job Broker

SE

Computing Element

Computing Element

SE

Computing Element

SE

Phase 3 – Distributed Analysis

Analysis Execution

From ROOT :

```
TDSet* dset = new TDSet(„TTree“,„mydataset“);
dset->AddQuery(„/glite/alice/“,“AOD.root“,“z<100“);
dset->SetProcessMode(kInteractive); // ^ kBatch
dset->Process(„mymacro.C“);

// to store the dataset for repetitive sessions:
dset->Write();
```

Phase 3 – Distributed Analysis
Analysis Model – Batch Analysis

- Processing Scheme:
  - Production of Data Sets

    ⇒ already possible as catalogue queries with AliEn/gLite
  - Data Sets are splitted into subsets and submitted as seperated jobs by a Job Optim.

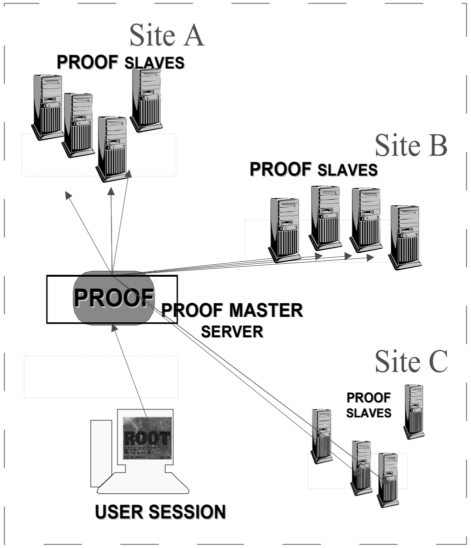
    ⇒ already possible with AliEn/gLite
  - Results are merged by a concurrent job which collects the output files

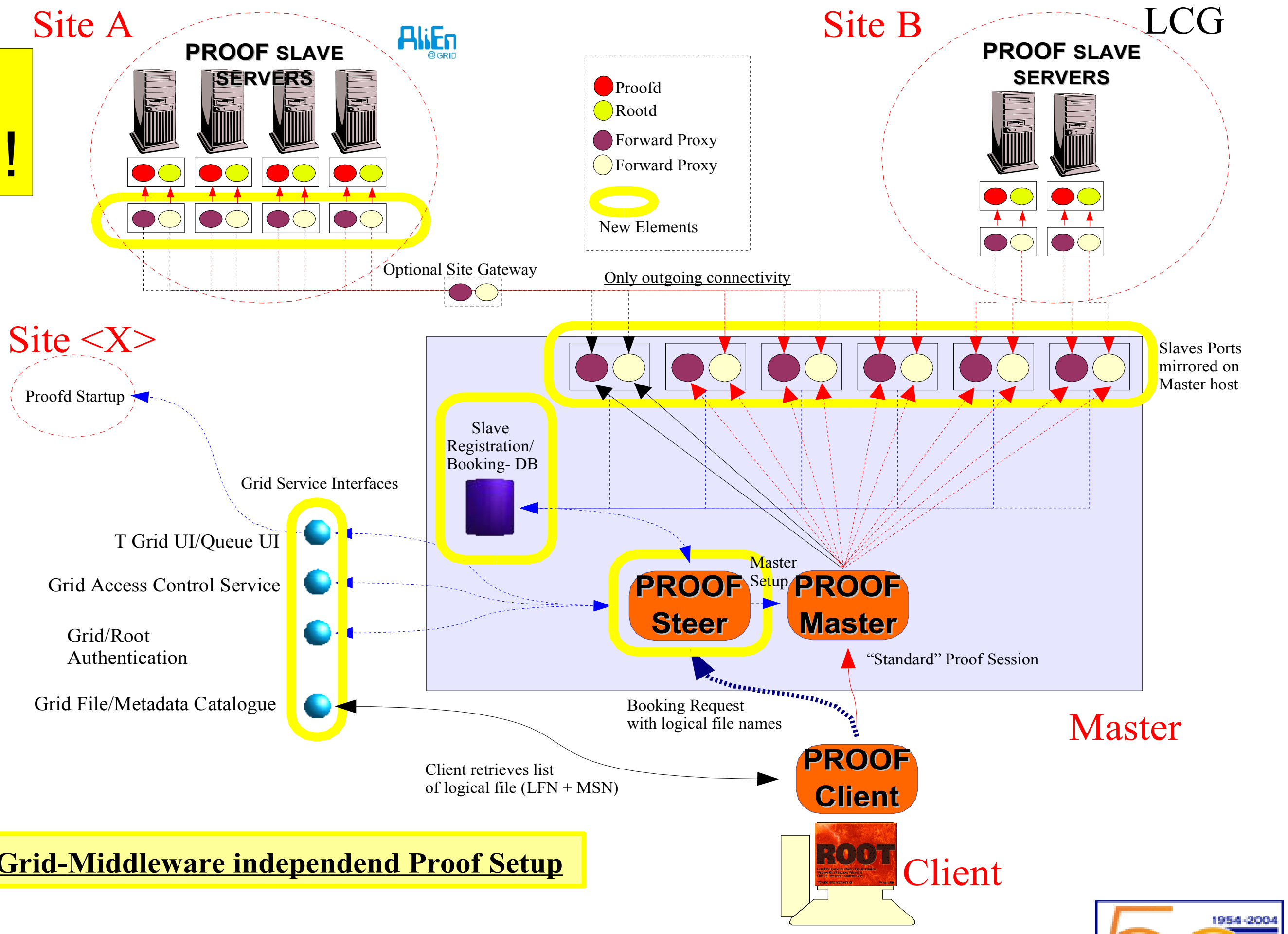
    ⇒ already possible with AliEn/gLite
  -

Remark: should add parallel and sequential job chains into gLite JDL syntax
JobChain = {"0:/glite/chep04/analysis.jdl","1:/glite/chep04/merge.jdl"}

Phase 3 – Distributed Analysis Interactive Analysis-PROOF

See talk by
Fons Rademakers!





Grid-Middleware independend Proof Setup

Summary+Outlook

- DC'04
    - Successfully running since 9 months with AliEn and continuing until the end of 2004
    - Many unexpected pitfalls and bottlenecks have been found, cured or circumvented
        - Permanent improvement of the system with the increasing requirements
            - System got more functionality,control and monitoring tools (master job handling, resubmission, MonaLisa...)
            - multi-user experiences give input for further developements
            - Demonstrated scalability of the AliEn design
    - Experienced gain and loss by federation of GRIDs (AliEn using LCG)
        - See poster of S.Bagnasco
    - The offline computing model has been tested and validated during the DC'04
    - ALICE will try to use the gLite prototype for Phase III (Analysis) from October on