Experiment Experiences in the 2004 Data Challenges

Dario Barberis

on behalf of the LHC Experiments

Outline

- Brief summary of 2004 Data Challenges
- General comments
- Specific comments on LCG-2
- Suggestions

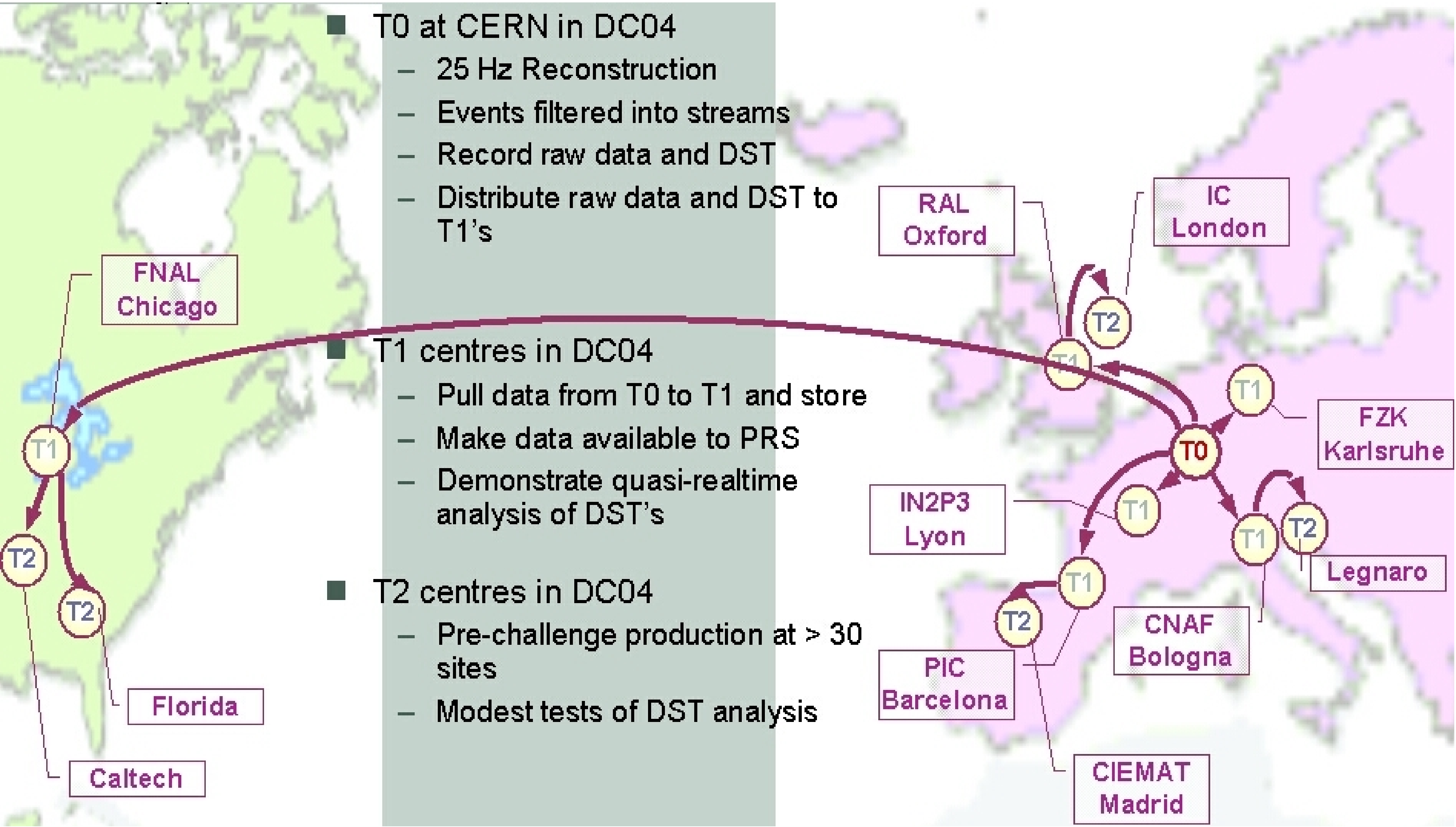
CMS: DC04 (1)

- Focused on organized (CMS-managed) data flow/access
- Functional DST with streams for Physics and Calibration
  - DST size ok, almost usable by "all" analyses; (new version ready now)
- Tier-0 farm reconstruction
  - 500 CPU. Ran at 25Hz. Reconstruction time within estimates.
- Tier-0 Buffer Management and Distribution to Tier-1's
  - TMDB: a CMS-built Agent system communicating via a Central Database.
  - Manages dynamic dataset "state", not a file catalog
- Tier-1 Managed Import of Selected Data from Tier-0
  - TMDB system worked.
- Tier-2 Managed Import of Selected Data from Tier-1
  - Meta-data based selection ok. Local Tier-1 TMDB ok.
- Real-Time analysis access at Tier-1 and Tier-2
  - Achieved 20 minute latency from Tier 0 reconstruction to job launch at Tier-1 and Tier-2
- Catalog Services, Replica Management
  - Significant performance problems found and being addressed

CMS: DC04 (2)

75 M Events
425 kSI2k-years
96 TB in POOL

- T0 at CERN in DC04
  - 25 Hz Reconstruction
  - Events filtered into streams
  - Record raw data and DST
  - Distribute raw data and DST to T1's
- T1 centres in DC04
  - Pull data from T0 to T1 and store
  - Make data available to PRS
  - Demonstrate quasi-realtime analysis of DST's
- T2 centres in DC04
  - Pre-challenge production at > 30 sites
  - Modest tests of DST analysis

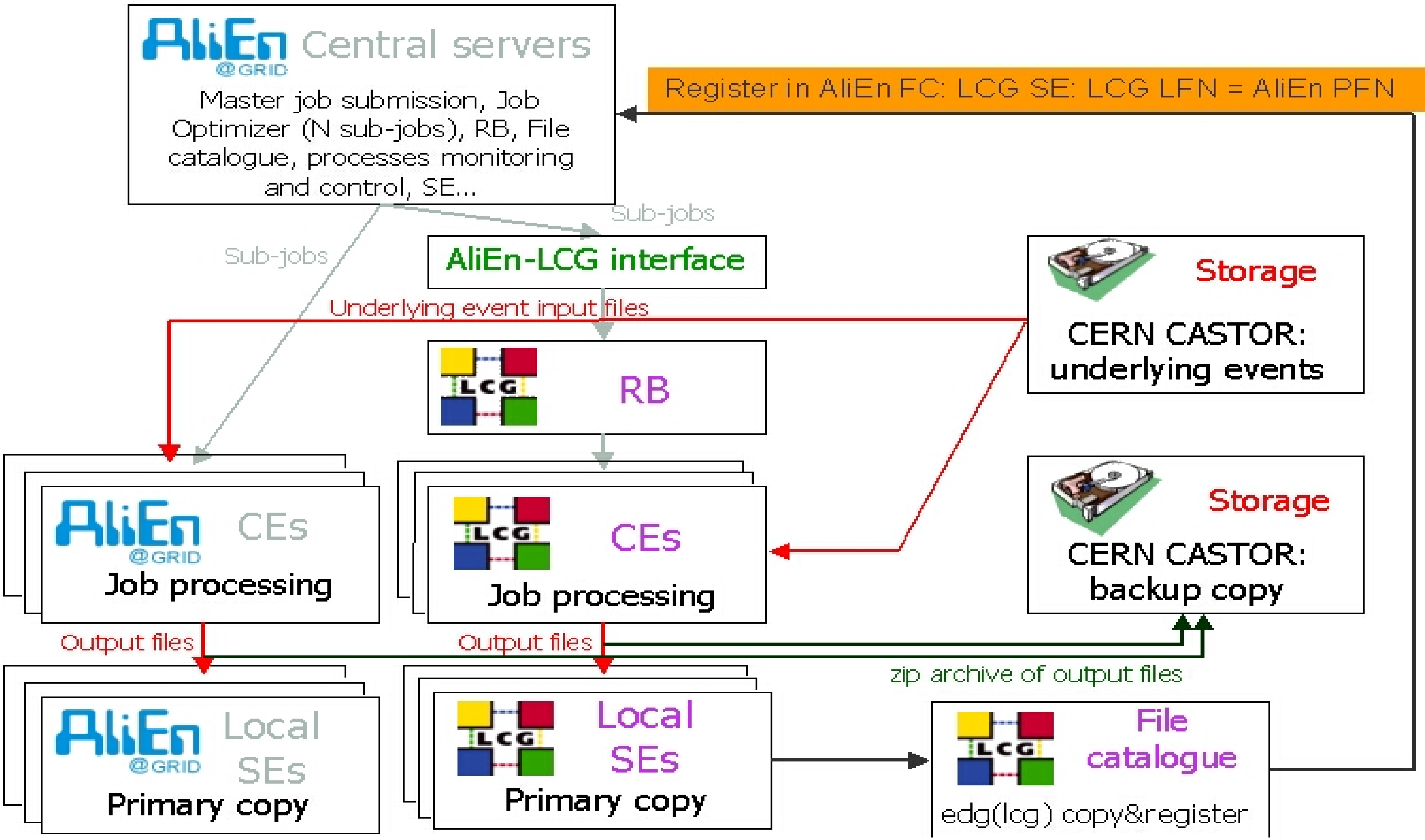


ALICE: PDC04 (1)

- 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 ALICE off-line framework: AliEn, AliRoot, LCG, PROOF…
  - Experiment with Grid enabled distributed computing
  - Triple purpose: test of the middleware, the software and physics analysis of the produced data for the Alice PPR
- Three phases
  - Phase I - Distributed production of underlying Pb+Pb events with different centralities (impact parameters) and of p+p events
  - Phase II - Distributed production mixing different signal events into the underlying Pb+Pb events (reused several times)
  - Phase III – Distributed analysis
- 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 and, through an interface, the LCG resources
  - In phase III GLite+ARDA

ALICE: PDC04 (2)

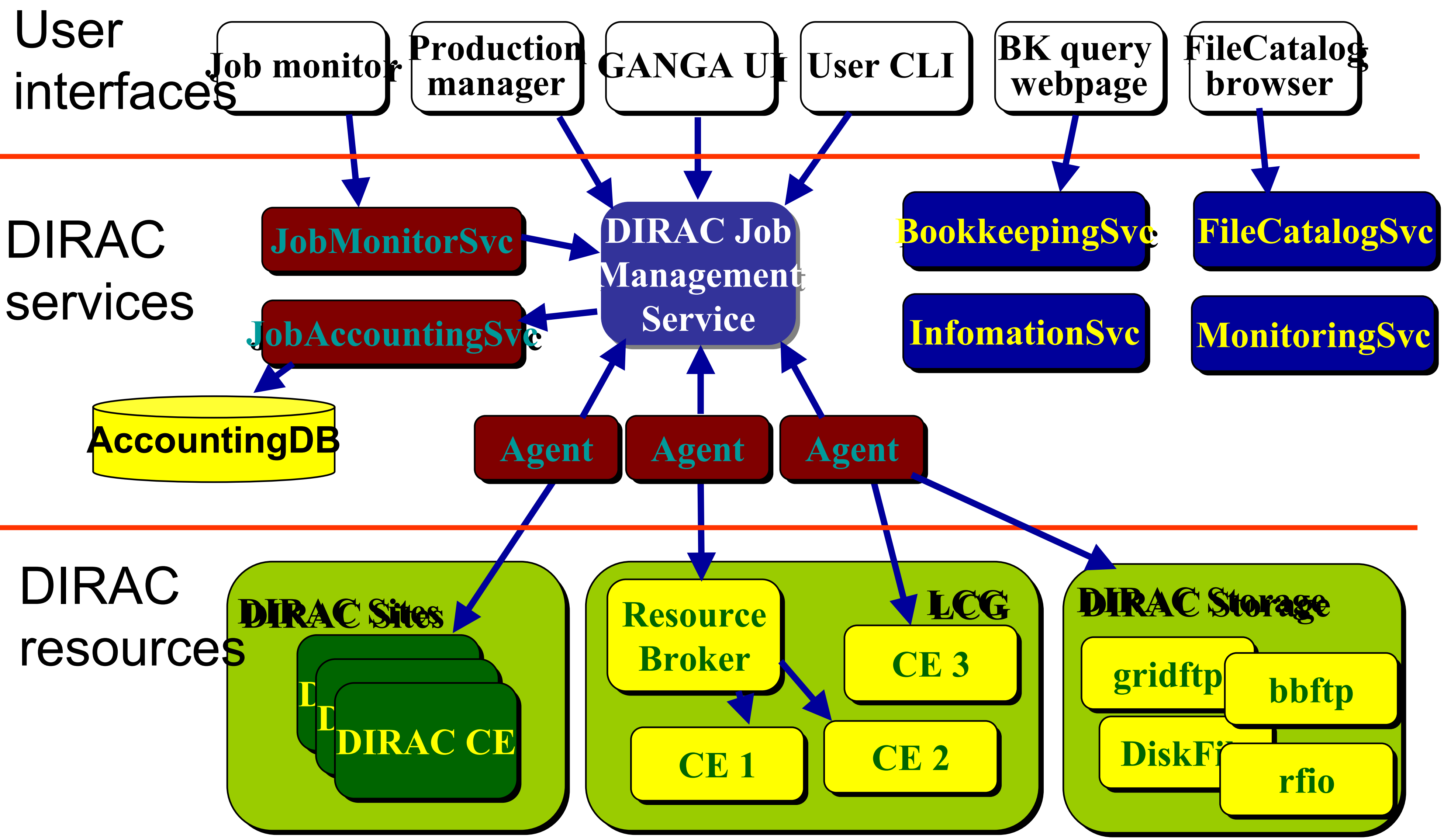
Structure of event production in Phase II



LHCb: DC04 (1)

- Gather information for LHCb Computing TDR
- Physics Goals:
  - HLT studies, consolidating efficiencies.
  - B/S studies, consolidate background estimates + background properties.
- Requires quantitative increase in number of signal and background events:
  - 30 10^6 signal events (~80 physics channels).
  - 15 10^6 specific backgrounds.
  - 125 10^6 background (B inclusive + min. bias, 1:1.8).
- Split DC'04 in 3 Phases:
  - Production: MC simulation (done).
  - Stripping: Event pre-selection (to start soon).
  - Analysis (in preparation).

LHCb: DC04 (2)


User interfaces
Job monitor
Production manager
GANGA UI
User CLI
BK query webpage
FileCatalog browser
DIRAC services
JobMonitorSvc
DIRAC Job Management Service
BookkeepingSvc
FileCatalogSvc
JobAccountingSvc
InfomationSvc
MonitoringSvc
AccountingDB
Agent
Agent
Agent
DIRAC resources
DIRAC Sites
DIRAC CE
Resource Broker
LCG
CE 3
CE 1
CE 2
DIRAC Storage
gridftp
bbftp
DiskFi
rfio

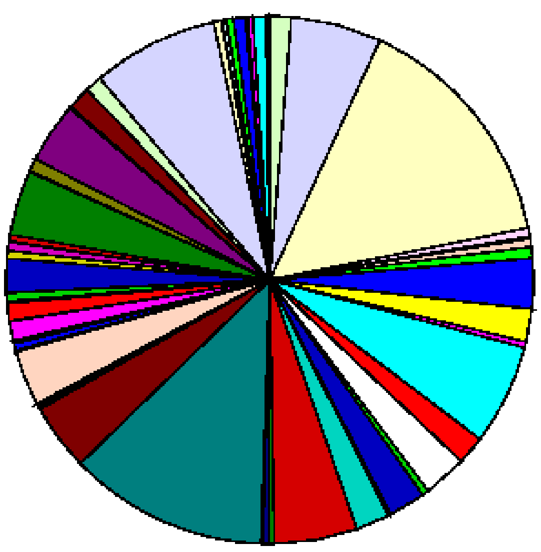
LHCb: DC04 (3)

186 M Events
350 kSI2k-years
61 TB in POOL

20 DIRAC Sites

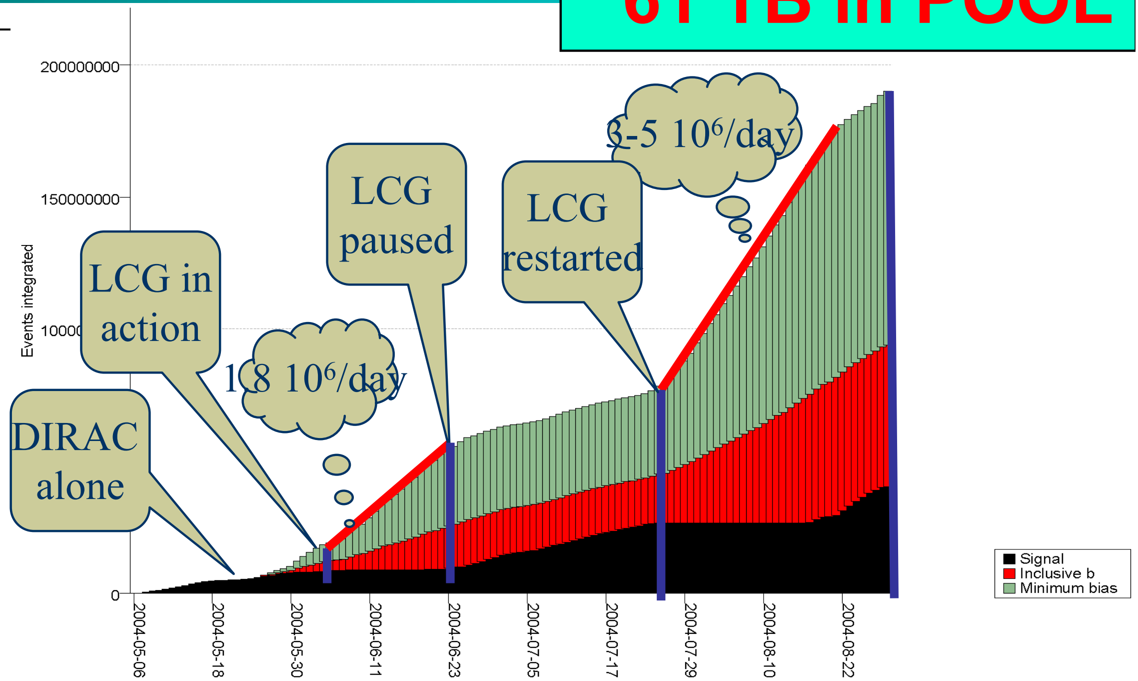
43 LCG Sites
(8 also DIRAC sites)

Events: 185.55 M

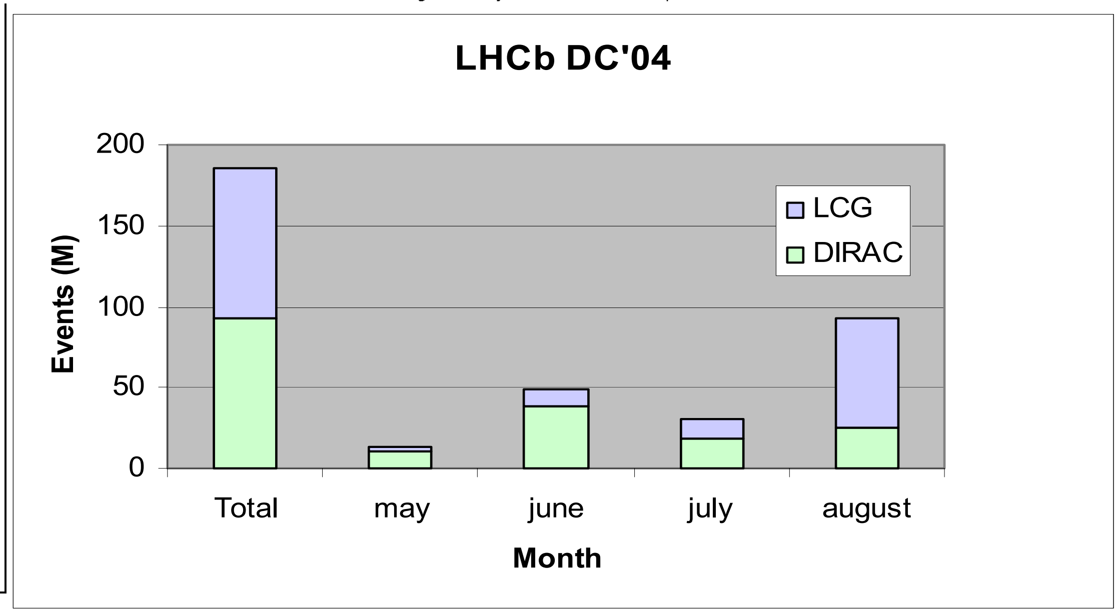


@2004-09-06 Between 20[illegible] – 2004-08-31

DIRAC.Barcelona.es	1.305%
DIRAC.Bologna.it	5.560%
DIRAC.CERN.ch	14.96%
DIRAC.CracowAgu.pl	0.532%
DIRAC.IF-UFRJ.br	0.124%
DIRAC.IHEP-Protvino.ru	0.504%
DIRAC.IHEP2-Protvino.ru	0.691%
DIRAC.ITEP-Moscow.ru	3.066%
DIRAC.Imperial.uk	2.017%
DIRAC.JINR-Dubna.ru	0.353%
DIRAC.Karlsruhe.de	6.181%
DIRAC.LHCBONLINE.ch	1.752%
DIRAC.Liverpool.uk	2.674%
DIRAC.Lpool.uk	0.405%
DIRAC.Lyon.fr	2.273%
DIRAC.Manno.ch	0.035%
DIRAC.Oxford.uk	0.137%
DIRAC.Santiago.es	2.053%
DIRAC.ScotGrid.uk	5.078%
DIRAC.Zurich.ch	0.355%
LCG.BHAM-HEP.uk	0.341%
LCG.Barcelona.es	0.106%
LCG.Bari.it	0.010%
LCG.CERN.ch	12.22%
LCG.CNAF.it	4.097%
LCG.Cagliari.it	0.049%
LCG.Cambridge.uk	0.128%
LCG.Carleton.ca	0.146%
LCG.Catania.it	0.031%
LCG.FNAL.us	0.017%
LCG.FZK.de	3.375%
LCG.Ferrara.it	0.094%
LCG.IN2P3.fr	0.419%
LCG.ITEP.ru	0.176%
LCG.Imperial.uk	1.188%
LCG.JINR.ru	0.021%
LCG.KFKI.hu	1.077%
LCG.Krakow.pl	0.127%
LCG.Lancashire.uk	0.515%
LCG.Legnaro.it	2.076%
LCG.Manchester.uk	0.473%
LCG.Milano.it	0.527%
LCG.Montreal.ca	0.041%
LCG.NCU.tw	0.408%
LCG.NIKHEF.nl	3.963%
LCG.Napoli.it	0.062%
LCG.Oxford.uk	0.791%
LCG.PIC.es	3.716%
LCG.Padova.it	0.099%
LCG.QMUL.uk	1.417%
LCG.RAL-HEP.uk	1.042%
LCG.RAL.uk	7.726%
LCG.RHUL.uk	0.463%
LCG.Roma.it	0.052%
LCG.SARA.nl	0.246%
LCG.SINP.ru	0.034%
LCG.Sheffield.uk	0.420%
LCG.Torino.it	0.722%
LCG.Toronto.ca	0.143%
LCG.Triumf.ca	0.317%
LCG.UCL-CCC.uk	0.795%
LCG.USC.es	0.193%
LCG.WEIZMANN.il	0.034%



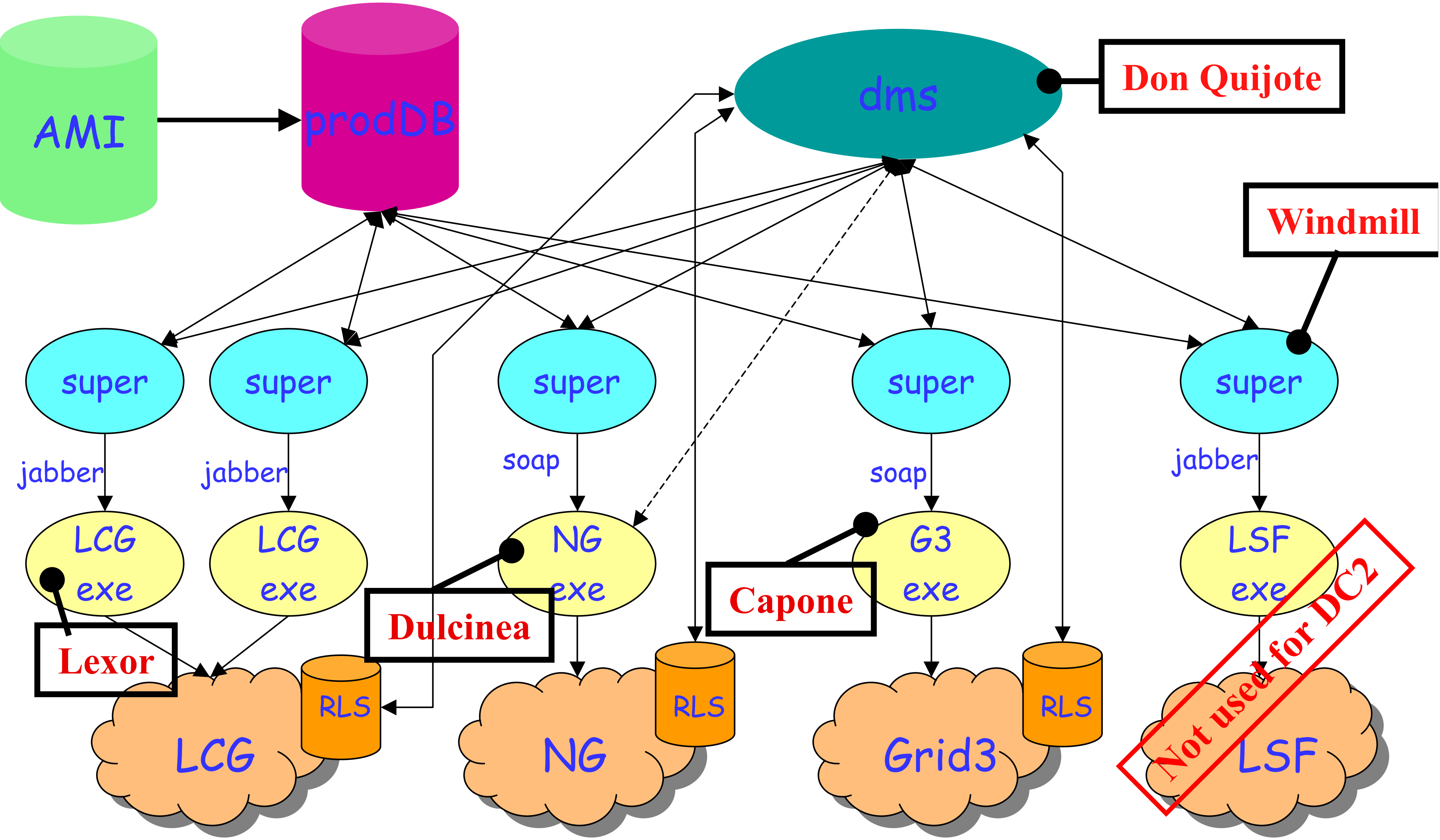
Events integrated : All jobs with ANY known replica



ATLAS: DC2 (1)

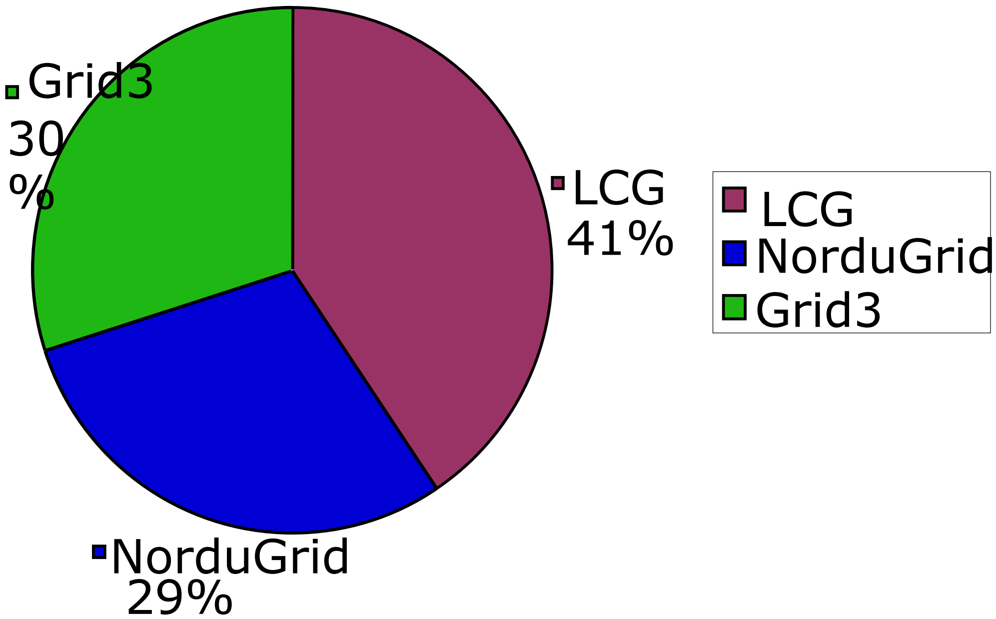
- DC2 is a three-part operation:
  - part I: production of simulated data (July-September 2004)
    - running on 3 Grids, worldwide
  - part II: test of Tier-0 operation (November-December 2004)
    - Do in 10 days what "should" be done in 1 day when real data-taking start
    - Input is "Raw Data" like
    - output (ESD+AOD) will be distributed to Tier-1s in real time for analysis
  - part III: test of distributed analysis on the Grid
    - access to event and non-event data from anywhere in the world both in organized and chaotic ways
- Requests
  - ~30 Physics channels (10 Million events)
  - Several millions of events for calibration (single particles and physics samples (di-jets))

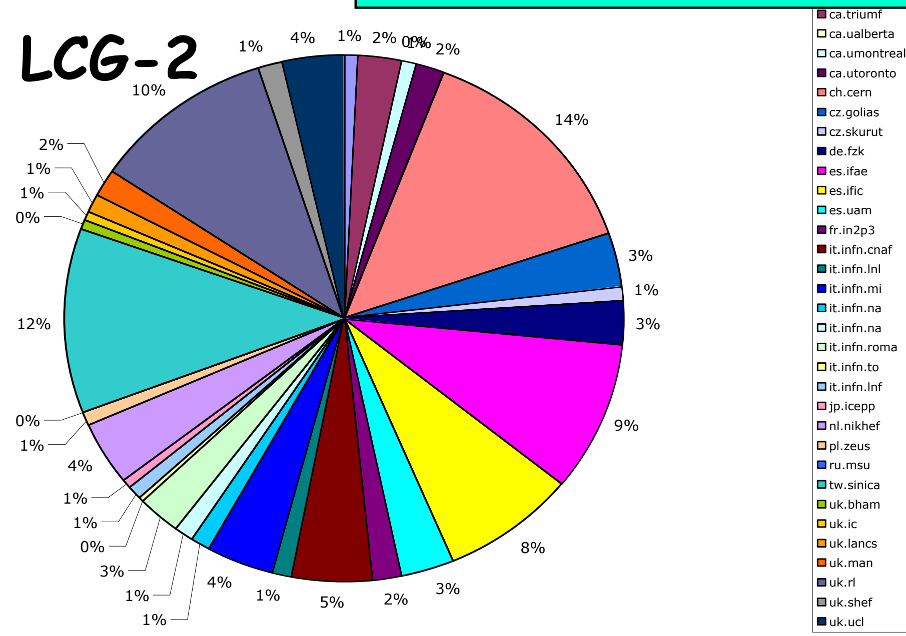
ATLAS: DC2 (2)

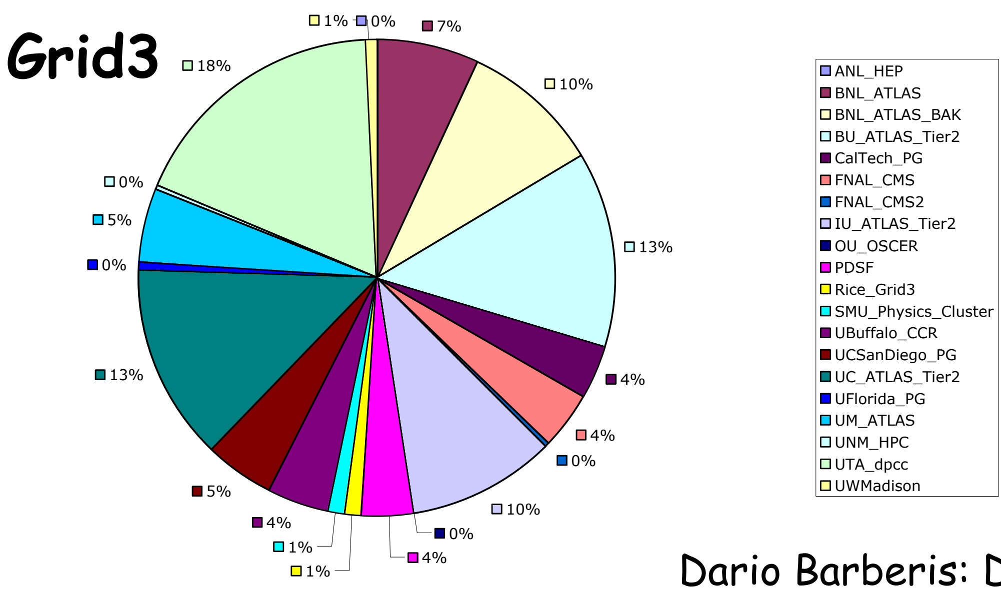

AMI
prodDB
dms
Don Quijote
Windmill
super
super
super
super
super
jabber
jabber
soap
soap
jabber
LCG exe
LCG exe
NG exe
G3 exe
LSF exe
Lexor
Dulcinea
Capone
RLS
RLS
RLS
LCG
NG
Grid3
LSF
Not used for DC2

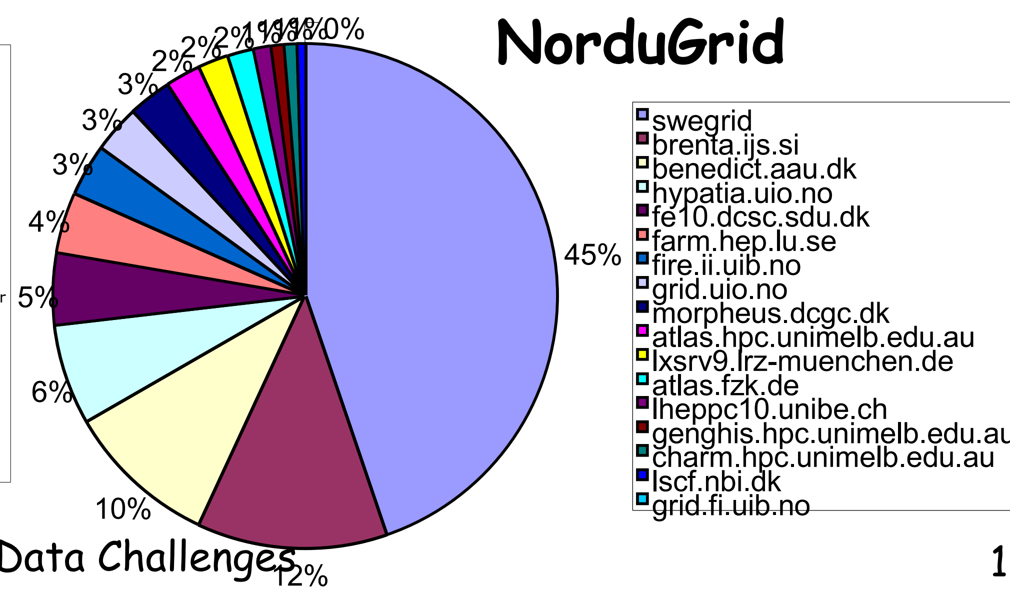
ATLAS: DC2 (3)

10 M Events
200 kSI2k-years
50 TB in POOL









General comments

- All experiments tried (and try) to use the LCG Grid and all other resources available to them
  - this fact will not change in the future
- ALICE and LHCb developed their own production systems and interfaced to the LCG-2 Grid through gateways
  - the whole of LCG-2 looked like a single, large Computing Element to ALICE
  - LHCb bypassed (or used in a special way) some of the critical components (Workload and Data Management)
- CMS ran before the full deployment of LCG-2 and concentrated on Data management
  - used pre-release LCG-0 for part of the simulation production in 2003
- ATLAS chose to use the 3 available Grids according to specs, developing only a higher-level job submission system
  - benefited, and suffered, accordingly!

Comments on performance

- As all Grid deployments are clearly in a prototype phase, inefficiencies are not unexpected
  - job success rates vary from 50% to 75% depending on the Grid and job type (and length)
  - it is difficult to imagine giving any system with a job success rate <<95% to 100's of physicists for analysis
- From here on I concentrate on the main sources of failures for LCG-2 (see GAG document in http://project-lcg-gag.web.cern.ch/project-lcg-gag/LCG_GAG_Docs/DCFeedBack.pdf):
  - experiment software installation and availability
  - site (mis)configuration
  - information system and monitoring
  - workload management system
  - data management

Experiment software installation

- Current practice is to have experiment software managers who are authorized to install software in dedicated areas and publish tags
- The lack of roles and priorities delays installation of new s/w versions wrt normal running jobs (installation jobs queue behind normal jobs)
- Frequent NFS failures, both at installation and running time, mostly at larger computing centres, make software unavailable to worker nodes (causing job failures)
  - this points also to general site management problems

Site configuration, IS and monitoring

- Site misconfiguration was responsible for a large number of job failures
- The information published through the Information System may not reflect reality at all times
  - the system is clearly not robust as human errors are possible, and indeed likely, and can be repeated in time
- NFS crashes and other communication problems are not detected by any automatic system
  - they can cause “black holes” for jobs
- Pro-active monitoring of the system as a whole was very basic as we started the DC’s
  - the GOCs start becoming operational only now
  - it is still not clear whose task it is to find out what goes wrong and fix it BEFORE we report massive job failures at a given site

Workload Management System

- Job submission time through the Resource Broker is very slow (typically 20 seconds/job for ATLAS)
  - this limits considerably the job throughput
  - no bulk operation is possible
  - sometimes job submission fails altogether (the RB rejects the job when it is too busy)
- Site ranking for job distribution based on too few parameters
  - jobs may end up queuing at a site that has free CPUs (but not for the right experiment) rather than going to another site
  - one work-around was the creation of VO-specific queues in each computing centre: this will not scale!
- Job distribution is very uneven, consecutive jobs tend to go to the same site as the info from the IS is not updated in real time
- The WMS can lose control of a job (declare it as "done" or "deleted" incorrectly) or just forget it altogether
- Lack of normalized CPU units means that jobs may go to wrong queues

Data Management System

- Many job failures were due to:
  1) failure to get input files (jobs killed manually after long wait time)
  2) failure to store output files
  3) failure to register output files
  4) correctly registered output files but data are corrupted during transfer
- All above conditions lead to considerable CPU time loss
- Reliable File Transfer systems could (should) fix most of the faults
- Underlying problem is the frequent loss of communication between processes running in remote installations

Final comments on Grids (1)

- So far only complaints…
  - it is easy to focus on items that cause trouble and forget the global results that have been nevertheless achieved
- In reality we all did manage to run productions of considerable size on Grid systems
- I do not think this amount of productions would have been possible otherwise
  - example of manpower difference:
    - **ATLAS DC1 in 2002 ran on non-Grid European sites with one production manager per site (for 3 months for the bulk Geant3 simulation)**
    - **ATLAS DC2 in 2004 ran on LCG sites (more sites than for DC1) with 4-5 people for the central operation, plus the LCG support team**
- On the other hand, most of the experiments got to the start of their DC exercise with only partially tested software
  - which did not make life easier when trying to understand the origin of failures

Final comments on Grids (2)

- Progress that was made on the LCG2 middleware this year was due mostly to the very cooperative attitude of the Grid Deployment team
  - unfortunately much less to the cooperation of the people who had developed it
- This situation should not be repeated with gLite/EGEE:
  - developers have to be exposed to feedback and work together with the users and the GD group

Final comments on Grids (3)

- We should perhaps move the focus from adding new features to making the systems more reliable
  - i.e.: my job may take longer to run but it will run and produce an output that goes to the correct place and gets catalogued
- On the Grid Middleware side, a lot of work was done during this year
  - many bug fixes were introduced during the summer
  - most causes of general job failures are at least understood, fixes for some of them are forthcoming
    - **more details in other talks in this session**
  - a lot was learned on the best way to configure our own production systems and to use the middleware available now
- Now we need stability and controlled evolution of the middleware
  - with the introduction of necessary improvements, but no upheaval

LCG-2.x vs gLite

- gLite development (mainly funded through EGEE) will lead to public releases relatively soon
  - current prototype still different from what is described in architecture and design documents
- It will be tested on testbeds of increasing size and complexity
- In the meantime, urgent fixes are needed for the LCG-2 system (the GD group at CERN is working on those)
  - some of the tools developed now are independent of Grid m/w
- All experiments support a transition to gLite-based m/w after appropriate testing and deployment of all components
- One thing to be avoided is the proliferation of Grid flavours:
  - we could not really cope with 3 this year, we do not want to have to support directly 4 next year!

My own comment on the number of Grids

- ATLAS is running on 3 Grids (LCG-2, NorduGrid and Grid3) with a high-level automatic job submission system
  - it turned out to be a much more manpower-intensive operation than anticipated
  - also for continuous (post-DC2) productions, we need to have production managers for each Grid flavour
- In reality, ATLAS used (uses) 4 Grids:
  - in Canada, the internal Grid (GridCanada) was interfaced to LCG-2 through a gateway at TRIUMF
    - **Canadian resources appear to LCG-2 as if they were concentrated at TRIUMF**
    - **internal configuration and middleware can differ from LCG-2**
    - **on the other hand, this gateway is not yet bi-directional**
    - **people in Canada do not yet see the whole of the LCG-2 Grid as if the resources were all located at TRIUMF: more work is needed**

My own comment on the number of Grids

- The number of Grids each experiment has to use is determined by the availability of resources
  - we have to use all the resources that are made available to our experiments
    - **for sure we will saturate any offered capacity as soon as we will start taking data**
  - we cannot dictate which middleware university computing centres or national/regional organizations will install
  - but we can ask that whatever they install conforms to a given set of interfaces and provides a given functionality
- In parallel with the deployment and support of one middleware flavour, we suggest that the LCG Project works towards
  - the definition of appropriate general interfaces to Grid systems
  - helping implementing them to make national/regional Grid systems available to LHC experiments