Grid Deployment and Regional Centres

Dominique Boutigny
LAPP – CNRS/IN2P3

LCG Comprehensive Review

Referees: Dominique Boutigny, Vladimir Kekelidze, Francesco Forti, Volker Guelzow, Patricia McBride

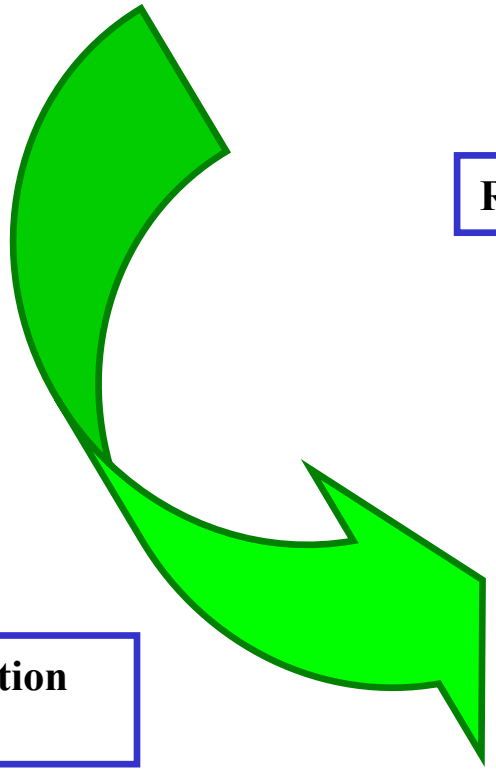
LHCC Meeting

November 25, 2004

References

- Material for this presentation has been taken from:
  - Ian Bird
    - "Introduction and Overview"
  - Markus Schulz
    - "Operational Experience and Status"
  - Dario Barberis
    - "Summary of Experiment Experiences in the Data Challenges "
  - Ian Bird
    - "Responses to data challenges and lessons learned "

The Grid deployment cycle



Strategic decision on a set of individual components

Component certification

Component integration

Extensive local testing

Packaging and installation procedure

Deployment in a few pilot sites

Pre-production service

Extensive certification

Large scale deployment

Service open to experiments' data challenges

Results and reviews

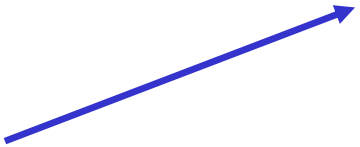
Cycle timing

- Jan 2003 GDB agreed to take VDT and EDG components
- September 2003 LCG-1
  - **Extensive certification process**
  - **Integrated 32 sites ~300 CPUs first use for production**
- December 2003 LCG-2
  - **Deployed in January to 8 core sites**
  - **Introduced a pre-production service for the experiments**
  - **Alternative packaging (tool based and generic installation guides)**
- Mai 2004 -> now monthly incremental releases (not all distributed)
  - **Driven by the experiences from the data challenges**
  - **Balance between stable operation and improved versions (driven by users)**
  - **2-1-0, 2-1-1, 2-2-0, (2-3-0)**
  - **(Production services RBs + BDIIs patched on demand)**
  - **> 80 sites (3-5 failed)**

Taken from M. Schulz

Key points

- Strategic choice of components at the beginning
  - Extensive testing and validation before deployment
  - Fast and easy feedback from experiments and remote sites
  - Constant communication channel between users and developers
  - Operation and User support



Include experiments in the loop as early as possible and at every stage

Developers should be committed to help in fixing problems at all stages of the deployment and production process

Installation in remote sites

Concern from the 2003 review: "Installation is too complex"

➔ **Installation complexity was almost exponential with the remote site size**

➔ **Relatively easy to install Grid software in a dedicated farm built from scratch**

➔ **Very difficult to deploy in an already large running computing center**

- **Compatibility with other software components**
- **Security**
- **Lack of flexibility of the installation tool (LCFG-ng)**
- **etc…**

Was a crucial problem for Tier-1s

➔ **The situation is much better today and is still improving**

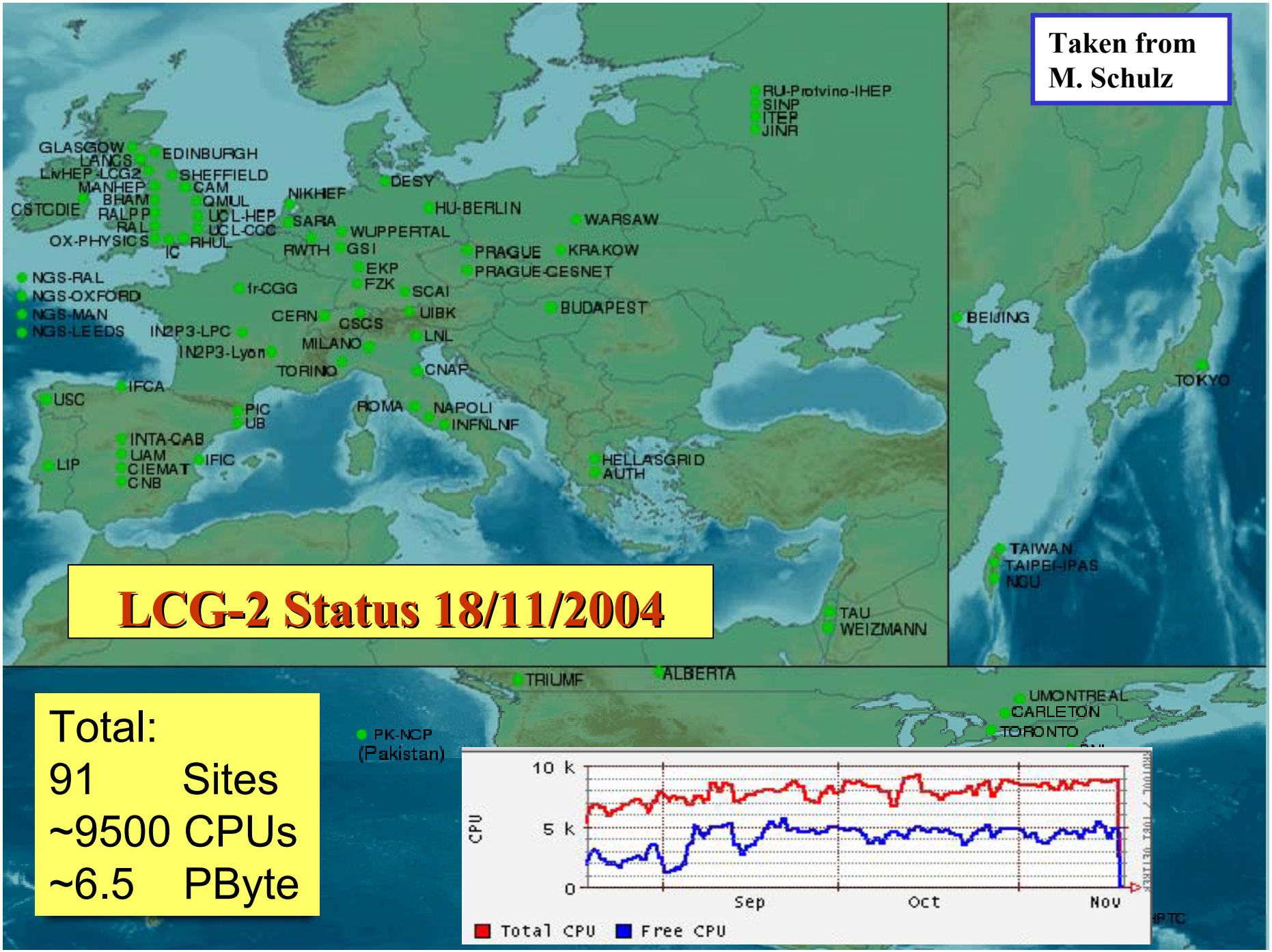
Adding a new site

- ❑ Software installation at remote sites and configuration
- ❑ Registration ➔ Virtual Organization - Grid Operation Center …
- ❑ Site certification
- ❑ Publication in the information system
- ❑ Site monitoring

➔ This procedure is now becoming routine

Taken from M. Schulz

Success: 80+ times – Failure: 3-5 times


Taken from
M. Schulz
RU-Protvino-IHEP
SINP
ITEP
JINR
GLASGOW
EDINBURGH
LANCS
LivHEP-LCG2
SHEFFIELD
MANHEP
CAM
BHAM
QMUL
CSTCDIE
RALPP
UCL-HEP
RAL
UCL-CCC
OX-PHYSICS
RHUL
IC
NIKHEF
DESY
HU-BERLIN
SARA
WARSAW
WUPPERTAL
RWTH
GSI
PRAGUE
KRAKOW
EKP
PRAGUE-CESNET
NGS-RAL
NGS-OXFORD
NGS-MAN
NGS-LEEDS
fr-CGG
FZK
SCAI
CERN
UIBK
BUDAPEST
CSCS
IN2P3-LPC
LNL
MILANO
IN2P3-Lyon
TORINO
CNAF
IFCA
USC
PIC
UB
ROMA
NAPOLI
INFNLNF
INTA-CAB
UAM
LIP
IFIC
CIEMAT
CNB
HELLASGRID
AUTH
BEIJING
TOKYO
TAIWAN
TAIPEI-IPAS
NCU
TAU
WEIZMANN
TRIUMF
ALBERTA
UMONTREAL
CARLETON
TORONTO
PK-NCP
(Pakistan)
LCG-2 Status 18/11/2004
Total:
91 Sites
~9500 CPUs
~6.5 PByte
10 k
5 k
0
CPU
Sep
Oct
Nov
Total CPU
Free CPU

Monitoring

In such a complex system, the monitoring is crucial

Taken from M. Schulz

TestZone tests reports for ce1.egee.fr.cgg.com

History of results for site: ce1.egee.fr.cgg.com

Colours definition

Job list match failed	#ffcc39
Replica Management failed	#cc3cff
OK	#99ff99
Test job still waiting for execution	#ffff33
Job Submission failed (Job Manager)	#cc3c00
Wrong LCG version (too old)	#c0c0c0

Test date	Version	Software Version	BrokerInfo	CSH test	BDII LDAP (RM)	PrintInfo	CopyAndReg. WN -> defaultSE	Copy defaultSE -> WN	Replicate defaultSE to castorgrid	3rd Party Rep. castorgrid to defaultSE	3rd Party cp castorgrid to WN	Delete Replica from defaultSE	GFAL infosys	lcg-cr -> defaultSE	lc de
2004-09-25 07:05:02	LCG-2_2_0	LCG-2_2_0	OK	OK	ldap://lxn1189.cern.ch:2170	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
2004-09-24 11:48:50	LCG-2_2_0	LCG-2_2_0	OK	OK	ldap://ce1.private.egee.fr.cgg.com:2135	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	OK	OK	OK
2004-09-24 07:05:10	LCG-2_2_0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/
2004-09-23 07:05:55	LCG-2_2_0	LCG-2_2_0	OK	OK	ldap://ce1.private.egee.fr.cgg.com:2135	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	OK	OK	OK
2004-09-22 07:05:33	LCG-2_2_0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/
2004-09-21 14:19:56	LCG-2_2_0	LCG-2_2_0	OK	OK	ldap://ce1.private.egee.fr.cgg.com:2135	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	FAILED	OK	OK	OK
2004-09-21 07:05:52	LCG-2_2_0	LCG-2_2_0	OK	OK	ldap://lxn1189.cern.ch:2170	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK
2004-09-20 07:05:29	FAILED	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/
2004-09-19															

Off-site problems are traced using a problem tracking tool – Daily remote site re-certification

A large effort has been undertaken to develop the necessary monitoring and site probing tools

Operational experience from Data Challenge

- Real production is different from certification tests
  - Usage characteristics of real production can't be simulated
- Time matters
  - Delays in support and operation are deadly during DCs
  - Several iterations needed to get it right
  - Communication between sites, operations, and experiments matters
    - not all players handled DCs with same priority (communication problem)

Taken from M. Schulz

The LCG-2 grid is large enough to see most of the operational problems

LCG-2 middleware has been improved continuously during operation

➔ Now reach a good level of stability

The whole grid has been successfully operated for months

➔ But still suffer from imperfections

➔ Some are related to fundamental architecture problems and will hopefully be addressed by EGEE future middleware

Global job efficiency is ~50-75%

Manageable for production type activity
Incompatible with analysis type activity

Mainly from D. Barberis

Data Challenge Overview (1)

The 4 LHC experiments were very active in DC but with different approaches

CMS

75 M events
96 TB in POOL

25 Hz reconstruction in Tier 0

Quasi real time DST analysis in Tier 1

Some tests of DST analysis in Tier 2

Focus on data flow management

Achieved 20 minute latency from Tier-0 Reco to job launch in Tier 1/2

Significant performance problems seen in Catalog Service and Replica Management

ALICE

Test and validation of the full computing model on 10% of a data sample collected in 1 year

Whole system running with ALIEN

LCG resources accessed through and ALIEN-LCG interface

Data Challenge Overview (2)

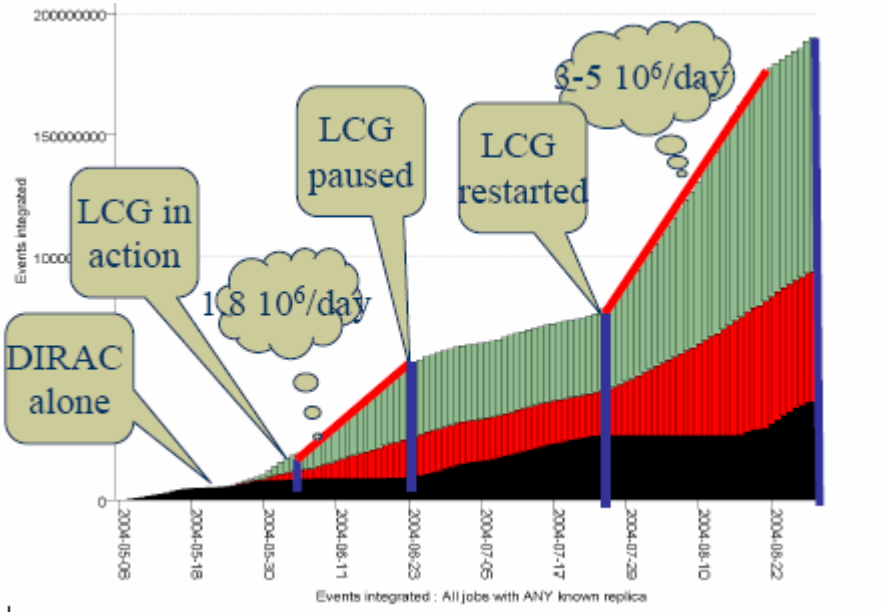
LHCb

3 phases:

- **MC Production (done)**
- **Pre-selection (to start soon)**
- **Analysis (in preparation)**

Some production done in DIRAC specific sites

Some in LCG-2 nodes through an LCG-DIRAC interface



ATLAS

10 M events
50 TB in POOL

Try to use the 3 available Grids: LCG-2, NorduGrid and Grid-3

Developed a high level job submission system to use all of them

Lessons learned from DC (1)

A huge amount of events have been processed on the LCG-2 Grid

➔ **Very Big SUCCESS !!!**

Main sources of failure:

- experiment software installation and availability
- site (mis)configuration
- information system and monitoring
- workload management system
- data management

Taken from D. Barberis

Workload Management System

Taken from D. Barberis

- 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

Taken from
D. Barberis

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

Lessons learned from DC (2)

As in many complex system, even if each component is working well individually, the problems appear in the interaction between the components

One cannot expect everything to be working well 100% of the time, the software should be fault tolerant

Most of the issues seen in Workload Management, Data Management and Information Systems will be addressed in the the next middleware generation (gLite)

➔ Should make sure that the current middleware problems are understood and actually addressed

The site mis-configuration problems which were a large source of job failure are addressed by improving the monitoring and site probing tools

➔ A real effort has been undertaken on this field and is starting to give positive results

It is very important to keep the current DC production grid up and running

- Continue to gain experience in operating this complex system
- Improve the diagnostic in problematic area in order to give a acute input to gLite developers

General comment

Given the complexity of the system

It may be a good idea to define a minimum set of basic functionalities to be achieved absolutely in order to preserve the LHC data quality

- **Devote the necessary effort and make priorities to implement this basis**
- **Focus on stability and reliability**
- **Add new functionalities only when they are absolutely necessary**
- **Delay what is not crucial**

Comments on using multiple Grids

Taken from D. Barberis

- 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

Taken from D. Barberis

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

Summary

An impressive achievement !

Proposed recommendations

Include experiments in the loop as early as possible and at every stage

Developers should be committed in fixing problems at all stages of the deployment and production process

Make sure that the current middleware problems are really understood and actually addressed in next generation

Keep the current LCG-2 grid up and running

Focus new developments on crucial issues

Work on interoperability of the grids and converge on a common interface