Calcolo distribuito nell'era di LHC

M. Paganoni
(Univ. Milano-Bicocca & INFN)

IFAE 2010
Roma, 7/4/2010

The LHC Computing Challenge

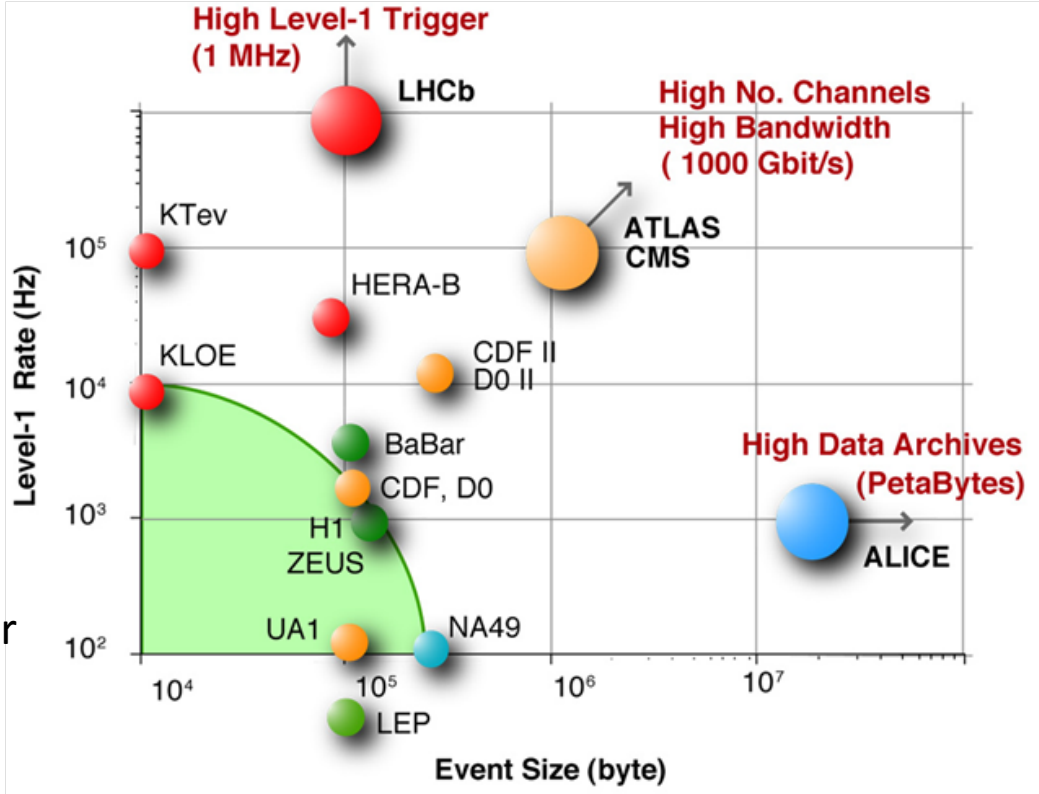
- Data volume
  - High rate * large number of channels * 4 experiments
  - Custody of the data for more than two decades

  ➔ **15 PetaBytes of new data / year**
- Compute power
  - Event complexity * Nb. events * thousands users

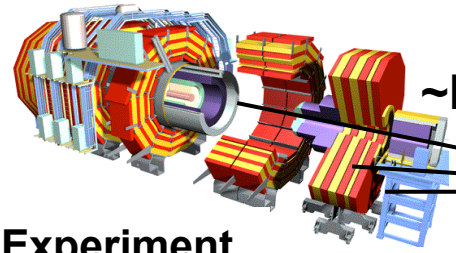
  ➔ **100 k of (today's) fastest CPUs**

  ➔ **45 PB of disk storage**
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient and coherent data access for analysis worldwide (8000 physicists in 50 countries)

➔ GRID technology



LHC Computing Hierarchy



Experiment

~PByte/sec

Online System

~100-1500 MBytes/sec

CERN/Outside ~ 1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~ 1:1:1

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

Tier 1

10 Gbps

IN2P3 Center | RAL Center | INFN Center | FNAL Center | ...

2.5-10 Gbps

Tier 2

Tier2 Center | Center | nter | Center | Center

~2.5-10 Gbps

Tier 3

Institute | tute | stitute | Institute

Physics data cache

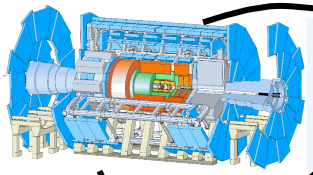
0.1 to 10 Gbps

Workstations

Tier 4

Tens of Petabytes by 2010.
An Exabyte ~5-7 Years later.

Global workflow



detector

event filter (selection & reconstruction)

reconstruction

raw data

event summary data

event reprocessing

processed data

batch physics analysis

analysis

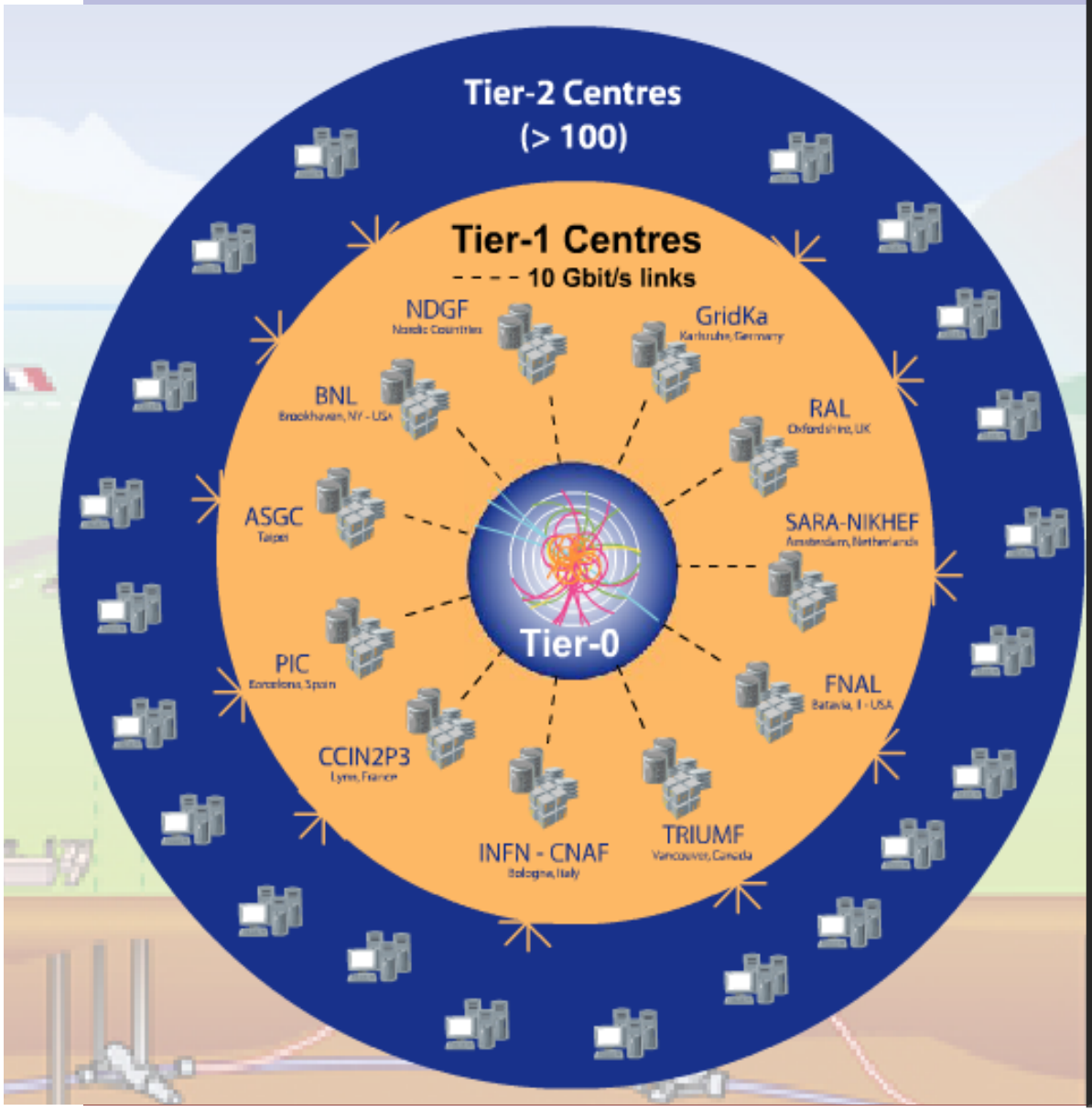
analysis objects (extracted by physics topic)

interactive physics analysis

event simulation

simulation

Data serving



Tier-0 (CERN):
- Data recording
- Prompt reconstruction
- Data distribution

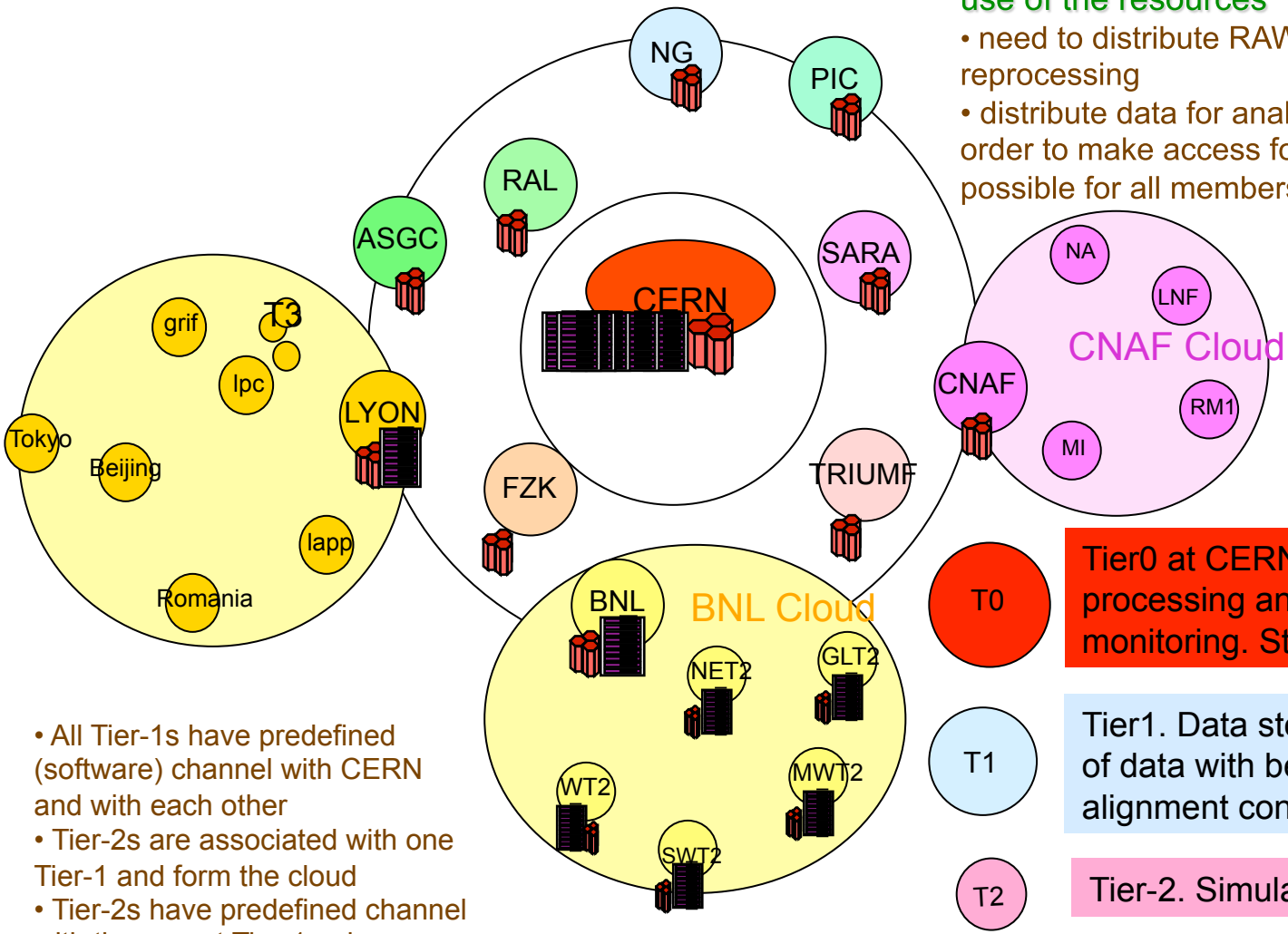
Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Skimmimg

Tier-2 (~130 centres):
- Simulation
- End-user analysis

ATLAS cloud model

Hierarchical Computing Model optimising the use of the resources

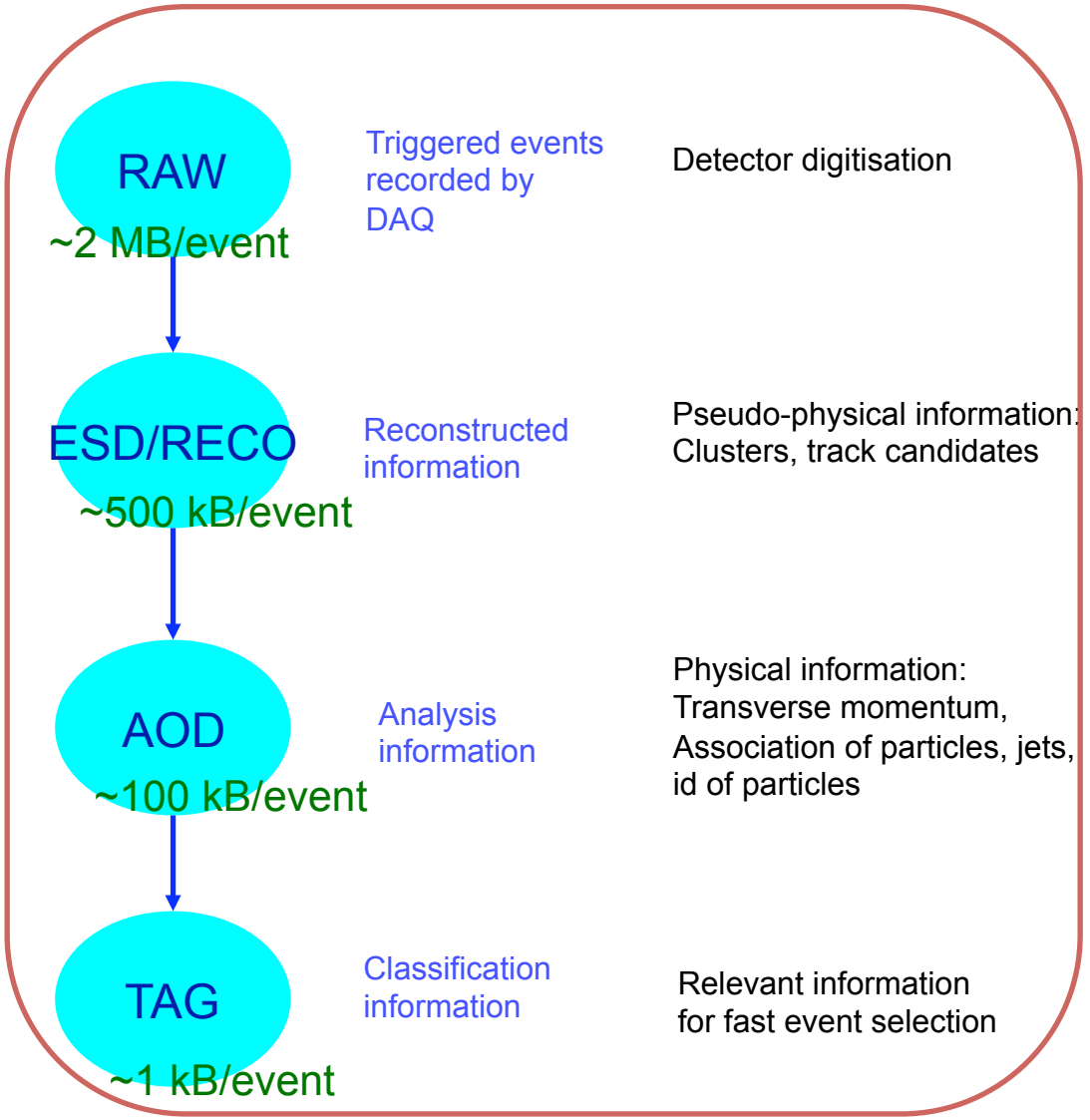
- need to distribute RAW data for storage and reprocessing
- distribute data for analysis in various formats in order to make access for analysis as easy as possible for all members of the collaboration



- All Tier-1s have predefined (software) channel with CERN and with each other
- Tier-2s are associated with one Tier-1 and form the cloud
- Tier-2s have predefined channel with the parent Tier-1 only.

T0	Tier0 at CERN. Immediate data processing and detector data quality monitoring. Stores on tape all data
T1	Tier1. Data storage and reprocessing of data with better calibration or alignment constants
T2	Tier-2. Simulation and user Analysis
T3	Tier-3. Non Grid user Analysis

Data Formats



Worldwide LHC Computing Grid

- ✓ A distributed computing infrastructure to provide the production and analysis environments for LHC
- ✓ Managed and operated by a worldwide collaboration between the experiments and the computer centres

A map of the worldwide LCG infrastructure operated by EGEE and OSG.

WLCG Collaboration Status
Tier 0; 11 Tier 1s; 64 Tier 2 federations
(124 Tier 2 sites)

Today we have 49 MoU signatories, representing 34 countries:

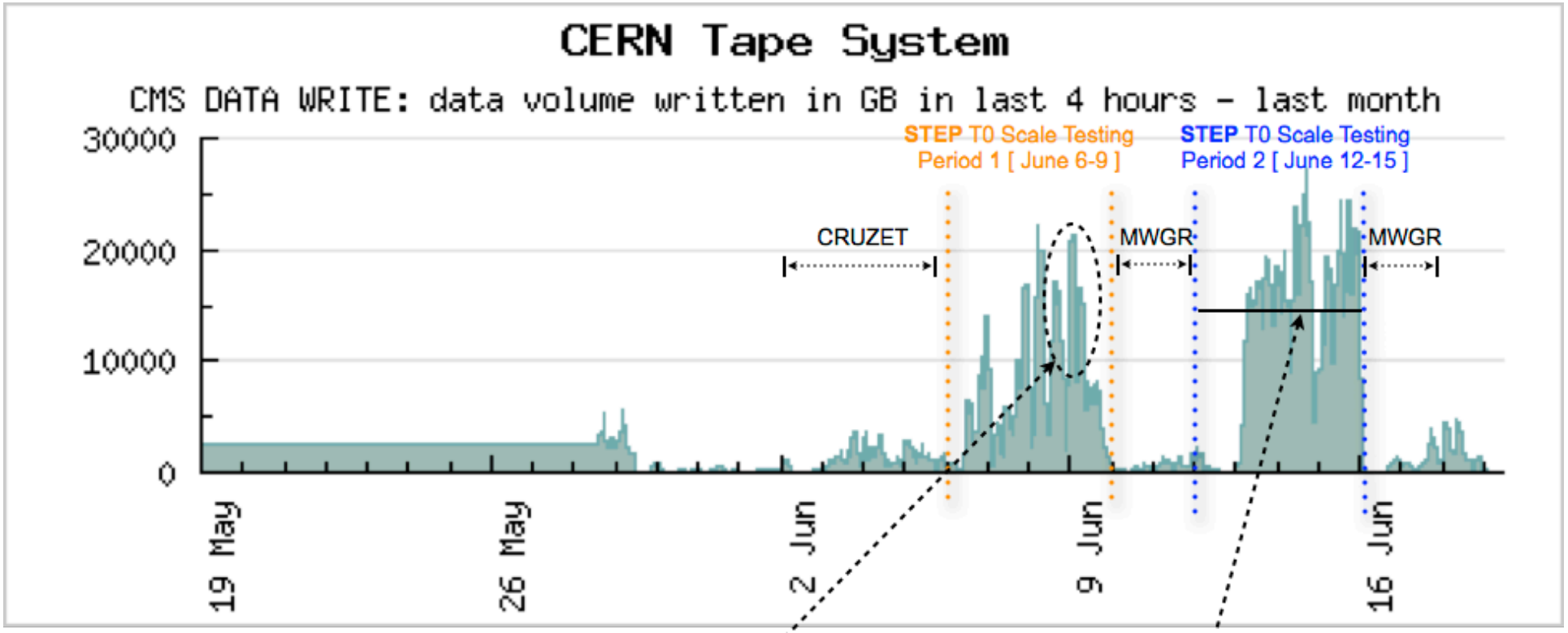
Australia, Austria, Belgium, Brazil, Canada, China, Czech Rep, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, India, Israel, Japan, Rep. Korea, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Russia, Slovenia, Spain, Sweden, Switzerland, Taipei, Turkey, UK, Ukraine, USA.

Preparation for LHC startup

- ✓ Since 2004 WLCG has been running a series of challenges with increasing targets for:
  - Data throughput
  - Workloads
  - Service availability and reliability
- ✓ Recent significant challenges
  - May 2008 – Combined Readiness Challenge
    - All 4 experiments running realistic work (simulating data taking)
    - Demonstrated that we were ready for real data
  - June 2009 – Scale Testing
    - Stress and scale testing of all workloads including massive analysis loads
- ✓ In essence the LHC Grid service has been running for several years
- ✓ Now we have moved from development to production

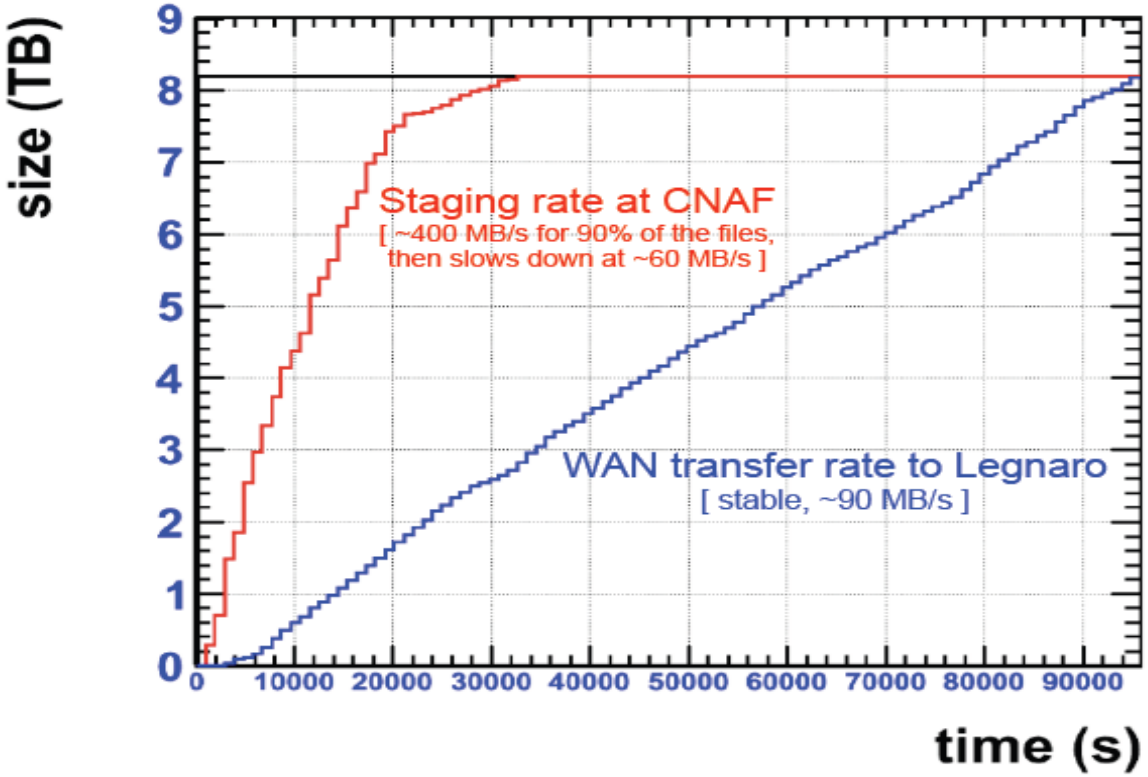
STEP09 and T0 (CMS)

- Can CMS archive the recorded RAW+RECO data on tape at T0 at sufficient rates (500 MB/s) ?
- Can this work when other VOs take data and write to tape?



STEP09 - data serving (CMS)

- ✓ T1 → T2 Data serving exercise: transfer from T1 tapes to T2s, i.e. put load on T1 tape-recall



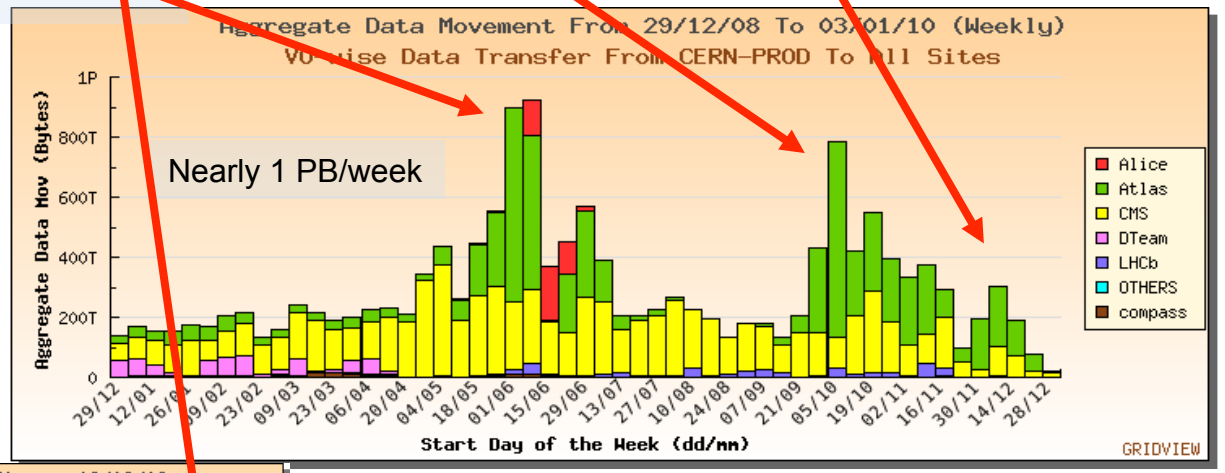
2009 Data Transfers

Final readiness test (STEP'09)

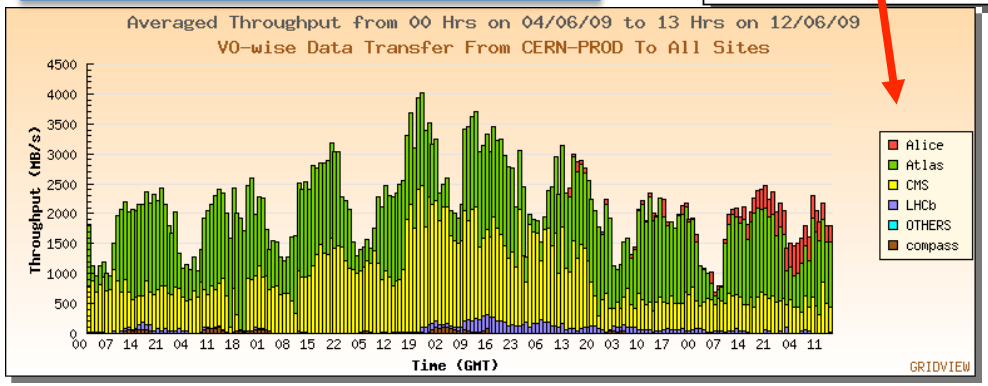
Preparation for LHC startup

LHC physics data

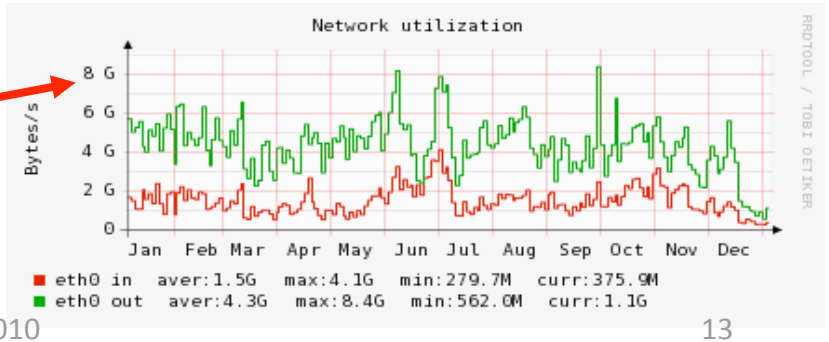
- ✓ Full experiment rate needed is 650 MB/s
- ✓ Desire capability to sustain twice that to allow for Tier 1 sites to shutdown and recover
- ✓ Have demonstrated far in excess of that, in a sustained way



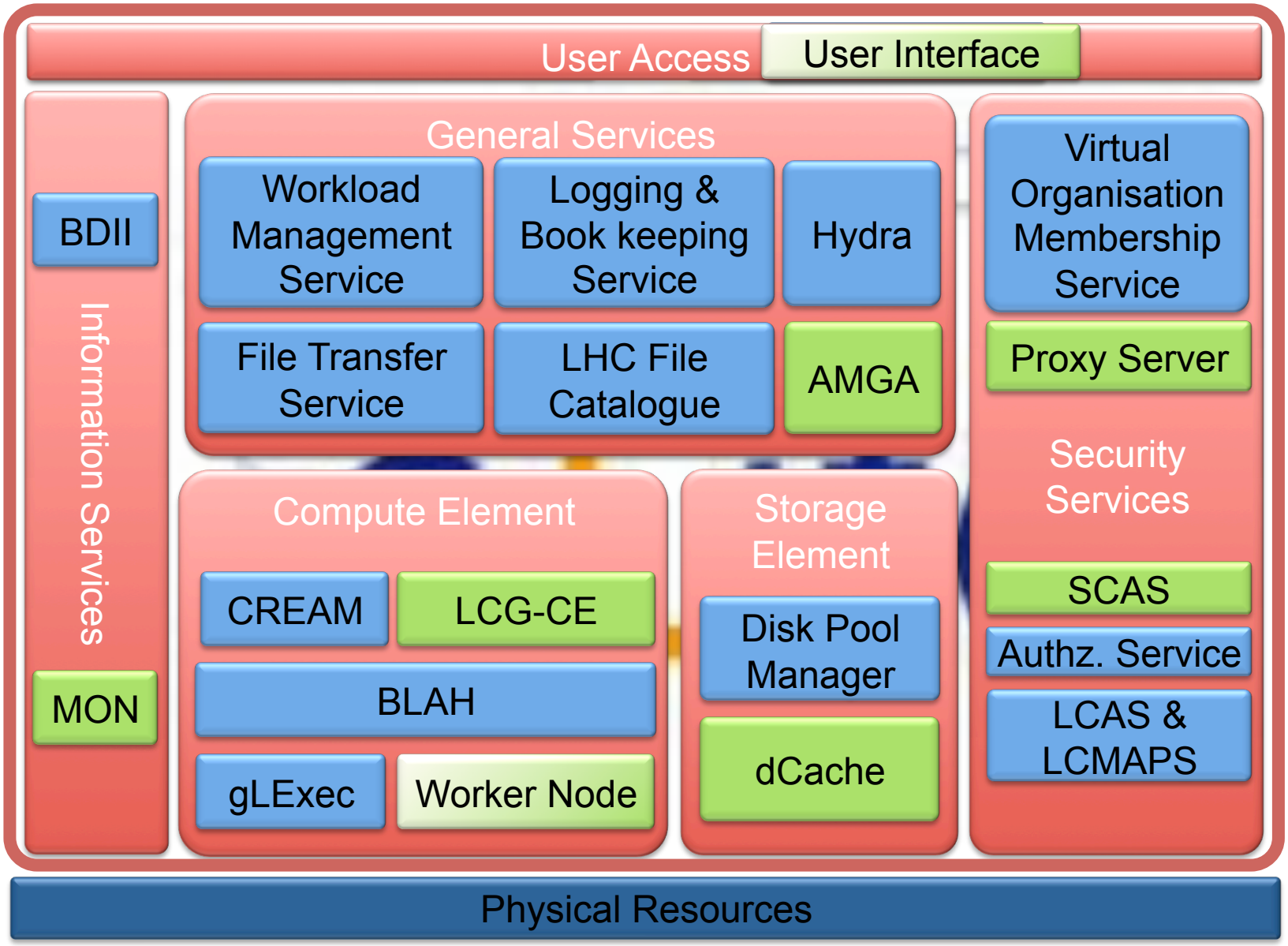
Nearly 1 PB/week



More than 8 GB/s peak transfers from Castor fileservers at CERN

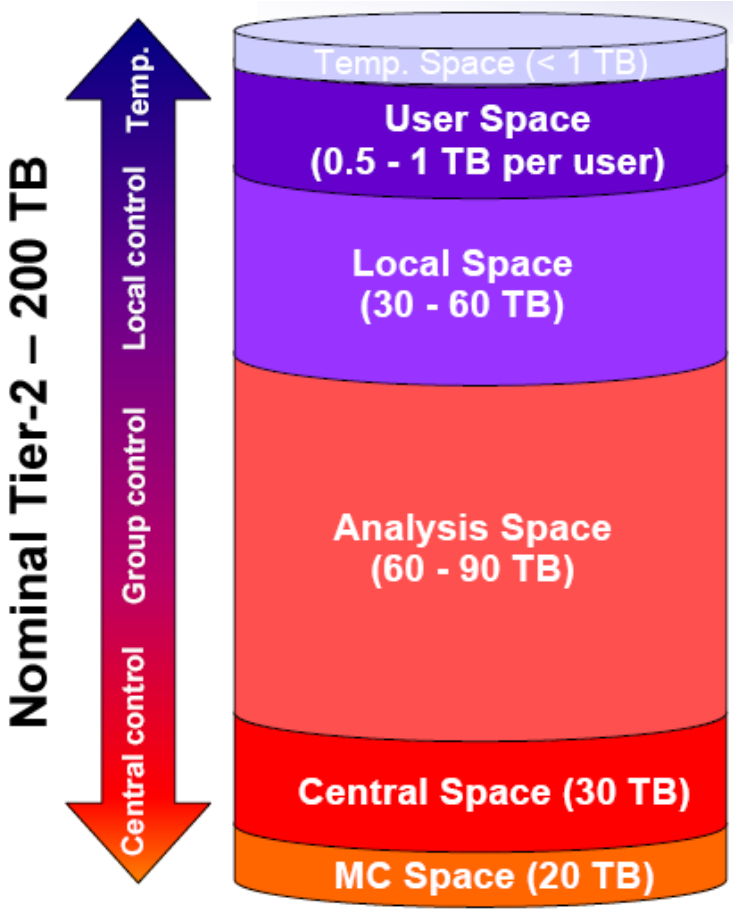


gLite Middleware



Physics analysis at Tier-2 (CMS)

- Hosting of user/group analysis
- Production of simulated events
- Currently 17 groups (physics analysis / detector performance) are associated with the Tier-2 and control 25 % of the resources:
  - space allocated
  - prioritization of jobs
- Current CMS total CPU at T2s:
  - 17k jobs slots
  - 50% for analysis

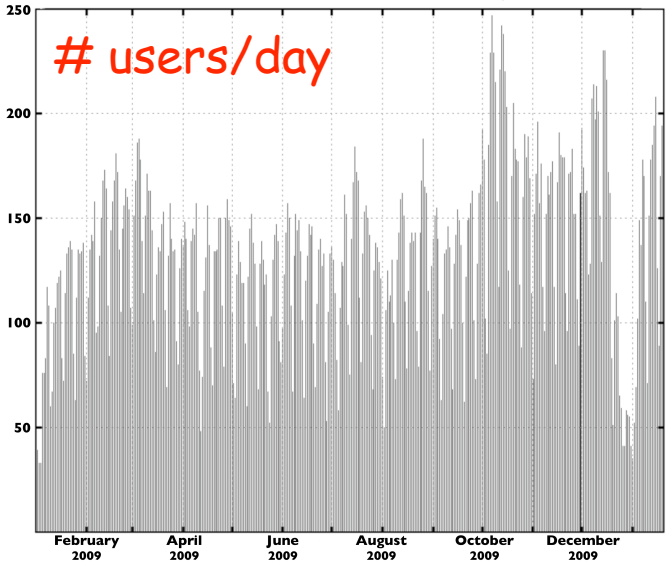


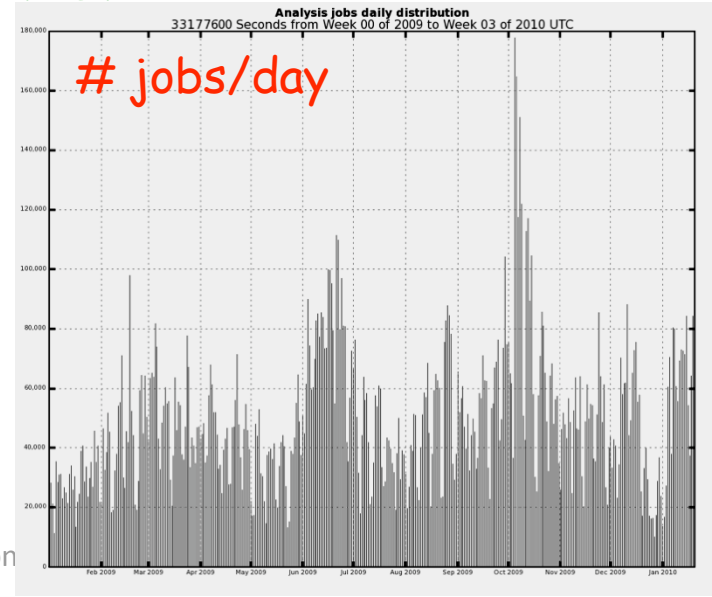
CRAB

- ✓ CMS Remote Analysis Builder:
  - Friendly interface for physics analysis on the Grid
  - CRAB takes care of the data discovery, ships the code , prepare and manage the jobs, retrieve the output
  - CRAB interacts with:

    the Data Placement system PhEDEx, the Data Bookkeeping system DBS, Grid flavours (OSG, EGEE, …) and local batch systems

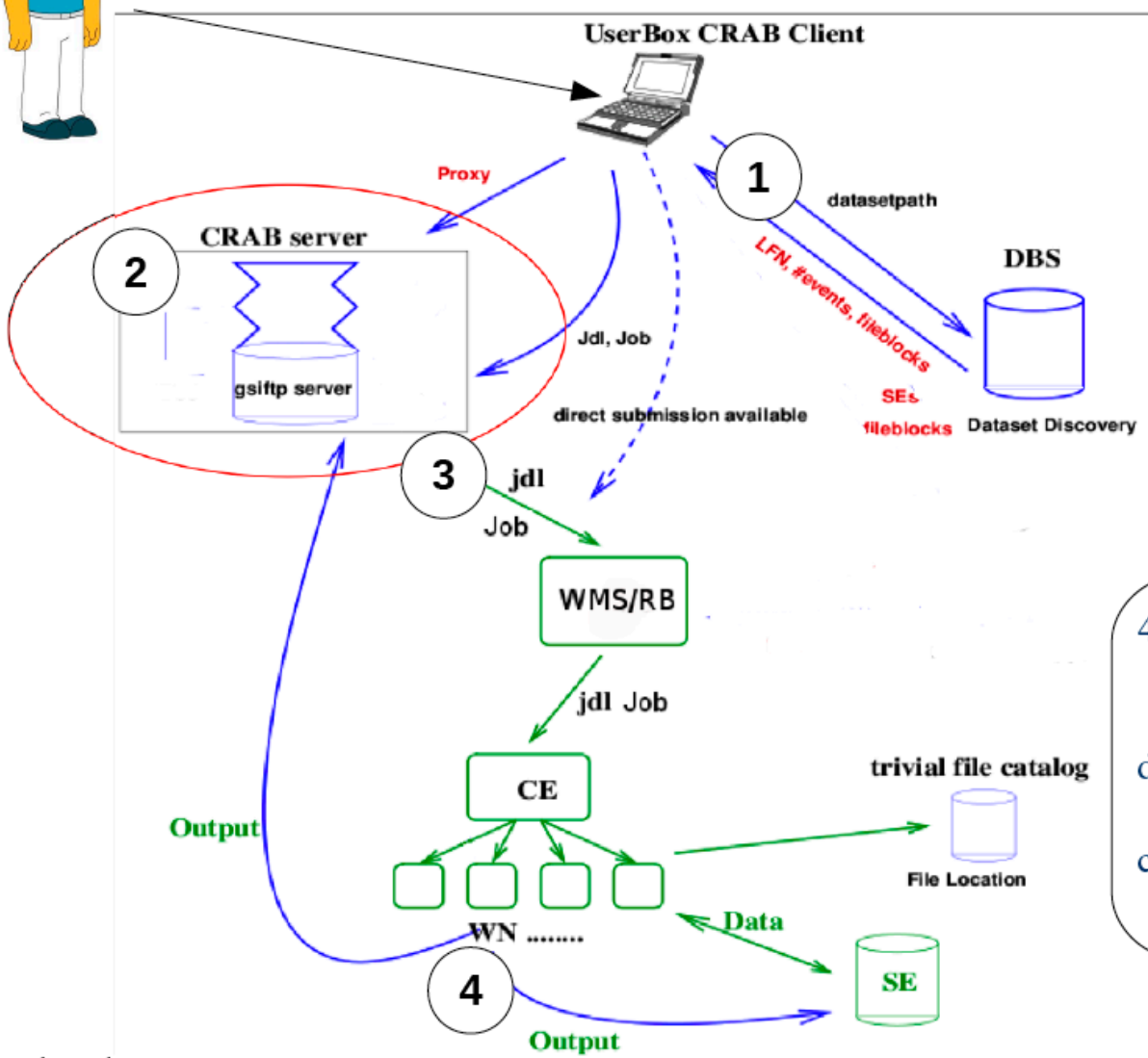
  - Scaling up to 200 kjob/day tested





CRAB Workflow

Write analysis code;
Select the dataset to be analyzed;
Decide job parameters (e.g. event number per job) and other CRAB configuration



1) Finds data information (physical files location and their content)

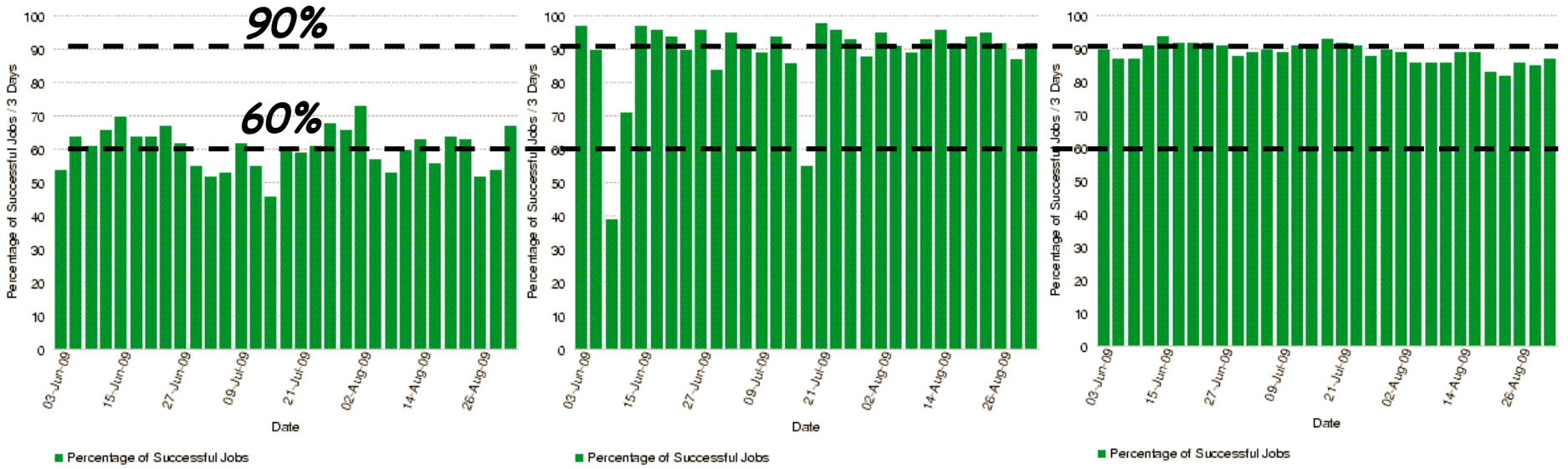
2) Splits the user task in many jobs, according to the user configuration and the data location Packs the user code

3) Submit jobs to the Grid: the data location is used by Grid Workload Management tools to match remote resources

4) through Grid tools:

- Track jobs status,
- Retrieve output or move it to a User defined SE,
- Allow more jobs action such as cancelling, resubmitting,
- Retrieve verbose log for job failures

Running on the Grid

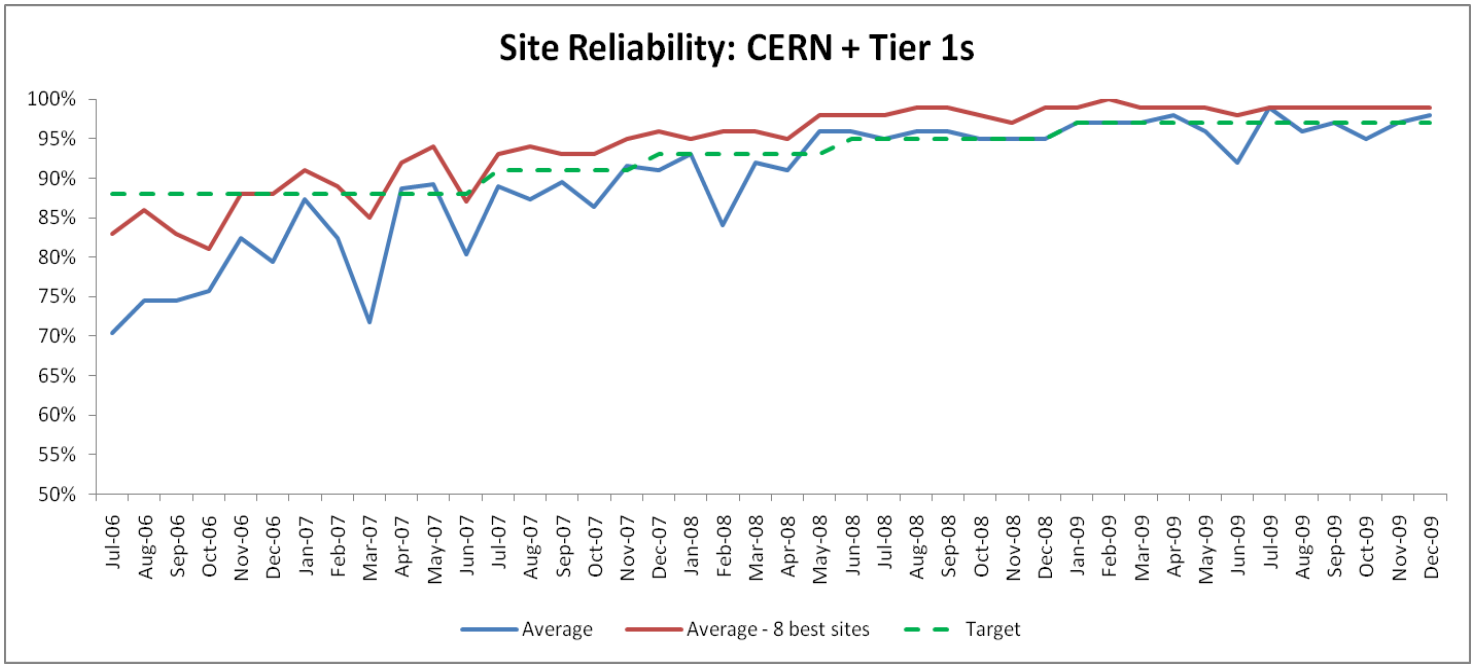


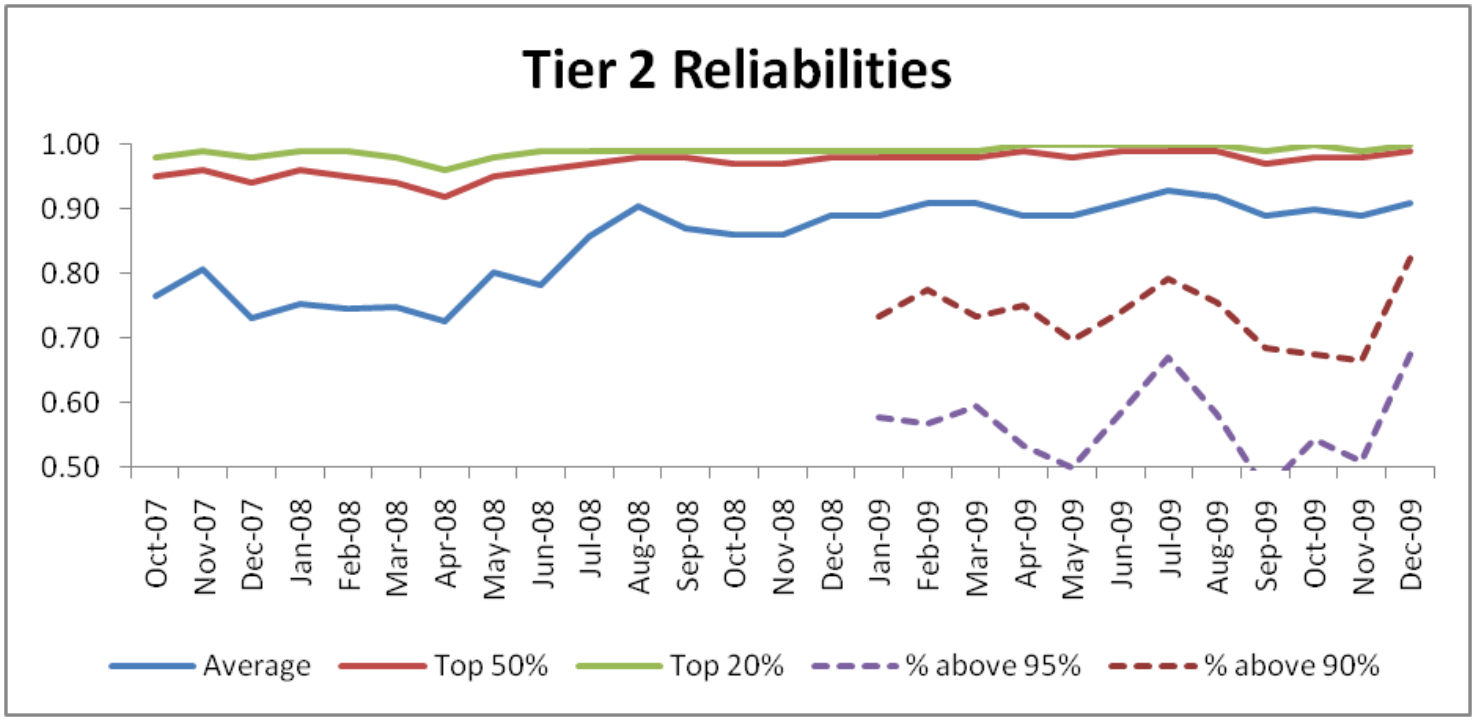
Analysis MC Production Job Robot

- ✓ For real analysis jobs, main failures are caused by application failures. Basic users cfg errors:
  - Output stageout issues
  - Data reading at hosting site

Smooth operations

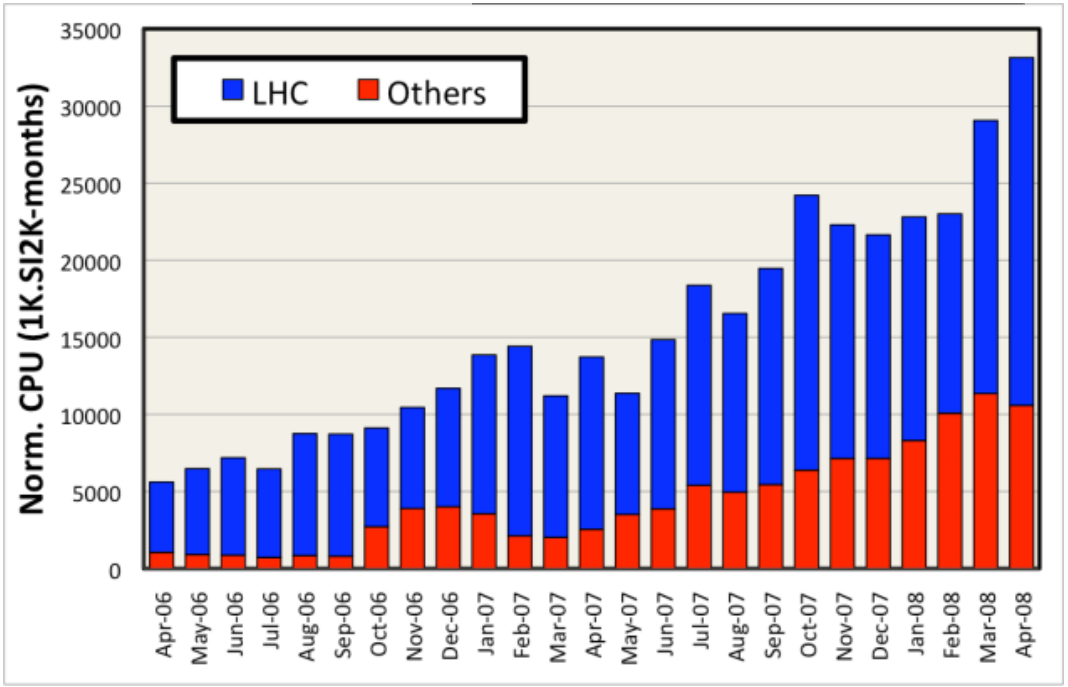
- ✓ Presently concentrating on tracking metrics for:
  - Performance
  - Reliability
  - Scalability
- ✓ Monitor site readiness and availability
  - Test all functionality required from experiments at each site in a continuous mode
  - Determine if the site is usable and stable, by testing:
    - Job submission
    - Local site configuration and software installation
    - Data access and data stage-out from batch node to storage
    - "Fake" analysis jobs
    - Quality of data transfers across sites
  - Site availability: fraction of time all functional tests in a site are successful
  - Job Robot efficiency: fraction of successful "fake" analysis jobs
  - Link quality: number of data transfer links with an acceptable quality between T1 and T2 centers


Site Reliability: CERN + Tier 1s
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
Jul-06
Aug-06
Sep-06
Oct-06
Nov-06
Dec-06
Jan-07
Feb-07
Mar-07
Apr-07
May-07
Jun-07
Jul-07
Aug-07
Sep-07
Oct-07
Nov-07
Dec-07
Jan-08
Feb-08
Mar-08
Apr-08
May-08
Jun-08
Jul-08
Aug-08
Sep-08
Oct-08
Nov-08
Dec-08
Jan-09
Feb-09
Mar-09
Apr-09
May-09
Jun-09
Jul-09
Aug-09
Sep-09
Oct-09
Nov-09
Dec-09
Average
Average - 8 best sites
Target

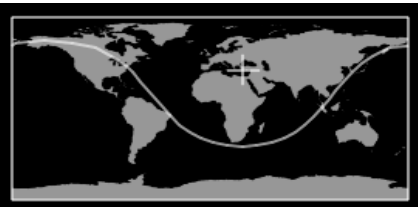

Tier 2 Reliabilities
1.00
0.90
0.80
0.70
0.60
0.50
Oct-07
Nov-07
Dec-07
Jan-08
Feb-08
Mar-08
Apr-08
May-08
Jun-08
Jul-08
Aug-08
Sep-08
Oct-08
Nov-08
Dec-08
Jan-09
Feb-09
Mar-09
Apr-09
May-09
Jun-09
Jul-09
Aug-09
Sep-09
Oct-09
Nov-09
Dec-09
Average
Top 50%
Top 20%
% above 95%
% above 90%

Impact of WLCG

- ✓ WLCG has been the driving force for the multiscience Grid EGEE, presently the largest Grid infrastructure worldwide
- ✓ Co-funded by the European Commission
  (Cost: ~170 M€ over 6 years, funded by EU ~100M€)
- ✓ EGEE already used for >100 applications in the fileds
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics
  - Earth Sciences
  - Fusion
  - High Energy Physics
  - Life Sciences



EGEE infrastructure



17000 users

136000 LCPUs (cores)

25PB disk

39PB tape

12 million jobs/month

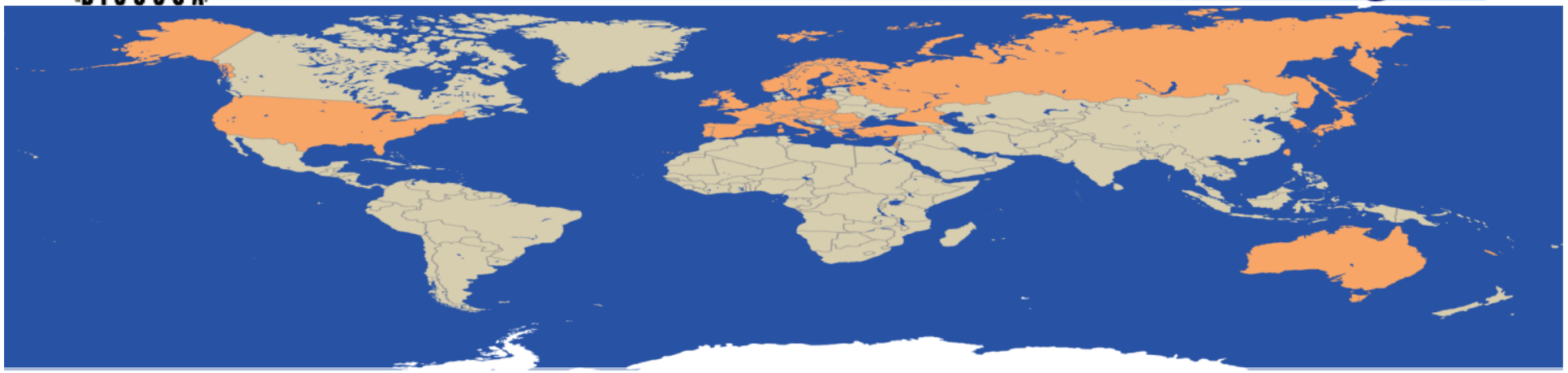
+45% in a year

268 sites, 162 VOs,

48 countries

Activity snapshot

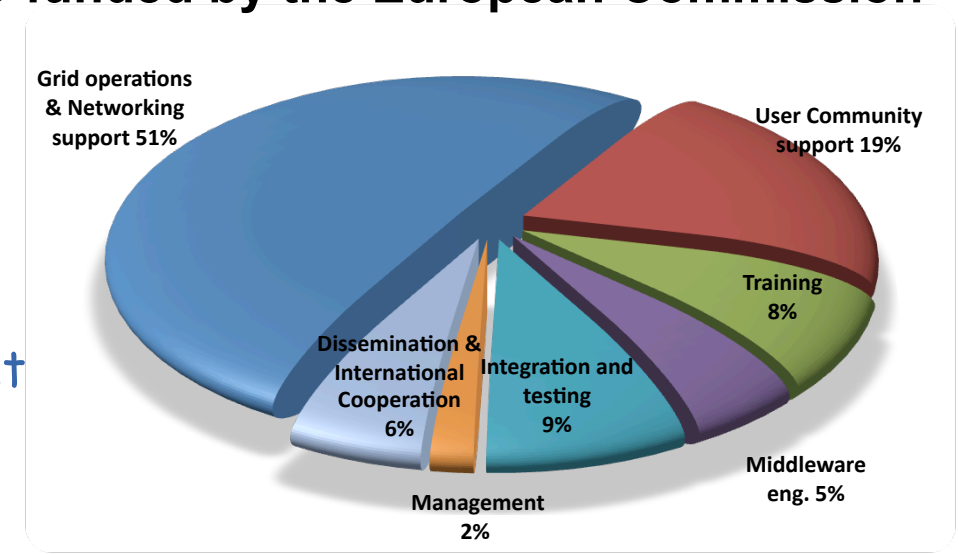
EGEE-III



Flagship Grid infrastructure project co-funded by the European Commission

Main Objectives

- Expand/optimise existing EGEE infrastructure, include more resources and user communities
- Prepare migration from a project based model to a sustainable federated infrastructure based on National Grid Initiatives



Duration: 2 years
Consortium: ~140 organisations across 33 countries
EC co-funding: 32Million €

National Grid Initiatives in Europe

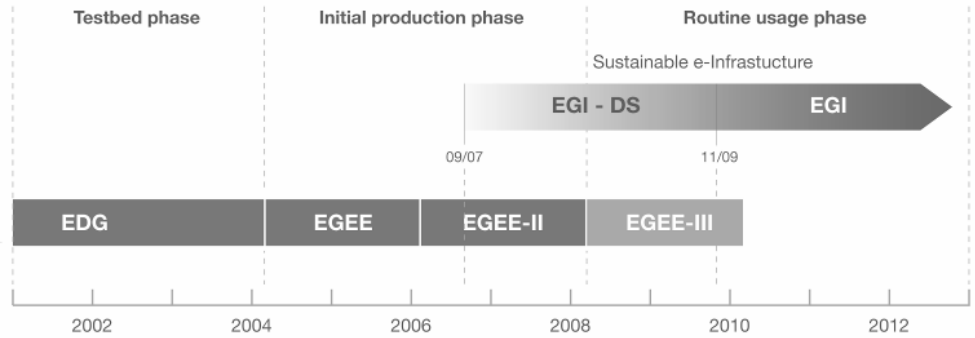
www.eu-egi.eu

Need to guarantee a worldwide robust and sustainable Grid infrastructure, by coordinating the National Grid Initiatives

22 Oct 2009:

33 NGIs + CERN, EMBL

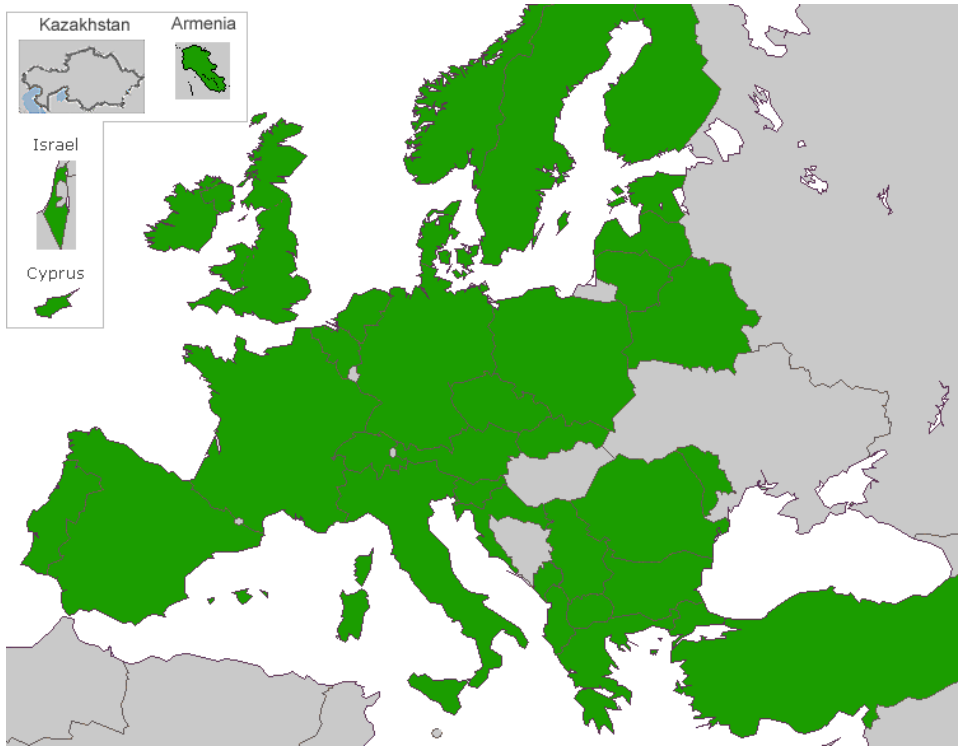
+ Observers



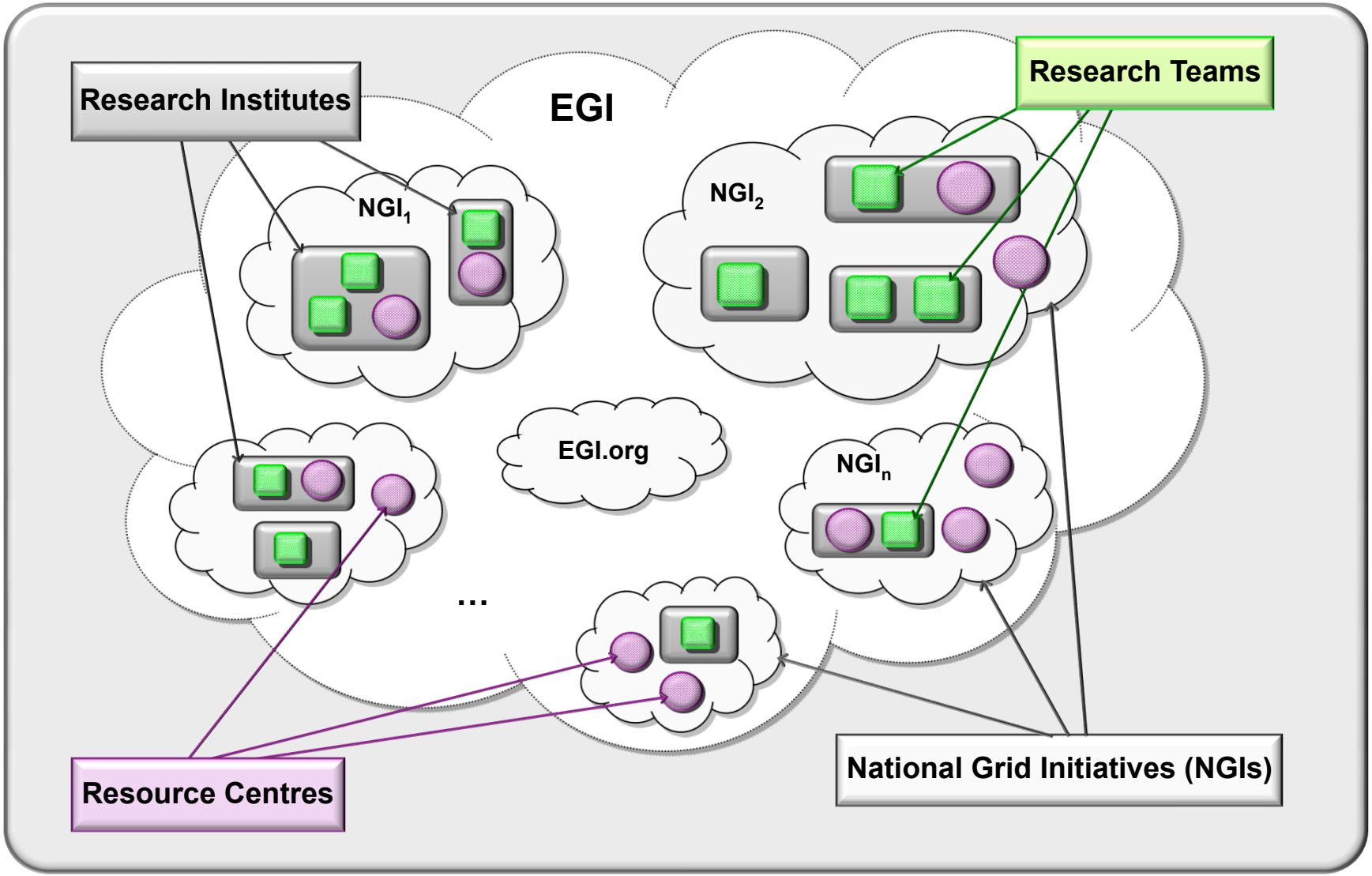
3 Feb 2010:

Agreement on the statutes

18 posts advertised to fill
manegement team in Amsterdam
www.eu-egi.eu



The EGI Actors



Conclusions

- ✓ Large experience acquired in the past 6 years with challenges and continuous operations on the WLGC Grid infrastructure
- ✓ Cosmic data taking have been a very useful exercise
- ✓ The next few years will see a continued evolution:
  - Virtualisation as a mechanism to improve the provision of grid services and to simplify application environments
  - Optimization of the experiment software for the multi-core architecture
- ✓ Computing is ready for LHC data taking:
  - Sustained data processing
  - Strong demand on site readiness
  - High demand on data accessibility by physicists