Enabling Grids for E-science in Europe

www.eu-egee.org

NA4 Applications

Marian Babik
Institute of Informatics, SAS
Dept. of parallel and distributed computing

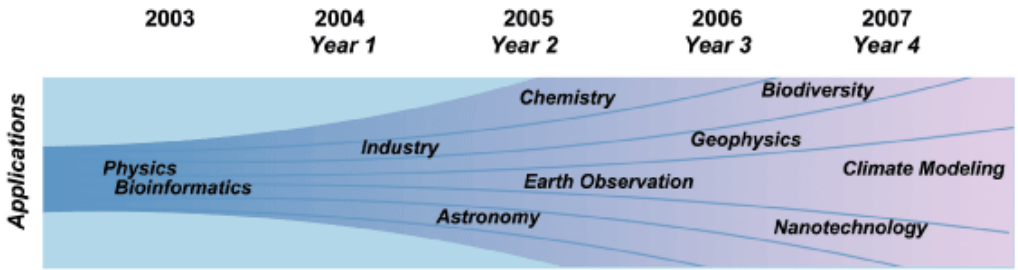
EGEE is a project funded by the European Union under contract IST-2003-508833

Talk Outline

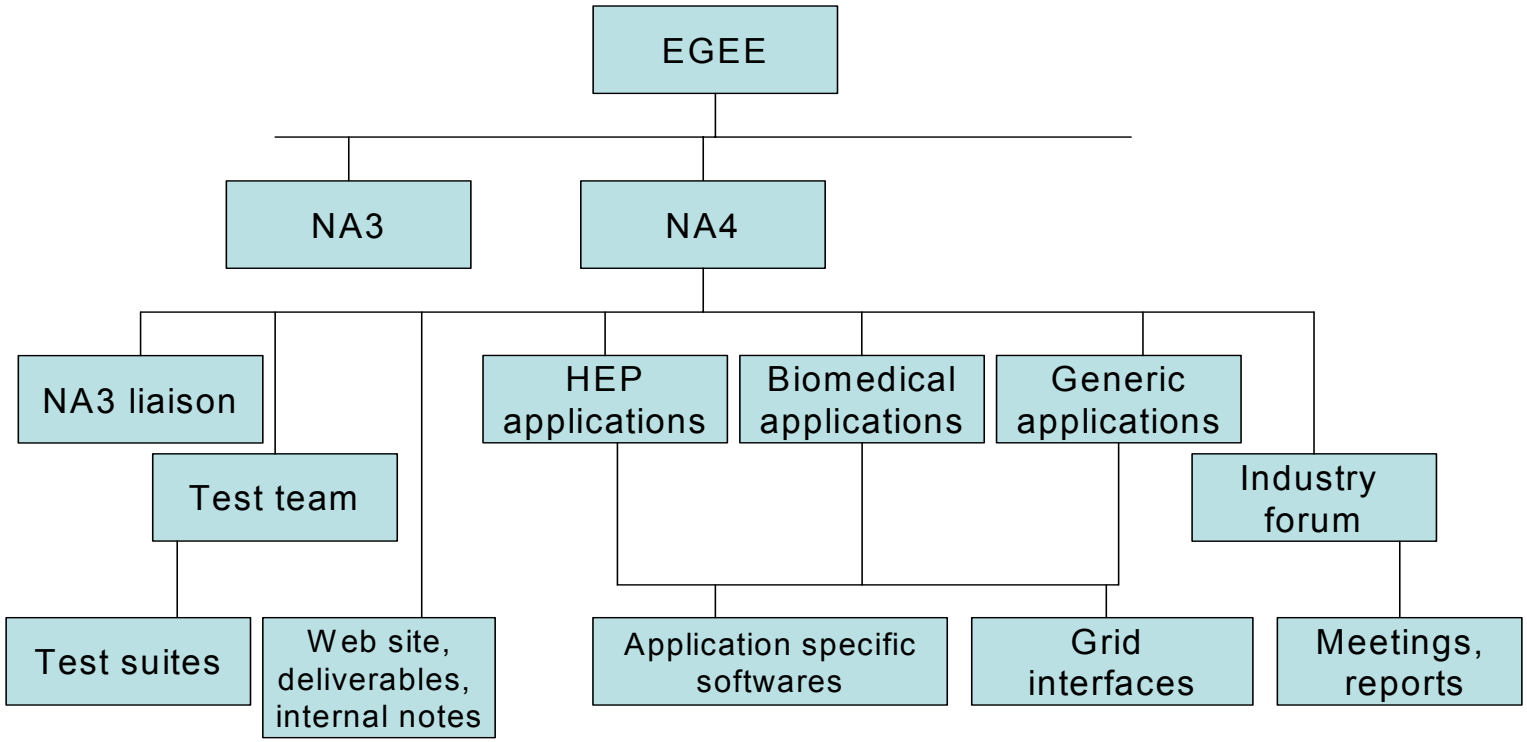
- **The basic goals of NA4**
- **The flavour of the work for the NA4 sub-groups**
  - *Biomedical applications*
  - *High Energy Physics*
  - *'Generic' applications*
  - *Astro-particle Physics*
  - *Computational Chemistry*
  - *Earth Science*
- **Concluding comments**

NA4: Identification and support of early-user and established applications on the EGEE infrastructure

- To identify through the dissemination partners and a well defined integration process a portfolio of early user applications from a broad range of application sectors from academia, industry and commerce.
- To support development and production use of all of these applications on the EGEE infrastructure and thereby establish a strong user base on which to build a broad EGEE user community.
- To initially focus on two well-defined application areas – Particle Physics and Life sciences, while developing a process for supporting other application areas



NA4 organisational structure



What do all applications expect from the grid?

- **Access to a world-wide virtual computing laboratory with vast resources**
- **Possibility to organise in VOs(virtual organisations) with members being given access rights according to their roles in the VO, and facilities to protect data**
- **Transparency in data and job management via easy to use application interfaces**
- **Definite added value in performance for both interactive and batch computation**

Biomedical Applications

- **Some key applications and their characteristics**
  - Bioinformatics: gene/proteome databases distributions
  - Medical applications (screening, epidemiology...): image databases distribution
  - Parallel algorithms for medical image processing, simulation, etc
  - Interactive application (human supervision or simulation)
  - Security/privacy constraints
  - Heterogeneous data formats (genomics, proteomics, image formats)
  - Frequent data updates
  - Complex data sets (medical records)
  - Long term archiving requirements

Biomedical pilot applications report

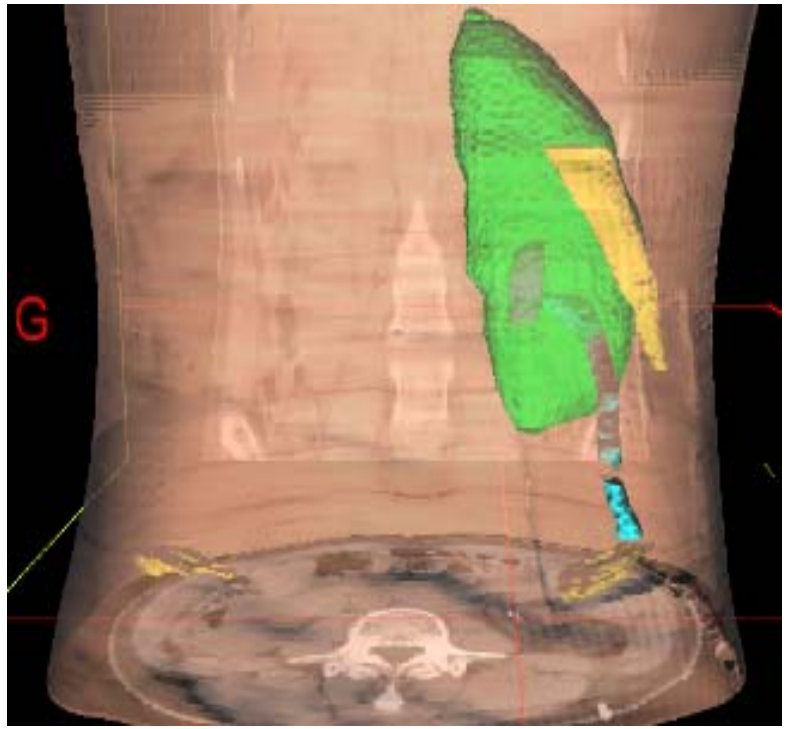
- **CDSS: Clinical Decision Support System (UPV)**
  - Application that Extracts Medically Relevant Knowledge from a Large Set of Information with the Objective of Guiding the Practitioners in their Clinical Practice.
  - http://egee-na4.ct.infn.it/biomed/CDSS.html
- **Sample Question Addressed by CDSS**
  - Which are the Genetic Factors that *Can* be Involved in the Schizophrenia?
- **The CDSS do not Substitute Human Medical Decision, but Improves Factors such as Sensitivity, Sensibility and Working Conditions.**
- **The CDSS is oriented to:**
  - Reinforce the Clinical Experience in Complex Cases.
  - Automatically Focus the Attention of the Expert on the Relevant Points.
  - Investigate New Correlations or Combine the Information from Different Sources.
  - Key Tool in Evidence-Based Medicine

Biomedical pilot applications report

- **GPS@: genomics web portal (IBCP)**
  - Already ported on EDG through an application C++ API
  - http://gpsa.ibcp.fr/
  - sequence homology search against proteic databases (blast)
  - patterns and signatures search
  - modeling of protein 3D structure (proscan, pattinprot)
- **NPSA/GPS@ submits a large number of short jobs. The long waiting time induced by LCG2 middleware event for short jobs is too penalizing and leads to server performance drops. GPS@ requires short execution time for short jobs. This includes:**
  - Resource Broker latency reduction.
  - Short queues for processing short jobs.
  - Possibly dedicated resources to short jobs.
- **Tested on LCG2 for 4 algorithms (PattInProt, secondary structures prediction)**

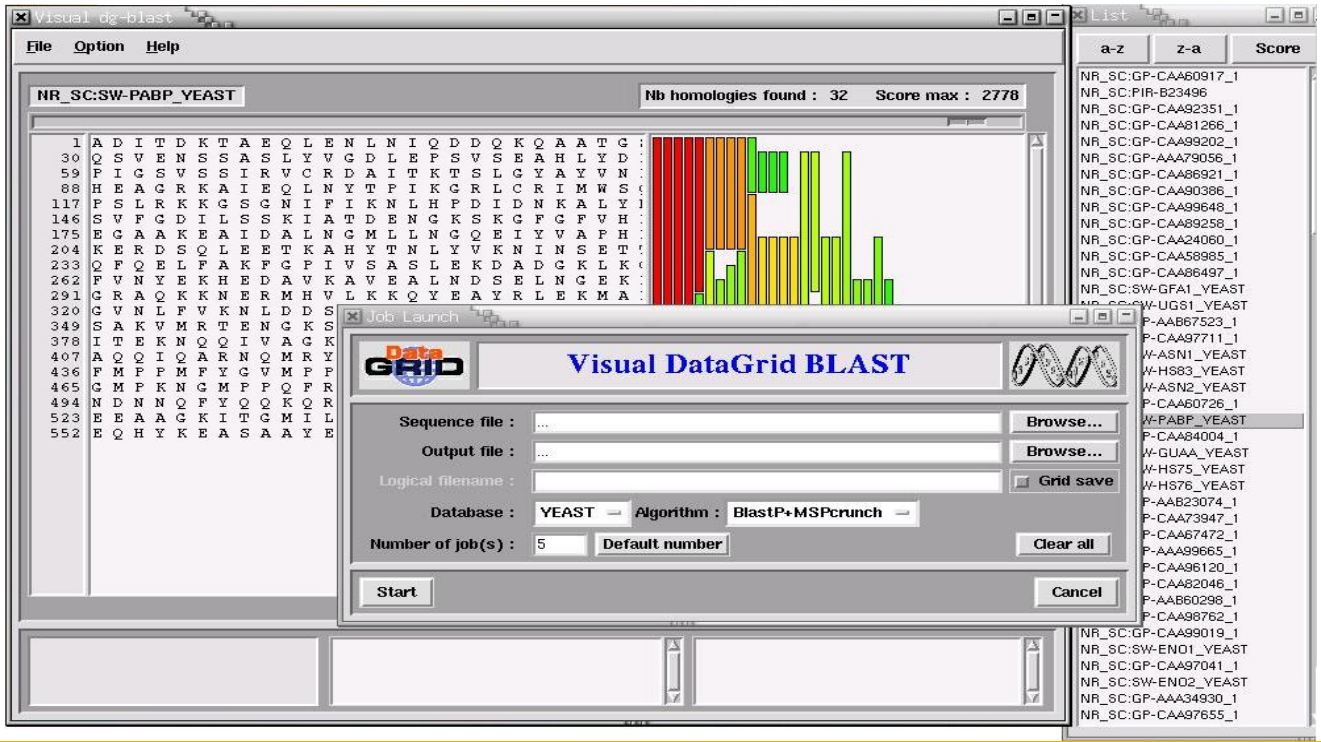
Biomedical pilot application report

- **g-PTM3D: interactive radiological images processing (LAL)**
  - radiological images manipulation
  - interaction and jobs execution
  - LCG2 pay off compensated by internal scheduler (scheduler jobs)
  - interaction through bypass unsatisfying (std input/output used for communication, far too slow)
  - http://egee-na4.ct.infn.it/biomed/gPTM3D.html

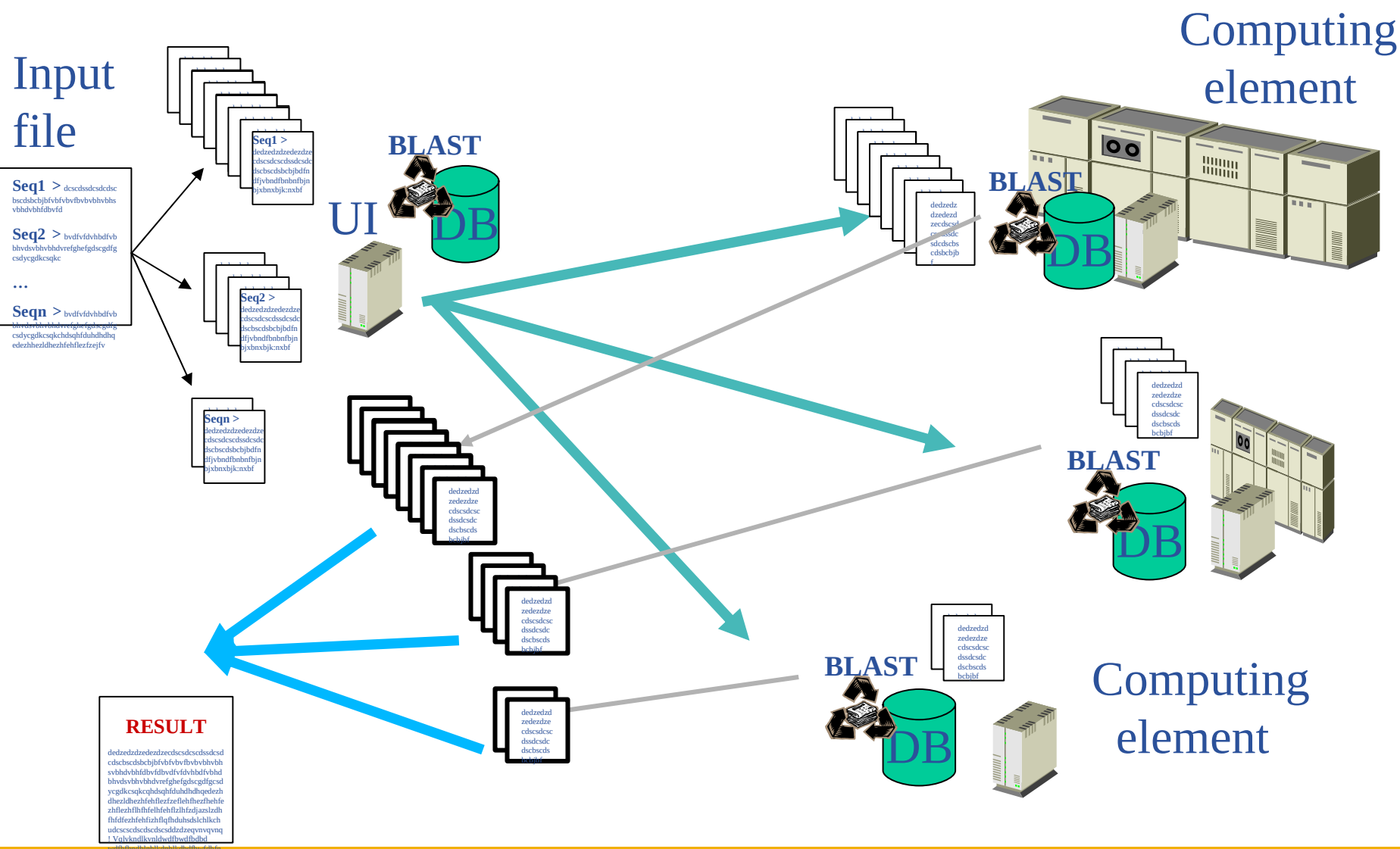


BLAST – comparing DNA or protein sequences

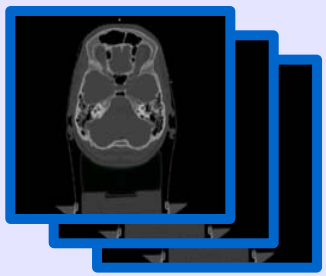
- **BLAST is the first step for analysing new sequences: to compare DNA or protein sequences to other ones stored in personal or public databases. Ideal as a grid application.**
  - Requires resources to store databases and run algorithms
  - Can compare one or several sequence against a database in parallel
  - Large user community



BLAST gridification



Monte carlo simulation for radiotherapy planning



Scanner slices: DICOM format

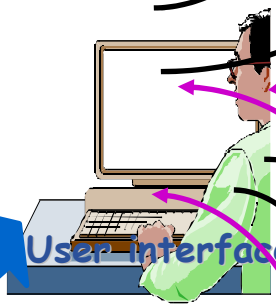
Concatenation

Image: text file

Binary file: Image.raw Size 19M

Retrieving of root output files from CEs

the CE



Anonymisation

Database

GATE

Computing Element

Storage Element

CCIN2P3

GATE

Computing Element

Storage Element

RAL

GATE

Computing Element

Storage Element

NIKHEF

GATE

Computing Element

Storage Element

MARSEILLE

HEP applications and the grid

- Have been running large distributed computing systems for many years
- Now the focus for the future is on computing for LHC and hence we have the LCG (LHC computing grid project)
- In addition to the 4 LHC experiments(ATLAS,ALICE,CMS,LHCb) other current HEP experiments use grid technology e.g. Babar,CDF,D0.., and don't forget Theory and other new HEP experiments..
- LHC experiments are currently executing large scale data challenges(DCs) involving thousands of processors world-wide and generating many Terabytes of data
- Moving to so-called 'chaotic' use of grid with individual user analysis (thousands of users operating within experiment VOs)..see ARDA

LHC Experiments

ATLAS

CMS

ALICE

LHCb

Storage –

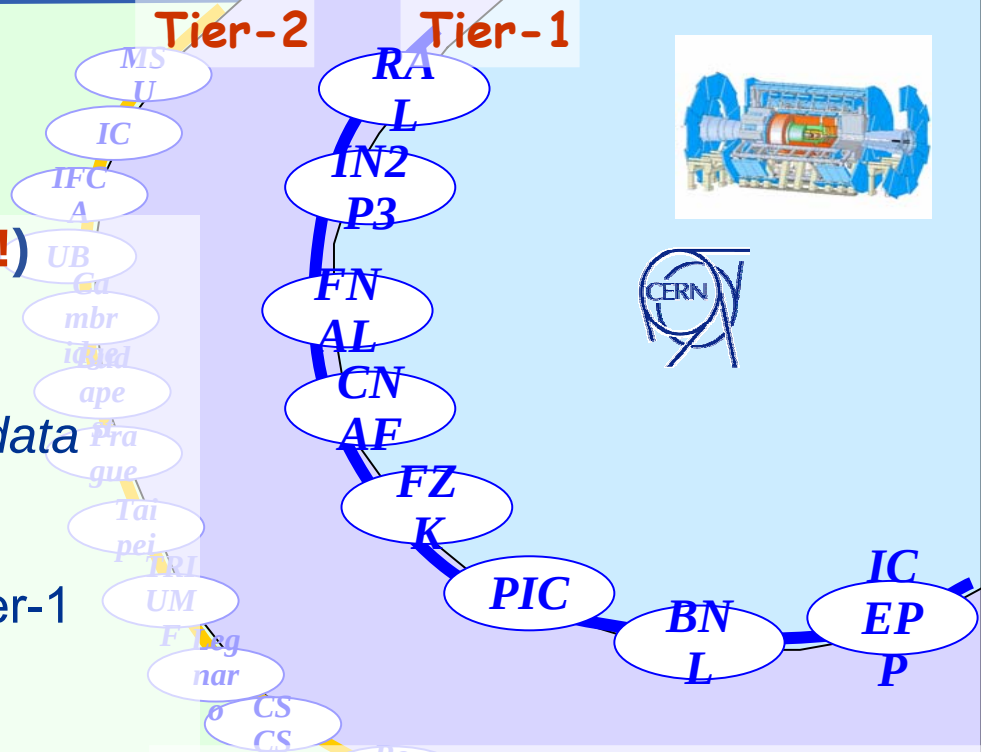
Raw recording rate 0.1 – 1 GByte/s

Accumulating at 5-8 PetaByte/year

10 PetaByte of disk

Processing –

200,000 of today's fastest PCs



LHC Computing Model (simplified!!)

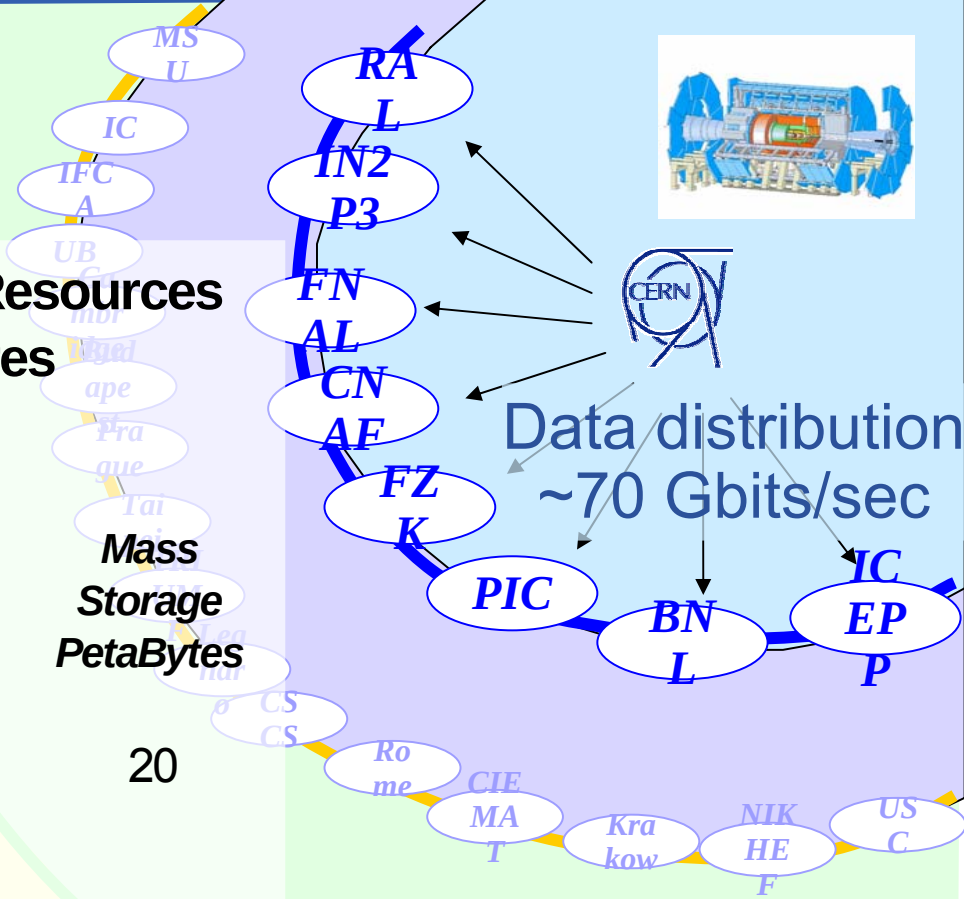
- **Tier-0 – the accelerator centre**
  - Filter → *raw data*
  - Reconstruction → *summary data* (ESD)
  - Record *raw data* and *ESD*
  - Distribute *raw* and *ESD* to Tier-1
- **Tier-1 –**
  - Permanent storage and **management** of *raw*, calibration data, meta-data, analysis data and databases
    → **grid-enabled data service**
  - Data-heavy analysis
  - National, regional support
- **"online" to the data acquisition process**
- **high availability, long-term commitment managed mass storage**

- Tier-2 –
  - Well-managed disk storage – grid-enabled
  - Simulation
  - End-user analysis – batch and interactive
  - High performance parallel analysis (PROOF)

Current estimates of Computing Resources needed at Major LHC Centres

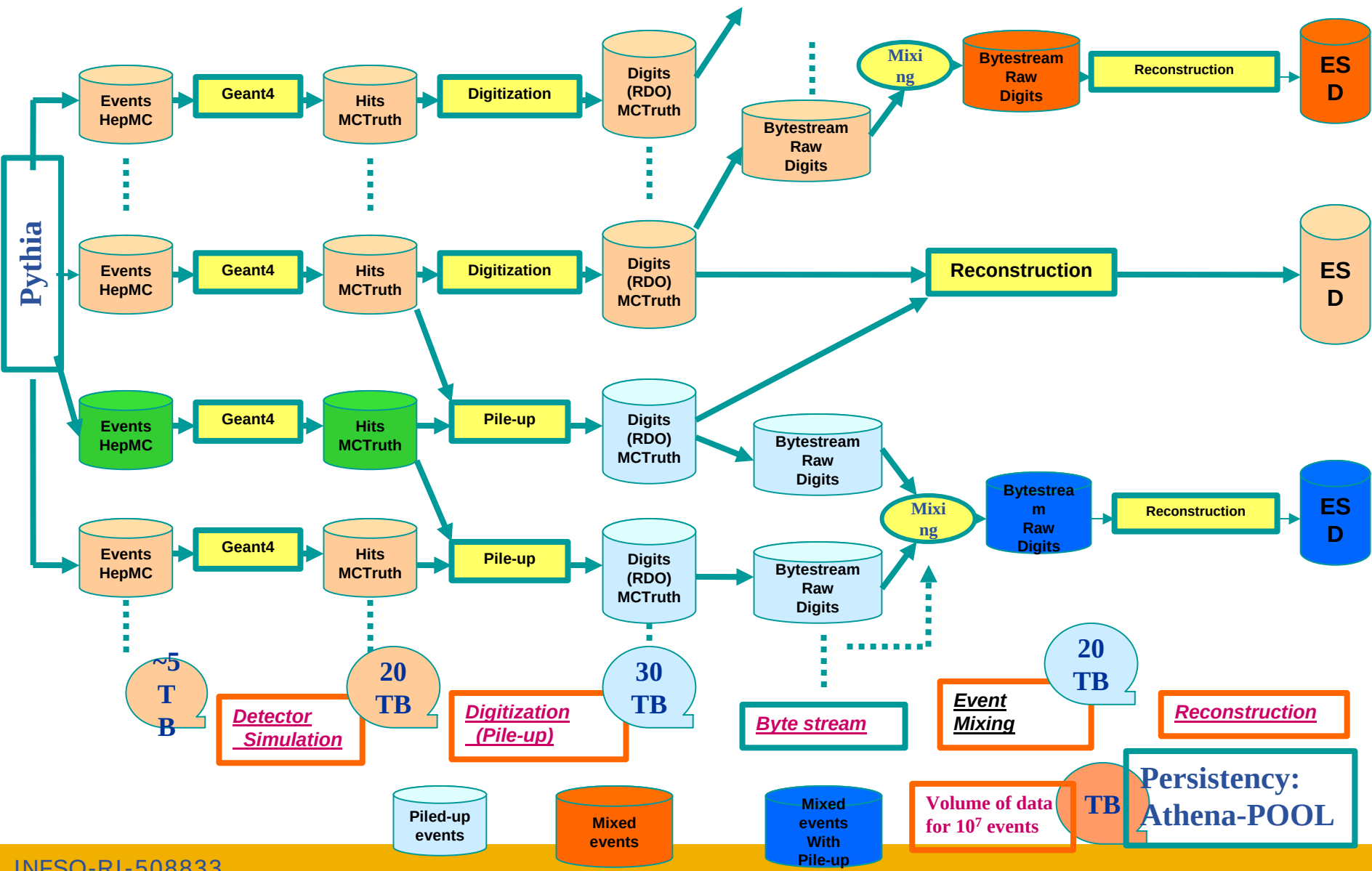
First full year of data - 2008

	*Processing M SI2000***	*Disk PetaBytes*	*Mass Storage PetaBytes*
CERN	20	5	20
Major data handling centres (Tier 1)	45	20	18
Other large centres (Tier 2)	40	12	5
Totals	**105**	**37**	**43**

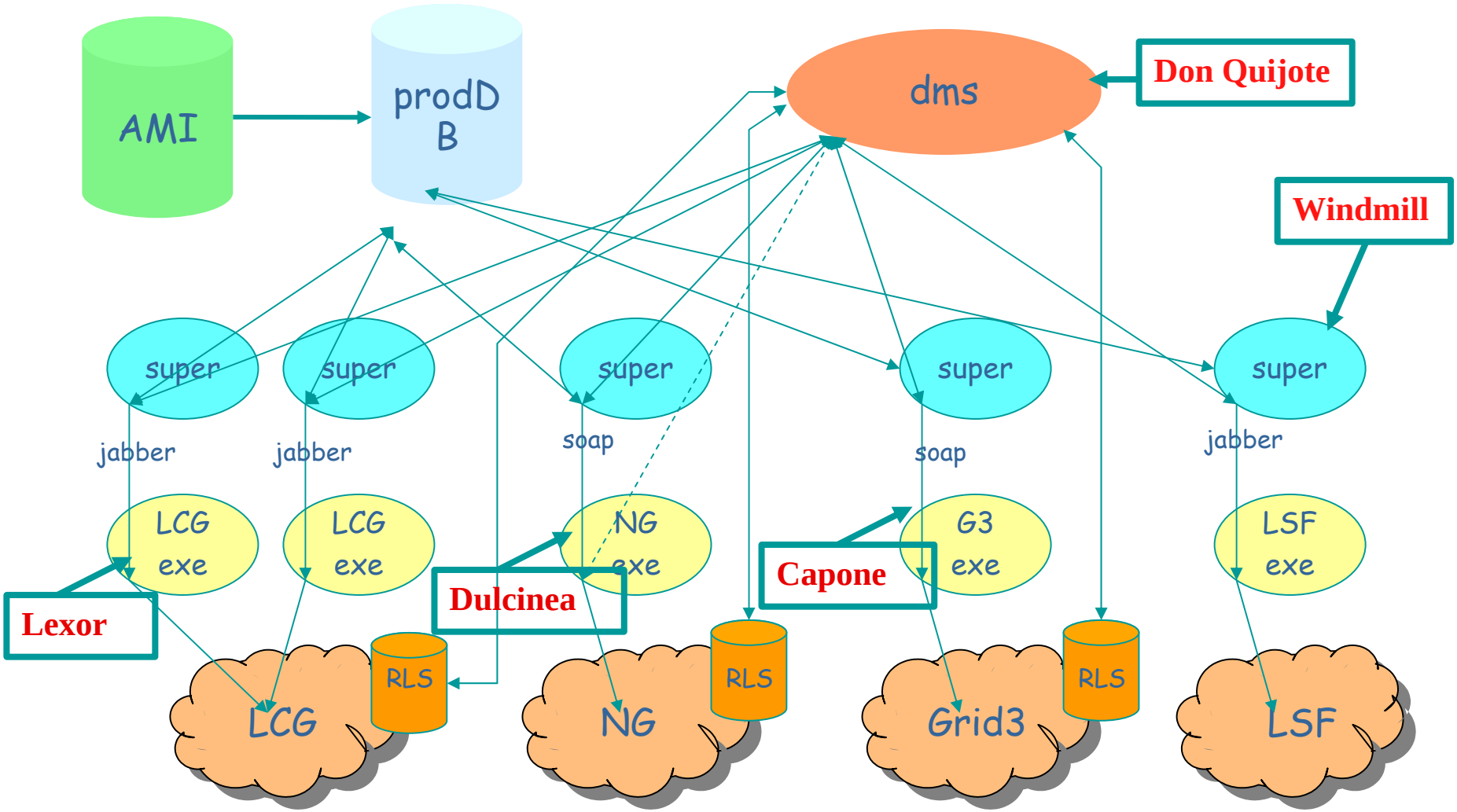


*** Current fast processor ~1K SI2000*

Atlas Data Challenge

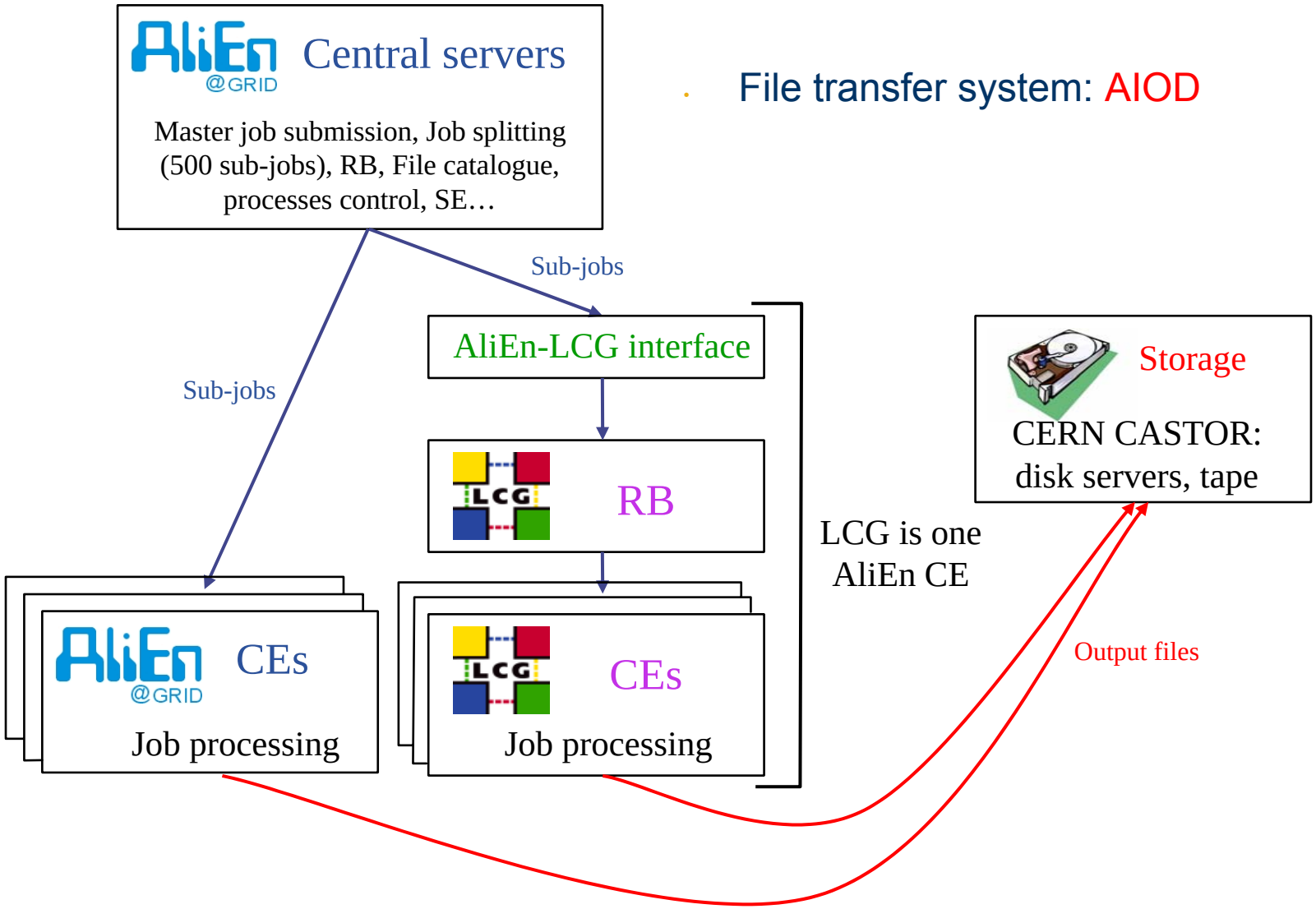


Atlas Data Challenge



Alice Data Challenge

- File transfer system: AIOD



HEP applications and data challenges using LCG-2

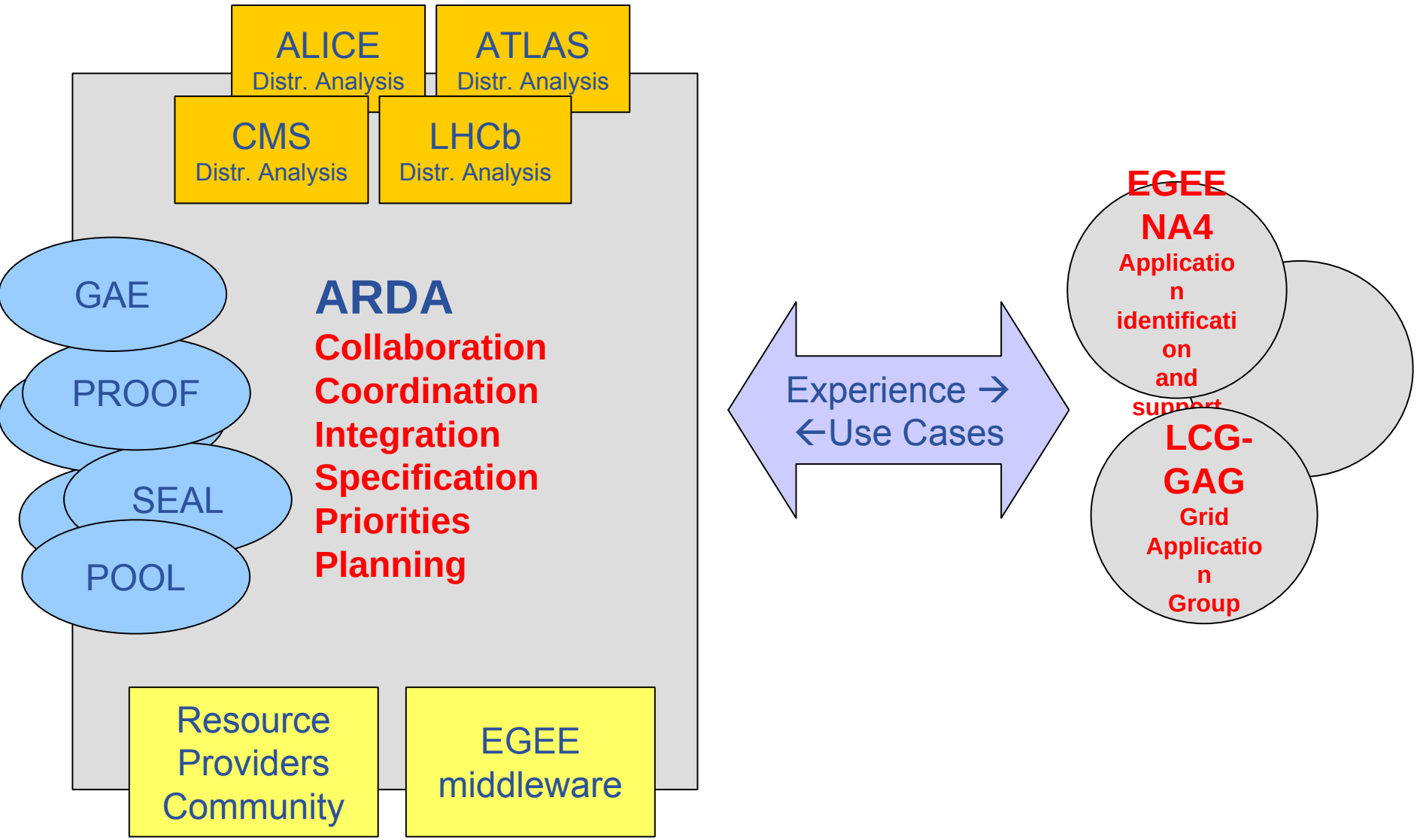
- **All have the same pattern of event simulation, reconstruction and analysis in production mode (with some limited 'user' analysis)**
- **All are testing their running models using Tier-0/1/2 with different levels of ambition, and doing physics with the data**
- **Larger scale user analysis to come with ARDA**
- **All have LCG-2 co-existing with use of other grids**

- ALICE and CMS started around February
- LHCb in May
- ATLAS pile-up
- D0 also making some use of LCG
- Regular reports in LCG GDB and PEB

  see reports of June 14 at LCG/GDB

  http://agenda.cern.ch/fullAgenda.php?ida=a04114
- All very happy about LCG user-support 'attitude' – very cooperative

ARDA (Architectural Roadmap for Distributed Analysis)



NA4 Generic Applications

- This is a key activity in the process of getting new scientific and industrial communities interested and committed to use the continental grid infrastructure built by the EGEE Project.
- GENIUS is a well established tool which will be fundamental in the process of interfacing new applications with the EGEE middleware hiding its complex internals to non-experts users from new communities.
- GILDA is a complete suite of grid elements (test-bed, CA, VO, monitoring system, web portal) and applications fully dedicated to dissemination purposes. This could also represent the ideal grid testbed where to start the porting of new generic applications.
- GILDA is the dissemination tool which will be used by NA3 during courses and tutorials so the important aspect of induction of the grid paradigm to new communities is also covered.
- It is now important to have the first meeting of the EGAAP board and define the first Generic Applications to be interfaced.

The GILDA home page (http://gilda.ct.infn.it)

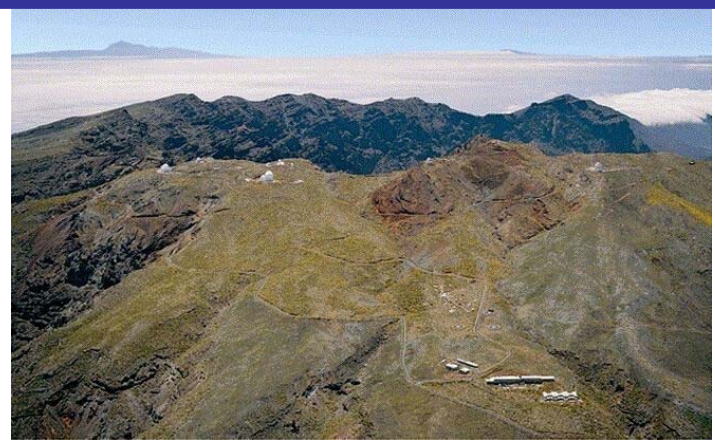
The Generic Application questionnaire

- Questionnaire to get information and first requirements from new communities interested in using the EGEE Infrastructure http://alipc1.ct.infn.it/grid/egee/na4/questionnaire/na4-genapp-questionnaire.doc
- Feed-backs received so far http://alipc1.ct.infn.it/grid/egee/na4/questionnaire:
  - Astrophysics (EVO and Planck satellite)
  - Earth Observation (ozone maps, seismology, climate)
  - Digital Libraries (DILIGENT Project)
  - Grid Search Engines (GRACE Project)
  - Video on demand
- Interest also from Computational Chemistry (Italy and Czech Republic), Civil Engineering (Spain), and Geophysics (Switzerland and France) communities

Astro-particle physics (MAGIC)

- **Ground based Air Cerenkov Telescope**
- **LaPalma, Canary Islands (28° North, 18° West)**
- **17 m diameter**
- **operation since autumn 2003 (part of CrossGrid)**
- **Collaborators:**

IFAE Barcelona, UAB Barcelona, Humboldt U. Berlin, UC Davis, U. Lodz, UC Madrid, MPI München, INFN / U. Padova, U. Potchefstrom, INFN / U. Siena, Tuorla Observatory, INFN / U. Udine, U. Würzburg, Yerevan Physics Inst., ETH Zürich



Video-on-demand

- VLC media player is a cross-platform media player and server. It works under Linux, Windows, Mac OS X, BeOS, BSD, Solaris, Familiar Linux and QNX.

- VLC (initially VideoLAN Client) is a highly portable multimedia player for various audio and video formats (MPEG-1, MPEG-2, MPEG-4, DivX, mp3, ...) as well as DVDs, VCDs, and various streaming protocols. It can also be used as a server to stream in unicast or multicast in IPv4 or IPv6 on a high-bandwidth network.

- Last release: VLC.0.7.2 installed on all GILDA sites
- More information about the VideoLAN streaming solution can be found in the streaming section http://www.videolan.org/.

Computational Chemistry

- **The problem of simulating in a realistic way complex processes on a molecular basis without resorting into phenomenological or ad hoc empirical approaches is a vital need for scientists**
  - investigating new materials
  - biological systems
  - life science
  - food and drug action
  - chemical processes …
- **Up to now only specific models limited to simplified examples have been considered due to the difficulty of dealing with the complexity of realistic systems.**

Computational Chemistry (2)

- **The rapid development of Grid technologies is making this possible by**
  - connecting the expertise, software and hardware needed to build a Molecular Simulator on a single distributed system
  - coordinating the various phases through a web based Workflow management environment
  - using a common User Interface (UI) to give to the users the access to the Grid facilities
  - enabling the usage of a large number of computer nodes
  - implementing a good security infrastructure
  - Grid based European Molecular Simulator, http://gems.simbex.org

Conclusions

- **NA4 is up and running now**
  - HEP is using LCG-2 for data challenges
  - ARDA is well under way and waiting for first new middleware prototype
  - Biomedicine has applications ready to go onto LCG-2 and pre-production services
  - Generic group is very active with GILDA and excellent relations with NA3
  - Testing group has active dialogue with JRA1 and ARDA for rationalising testing effort
  - Industry forum has developed links with several companies (see EGEE Cork presentations)
- **EU demos NA4 session:**
  - (http://agenda.cern.ch/fullAgenda.php?ida=a044621)
- **NA4 Web site** **http://egee-na4.ct.infn.it**