Enabling Grids for E-science in Europe

www.eu-egee.org

NA4 Applications

F.Harris(Oxford/CERN)
NA4/HEP coordinator

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

Talk Outline

Enabling Grids for E-science in Europe

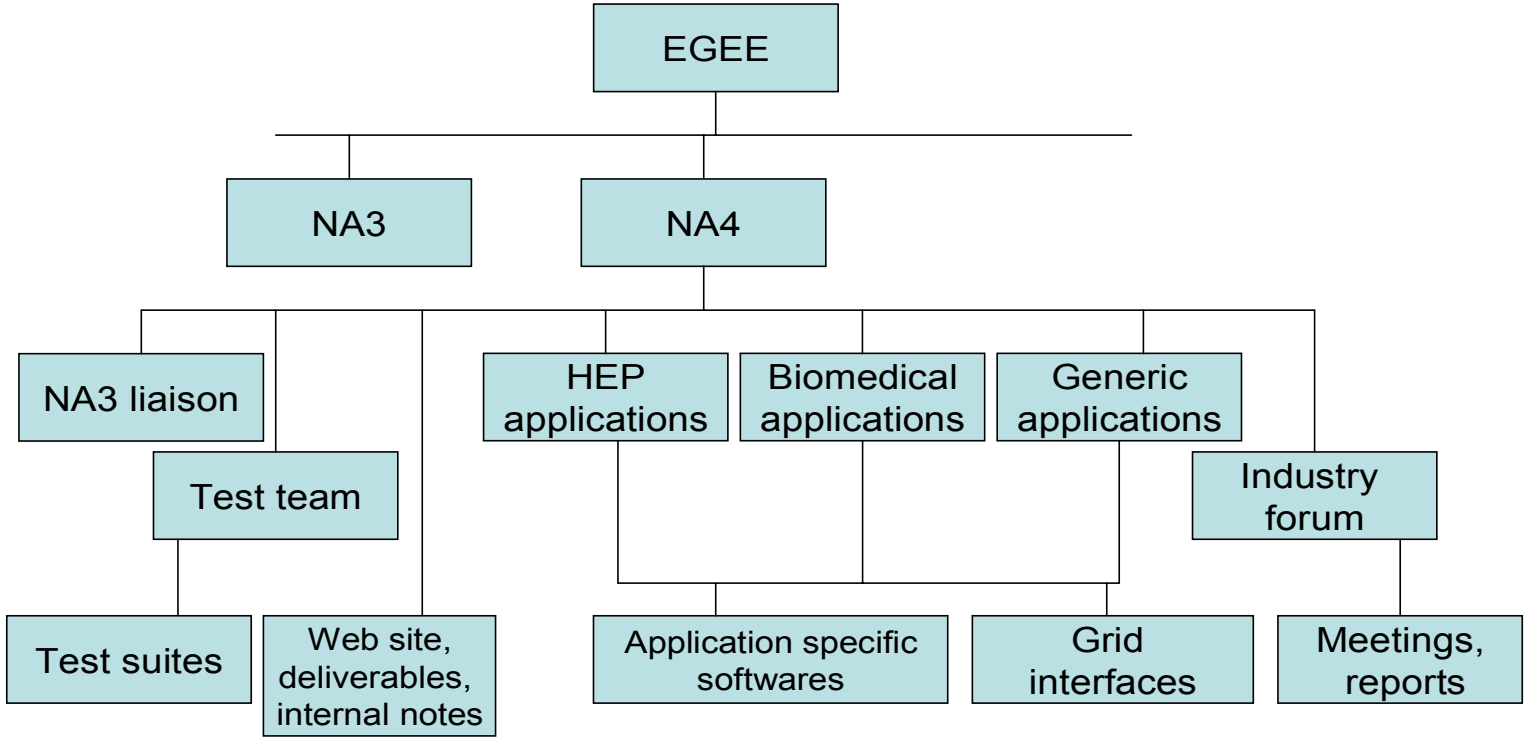
- The basic goals of NA4
- The organisation
- The participants and their roles
- The flavour of the work for the NA4 sub-groups
  - *Biomed*
  - *HEP*
  - *'Generic' applications*
  - *Testing*
  - *Industry Forum*
- Milestones and deliverables
- Relations with other EGEE activities
- Concluding comments

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

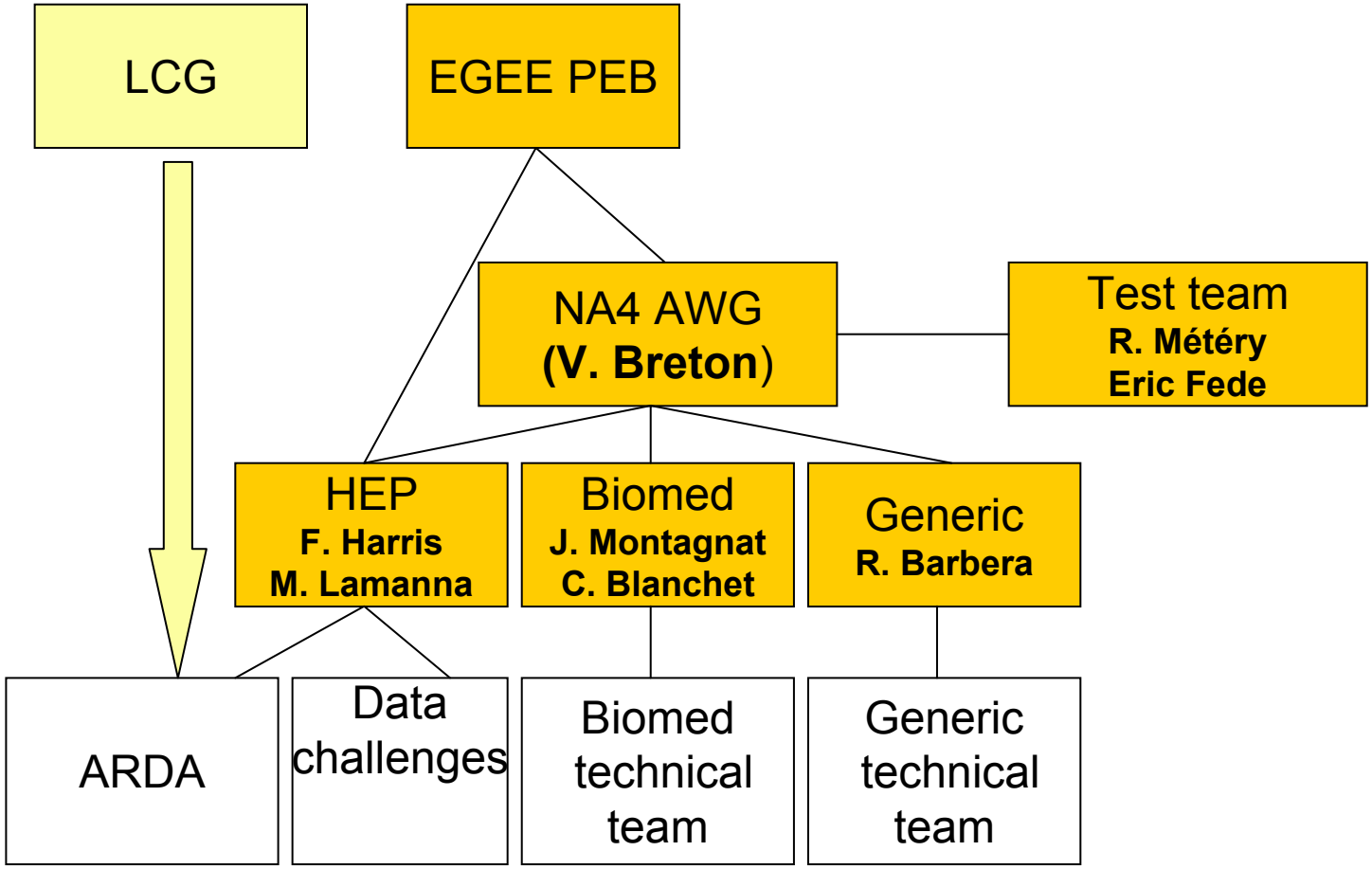
Enabling Grids for E-science in Europe

- 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



NA4 technical organization



Roles and staffing

Enabling Grids for E-science in Europe

Federation	Role	FTE Funded	FTE Unfunded
CERN	HEP Applications (coord.)	4	4 (9)
UK+Ireland	NA3 Liaison	0,5	0,5
Italy	Generic app (coord)	2	2
France	General coord., BioMed, Test team, Industry diss.	7	7
Northern Europe	Generic applications	1	1
Germany + Switzerland	Generic applications	1	1
Central Europe	Generic applications	1	1
South West Europe	BioMed	2	2
Russia	HEP, BioMed	3	3
Totals		**21,5**	**21,5**

What do all applications expect from the grid?

Enabling Grids for E-science in Europe

- 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

Enabling Grids for E-science in Europe

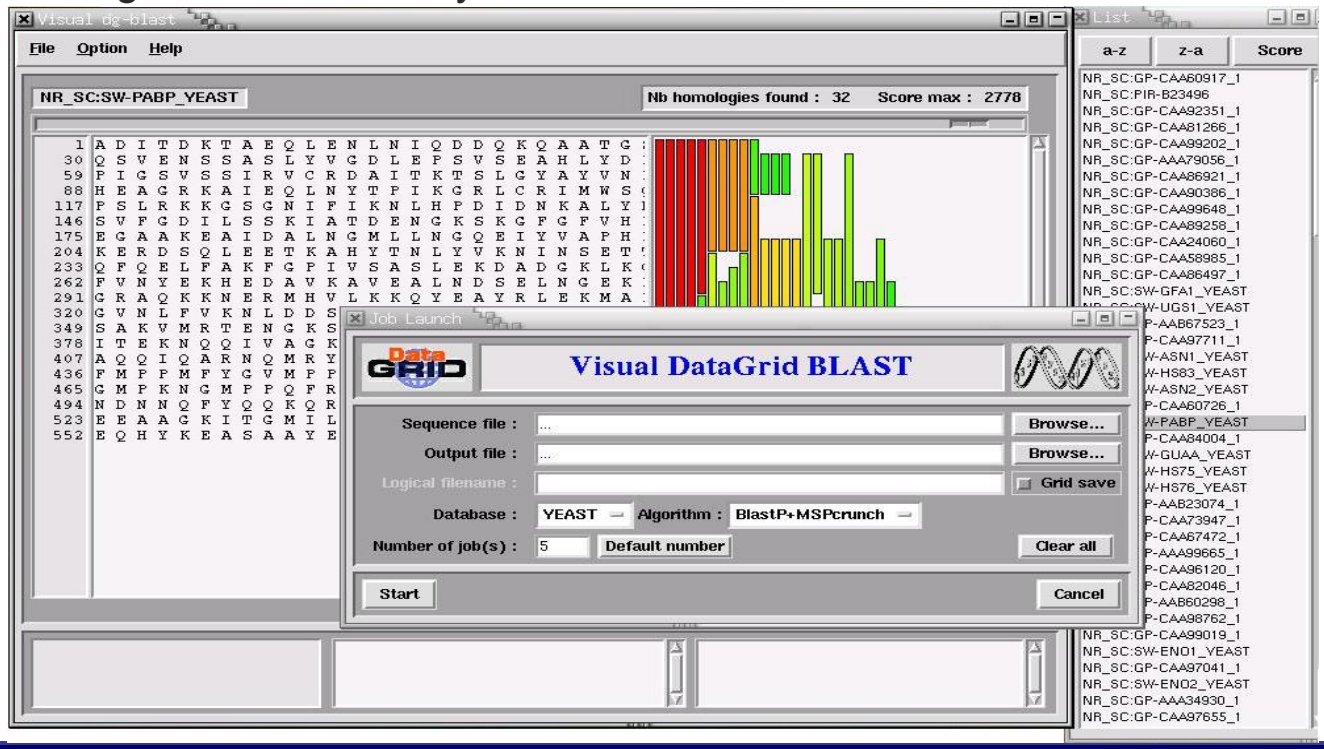
- Some key applications and their characteristics
  - Bioinformatics: gene/proteome databases distributions
  - Medical applications (screening, epidemiology...): image databases distribution
  - Parallel algorithms for medical image processing, simulation, etc
  - Interactive application (human supervision or simulation)

  - Security/privacy constraints
  - Heterogeneous data formats (genomics, proteomics, image formats)
  - Frequent data updates
  - Complex data sets (medical records)
  - Long term archiving requirements

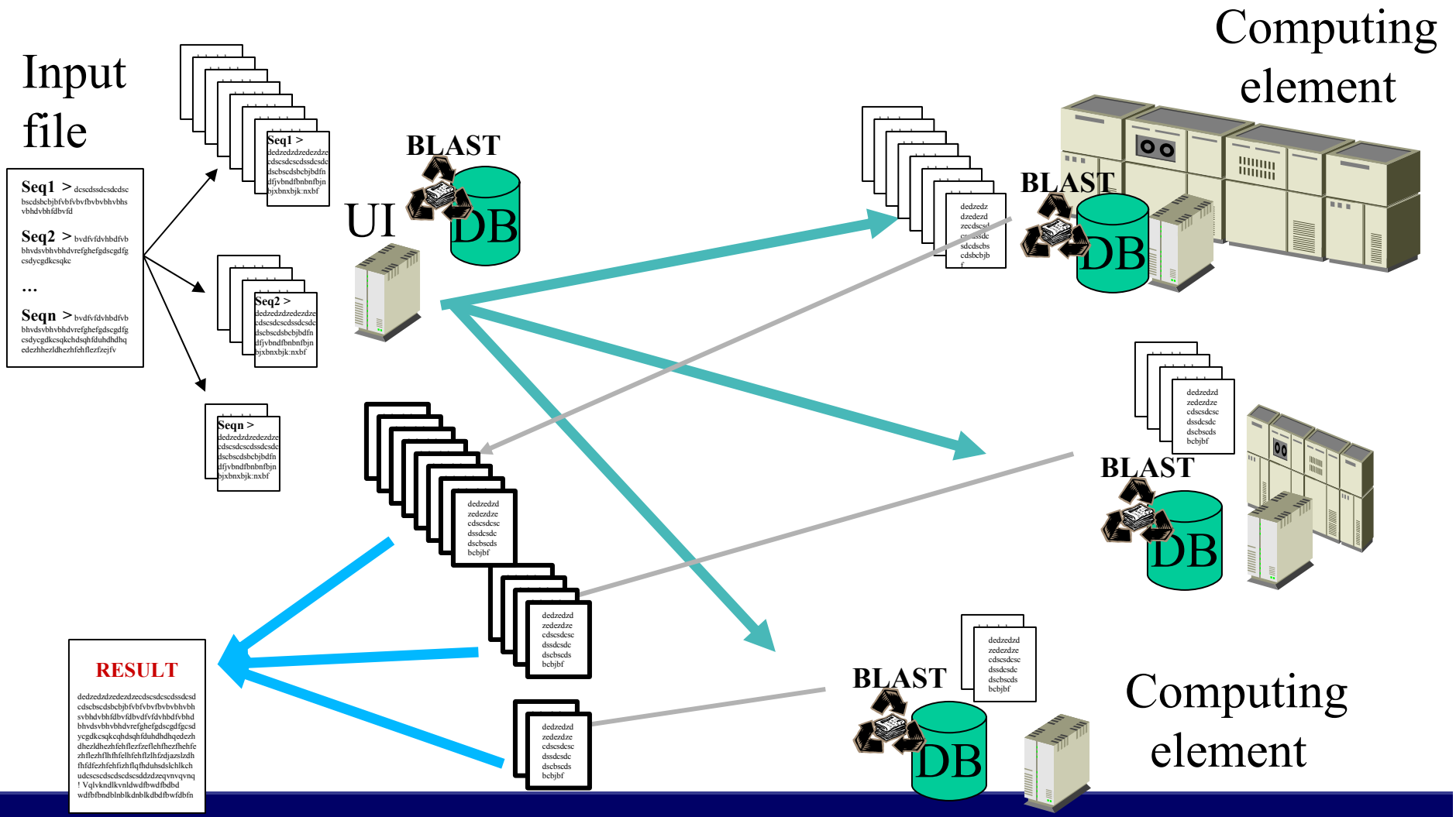
BLAST – comparing DNA or protein sequences

Enabling Grids for E-science in Europe

- 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

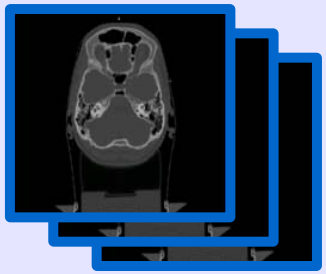


Monte carlo simulation for radiotherapy planning

Enabling Grids for E-science in Europe

Retrieving of root output files from CEs

the CE



Scanner slices: DICOM format

Concatenation

User interface

GATE

Computing Element

Storage Element

CCIN2P3

GATE

Computing Element

Storage Element

RAL

GATE

Computing Element

Storage Element

NIKHEF

GATE

Computing Element

Storage Element

MARSEILLE

Anonymisation

Databa se

Image: text file

Binary file: Image.raw Size 19M

HEP applications and the grid

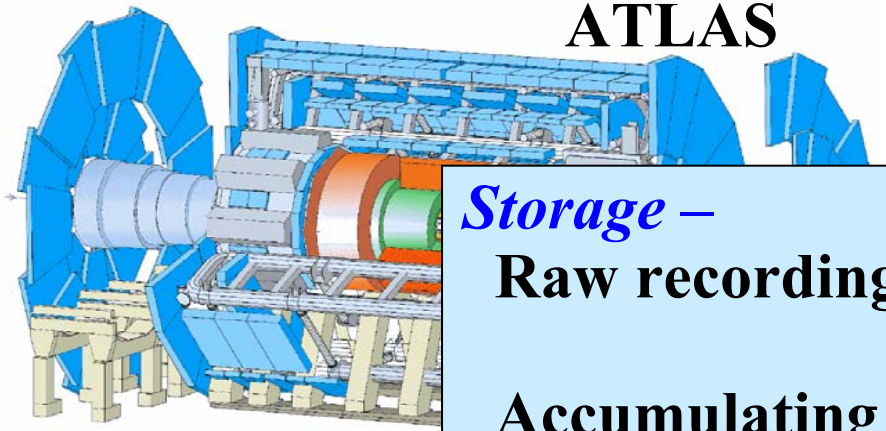
Enabling Grids for E-science in Europe

- 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

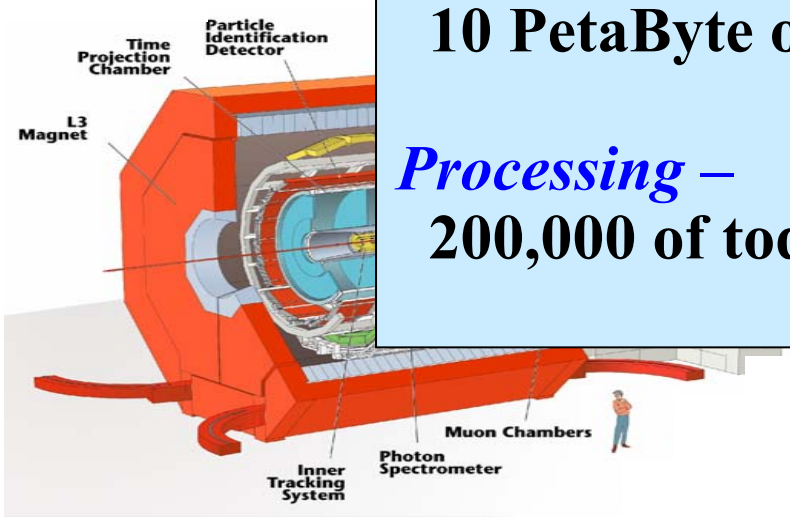
Enabling Grids for E-science in Europe

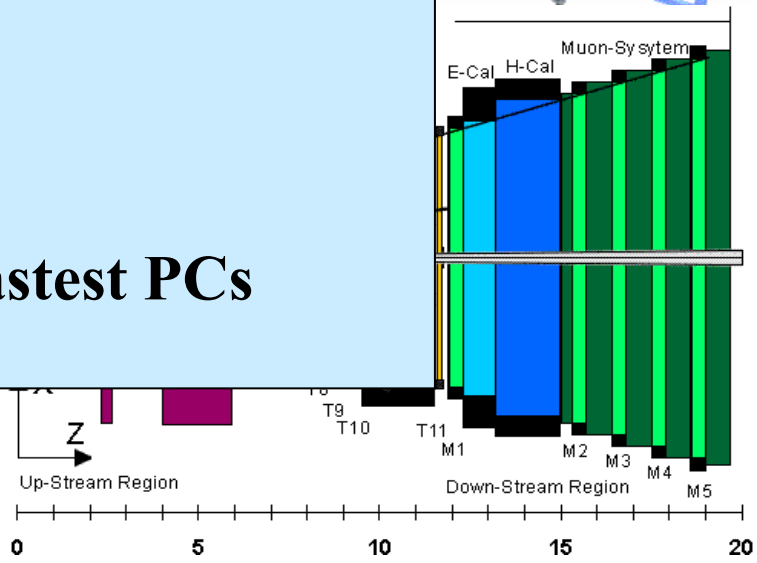
ATLAS



CMS

ALICE





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 Grid Project - a Collaboration

Enabling Grids for E-science in Europe

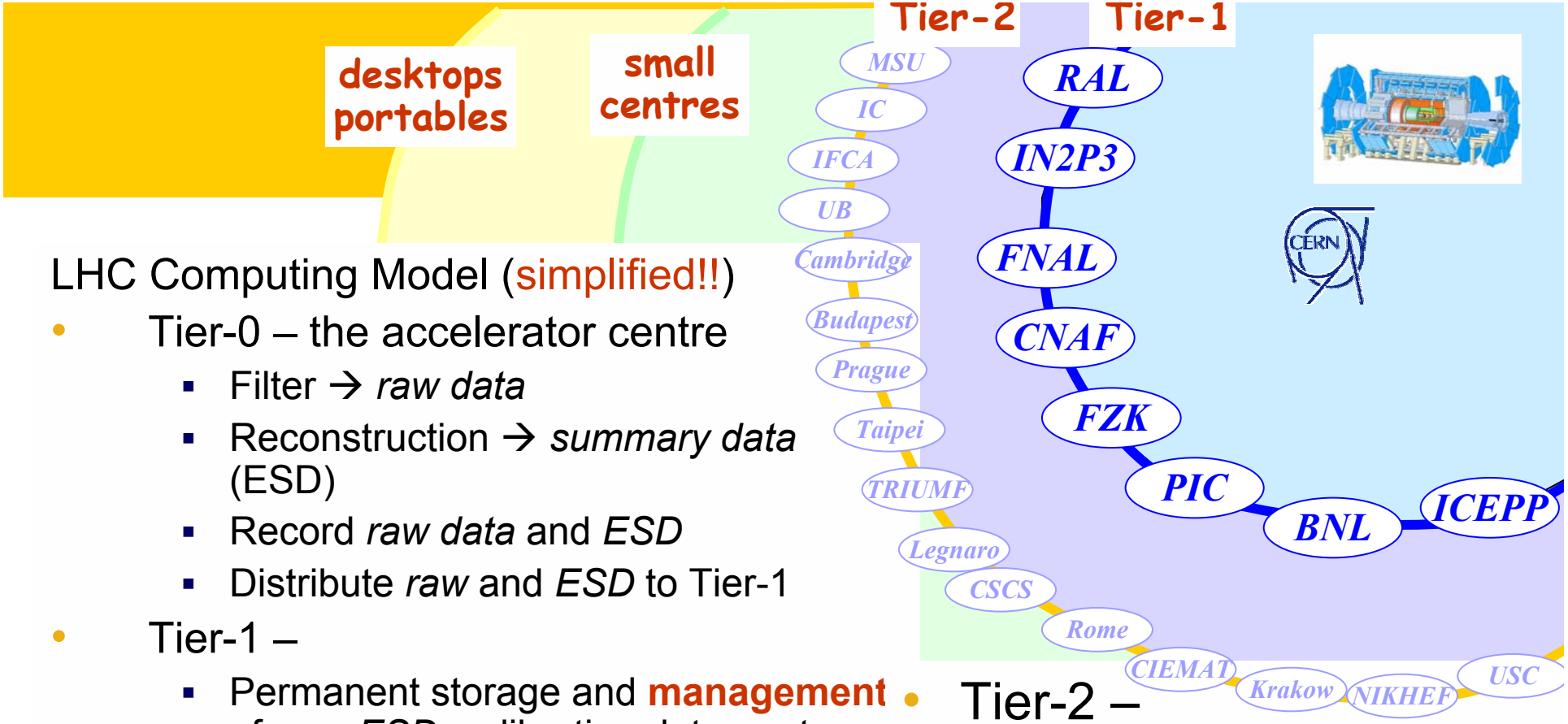
Building and operating the LHC Grid – a collaboration between

- The physicists and computing specialists from the LHC experiment
- The projects in Europe and the US that have been developing Grid middleware
- The regional and national computing centres that provide resources for LHC
- The research networks

Researchers

Software Engineers

Service Providers



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*, *ESD*, calibration data, meta-data, analysis data and databases → **grid-enabled data service**
  - Data-heavy analysis
  - Re-processing raw → ESD
  - National, regional support

"online" to the data acquisition process high availability, long-term commitmen 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*

Characteristics of CMS Data Challenge DC04 (just completed)……run with LCG-2 and CMS resources world-wide (US Grid3 was a major component)

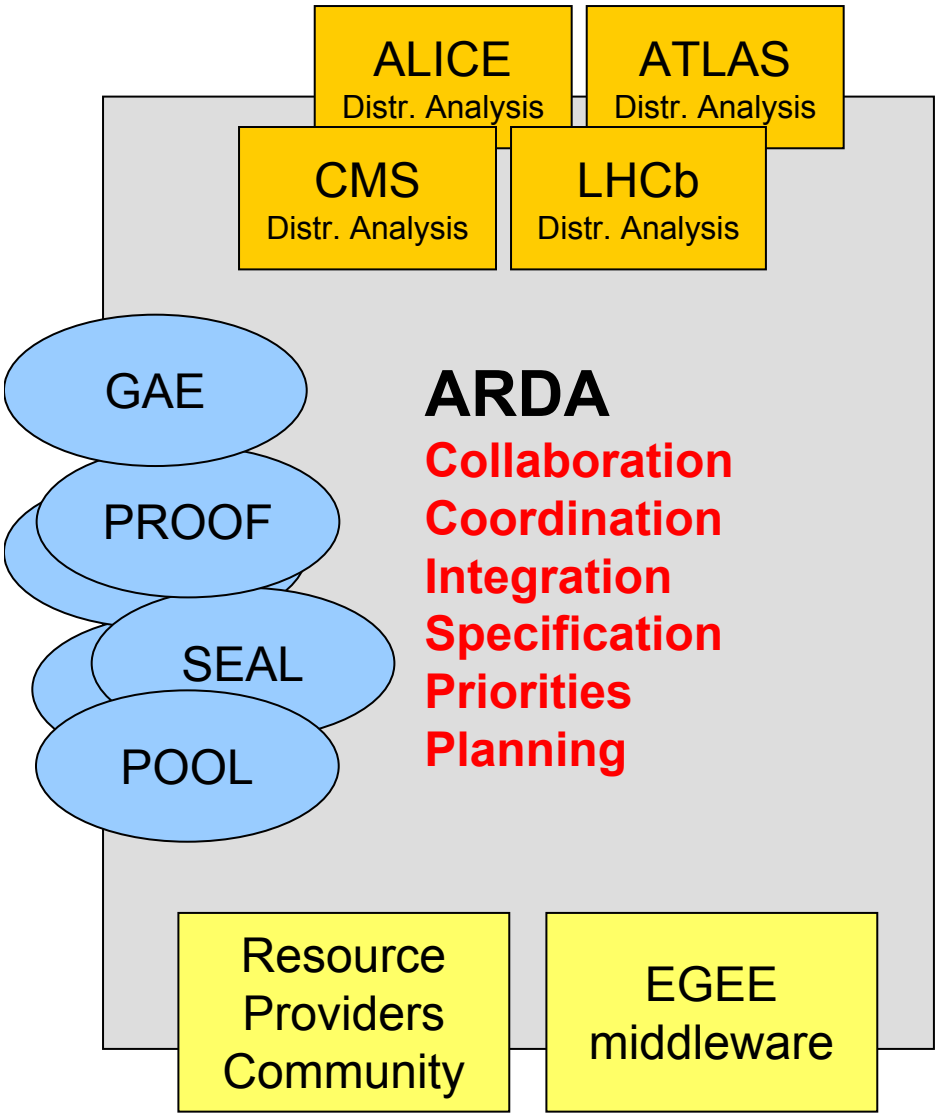
Enabling Grids for E-science in Europe

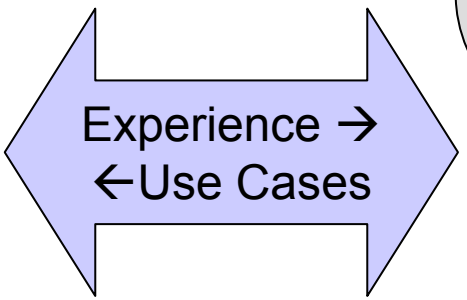
- **Pre-Challenge Production (Phase 1) – simulation generation and digitisation**
  - After 8 months of continuous running:
    - 750,000 jobs
    - 3,500 KSI2000 months
    - 700,000 files
    - 80 TB of data

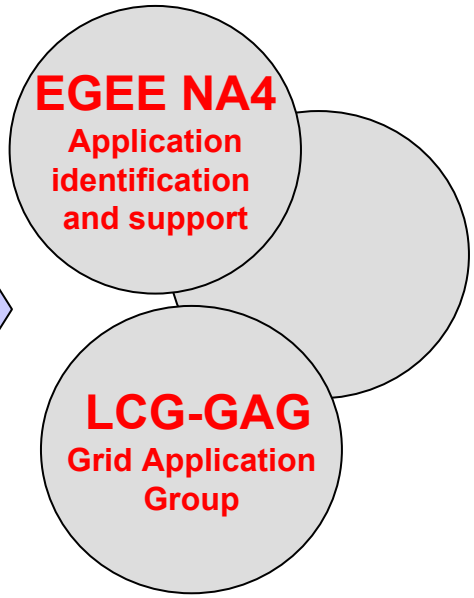
- **Data Challenge (Phase 2)**
  - **Ran the full data reconstruction and distribution chain at 25 Hz**
  - **Achieved**
    - 2,200 jobs/day (about 500 CPU's) running at Tier-0
    - Total 45,000 jobs Tier-0 and 1
    - 0.4 files/s registered to RLS (with POOL metadata)
    - Total 570,000 files registered to RLS
    - 4 MB/s produced and distributed to each Tier-1

ARDA (Architectural Roadmap for Distributed Analysis)

Enabling Grids for E-science in Europe







NA4 Generic Applications

Enabling Grids for E-science in Europe

- 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)

Enabling Grids for E-science in Europe

The Generic Application questionnaire (contribution to MNA4.1)

Enabling Grids for E-science in Europe

- 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)**
  - **Industrial applications (SIMDAT Project)**
- Interest also from Computational Chemistry (Italy and Czech Republic), Civil Engineering (Spain), and Geophysics (Switzerland and France) communities

EGEE Industry Forum Objectives

Enabling Grids for E-science in Europe

- The main role of the Industry Forum in the Enabling Grids for E-Science in Europe (EGEE) project is to raise awareness of the project amongst industry and to encourage businesses to participate in the project. This will be achieved by making direct contact with industry, liaising between the project partners and industry in order to ensure businesses get what they need in terms of standards and functionality and ensuring that the project benefits from the practical experience of businesses.

- The members of the EGEE Industry Forum are companies of all sizes who have their core business or a part of their business within Europe.

- The Industry Forum will be managed by a steering group consisting of EGEE project partners and representatives from business

- http://public.eu-egee.org/industry-forum/information

NA4 Testing Group

Enabling Grids for E-science in Europe

Three types of tests will be developed:

(based on user requirements and on the experience gathered by the ongoing activities like LHC DCs and ARDA prototyping)

- **Tests of service availability:** This set of tests will check the EGEE services availability. All the services providing by the GRID should be tested : Job submission and management, files management, Information service, ...
- **Tests of functionality:** To verify that the functionalities required are available , usable and complete; for example, file creation, moving and deletion, information publication, errors recovery.
- **Tests to measure performances:** Their goal is to characterize the testbed from the end users/application perspective. Part of them will be time measurements (time to submit X job, time to replicateY files,...), others will address scalability measurements (how many jobs can be accepted by service Z, files limits size,...) while others will be more abstract (information availability, errors message access,...).
- This work should be done in close collaboration with ARDA , JRA1 and SA1

Milestones for NA4 applications

Enabling Grids for E-science in Europe

MNA4.1	M6	First applications migrated to the EGEE infrastructure •HEP data challenges for 4 LHC experiments and for D0 •Biomedicine – GATE simulation in nuclear medicine + others •Plus the first 'generic' applications
MNA4.2	M12	First external review of Applications Identification and Support with feedback
MNA4.3	M24	•Second external review of Applications Identification and Support with feedback

NA4 relations with other EGEE activities and other bodies (1)

Enabling Grids for E-science in Europe

- SA1 grid operations
  - How to get new VOs onto LCG from different domains?
  - How to integrate new resources(sites) into LCG coming from different application areas?
  - Rationalistion of test procedures
  - Working with national agencies (e.g. GridPP Application monitoring)
- NA3 training
  - Estimating requirements for courses
  - Design and implementation of courses
- JRA1 middleware
  - All applications input requirements and monitor their satisfaction with feedback to middleware (process goes through the PTF-Project Technical Forum)
- JRA2 quality assurance
  - NA4 have a representative on this group to define process for monitoring quality of EGEE services

NA4 relations with other EGEE activities and other bodies (2)

Enabling Grids for E-science in Europe

- JRA3 security
  - Security of medical (and other application) data
  - Security for sites
- SA2,JRA4 networking
  - Global HEP requirements through LCG
  - Biomed and other applications must similrly give global needs
  - NA4 will give individual application use cases especially where problems have been encountered
- LCG
  - NA4/HEP are presented on the LCG/GAG(grid Applications Group)
    - This is HEP source of requirements and giving feedback to middleware on ‘customer satisfaction’. Some GAG people are on the PTF.

Conclusions

Enabling Grids for E-science in Europe

- 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)
- NA4 open meeting Jul 14-16 at Catania with emphasis on inter-activity dialogue (with middleware,operations,security,networking)
- NA4 Web site http://egee-na4.ct.infn.it