Applications and Use Cases

The European DataGrid Project Team

http://www.eu-datagrid.org

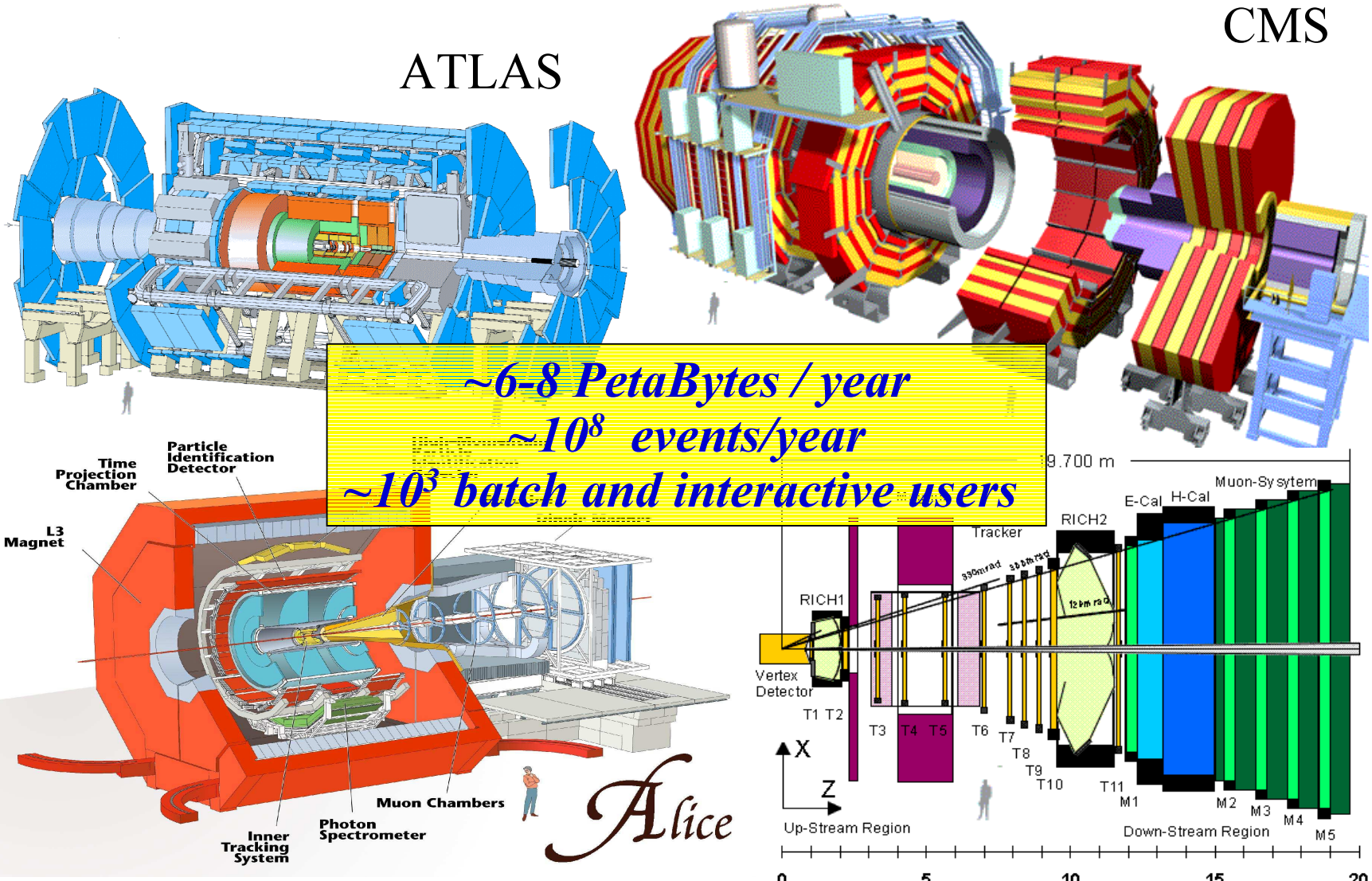
Overview

- **High Energy Physics**
  - Why we need to use GRIDs in HEP ?
  - Brief mention of the Monarc Model
  - Underlying network supporting the GRID
  - Testbed 1 validation : what has already been done on the testbed ?
  - Long terms plans for the GRID : HEP use cases
- **Earth Observation**
  - Mission and plans
  - What do typical Earth Obs. applications do ?
  - Testbed 1 demonstrator
- **Biology**
  - dgBLAST

The LHC Detectors

ATLAS

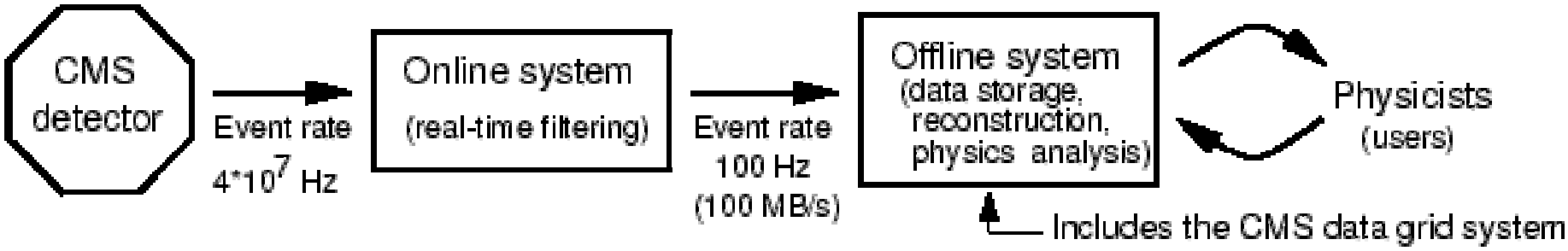
CMS

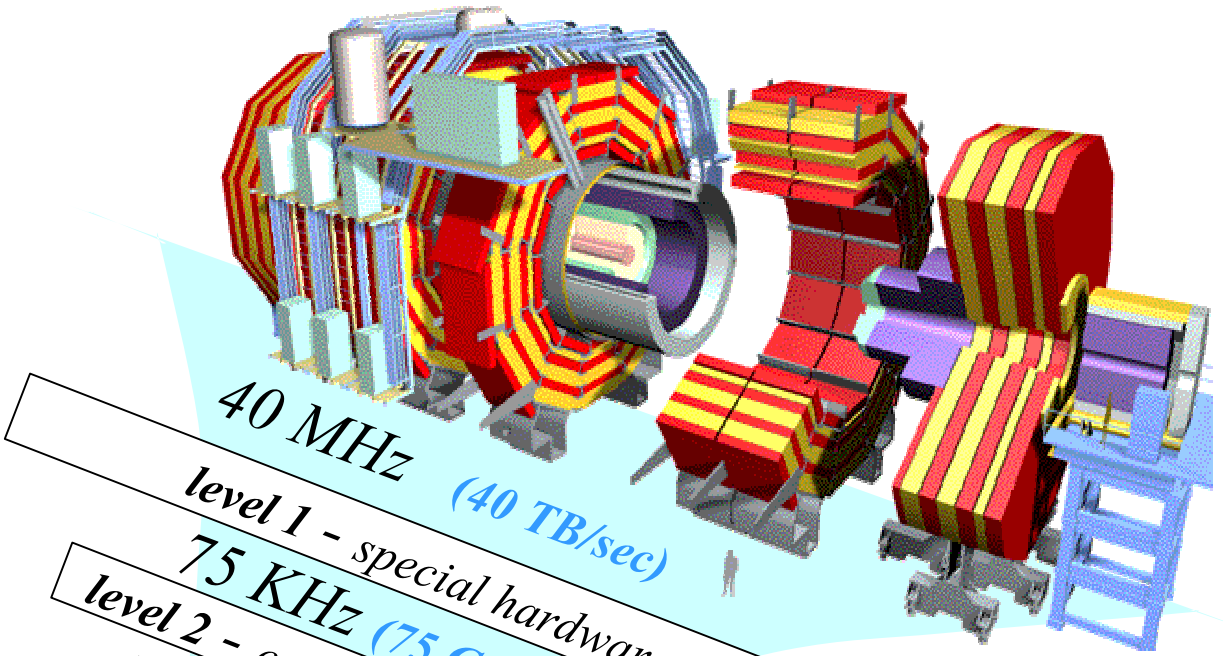


Federico.carminati , EU review presentation

Why using GRIDs in HEP ?

- Distributed nature of the problem : the world community of HEP users has to analyse an unprecedented amount of experimental data
- Every physicist should have equal rights access to the distribued set of data and have transparent access to dynamic resources .
- The system will be extremely complex
  - Number of sites and components in each site
  - Different tasks performed in parallel: simulation, reconstruction, scheduled and unscheduled analysis
- Example: the CMS experiment at LHC:
  - On line : ~ 40 TB /sec (~ 1 raw evt size * $4*10^7$ evts/s)
  - Off-Line : input to Tier 0 : **100 - 200 MB/s**



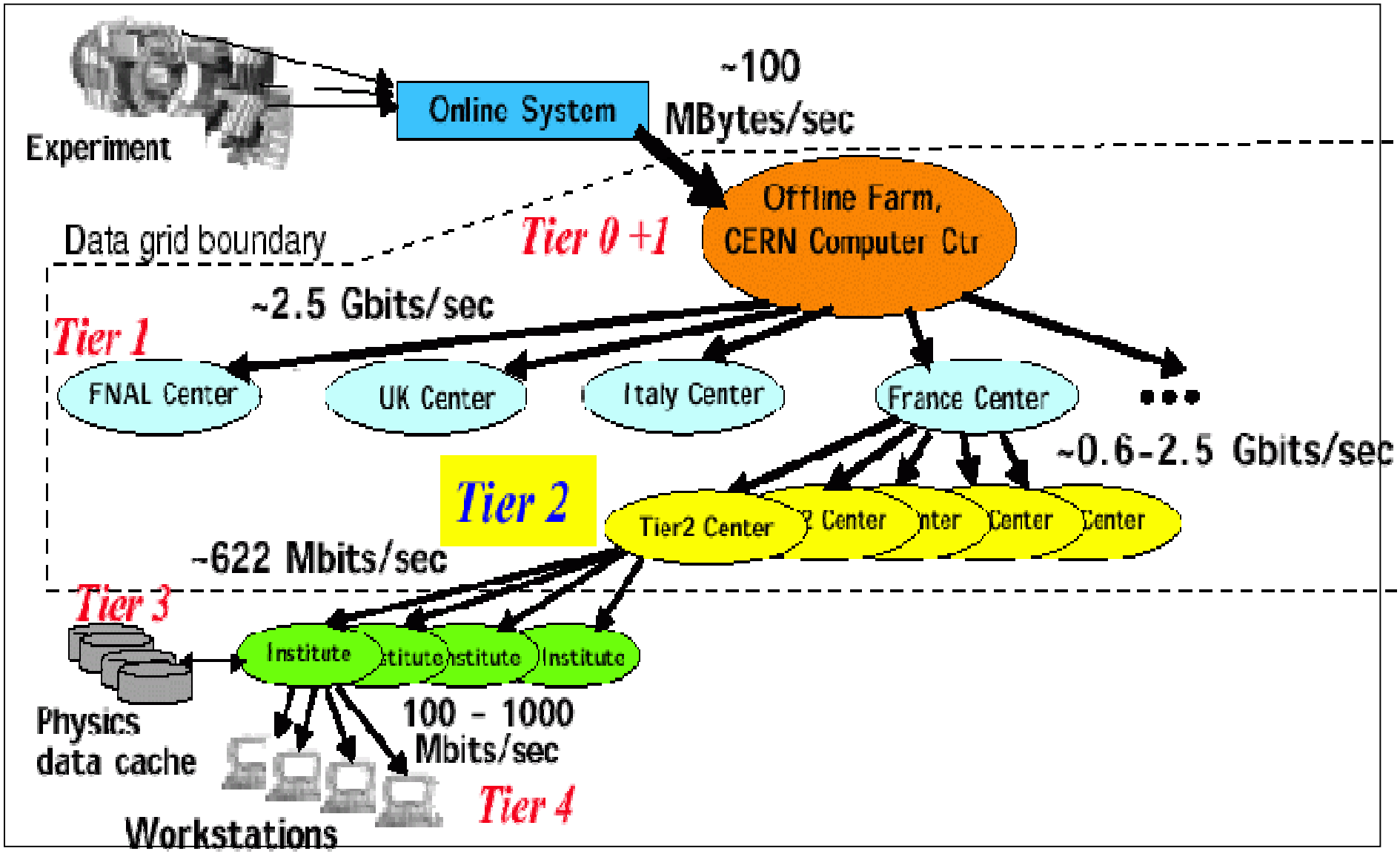


online system

multi-level trigger
filter out background
reduce data volume

les.robertson@cern.ch

The Monarc Model : centre to boundaries view (acquired data flow)

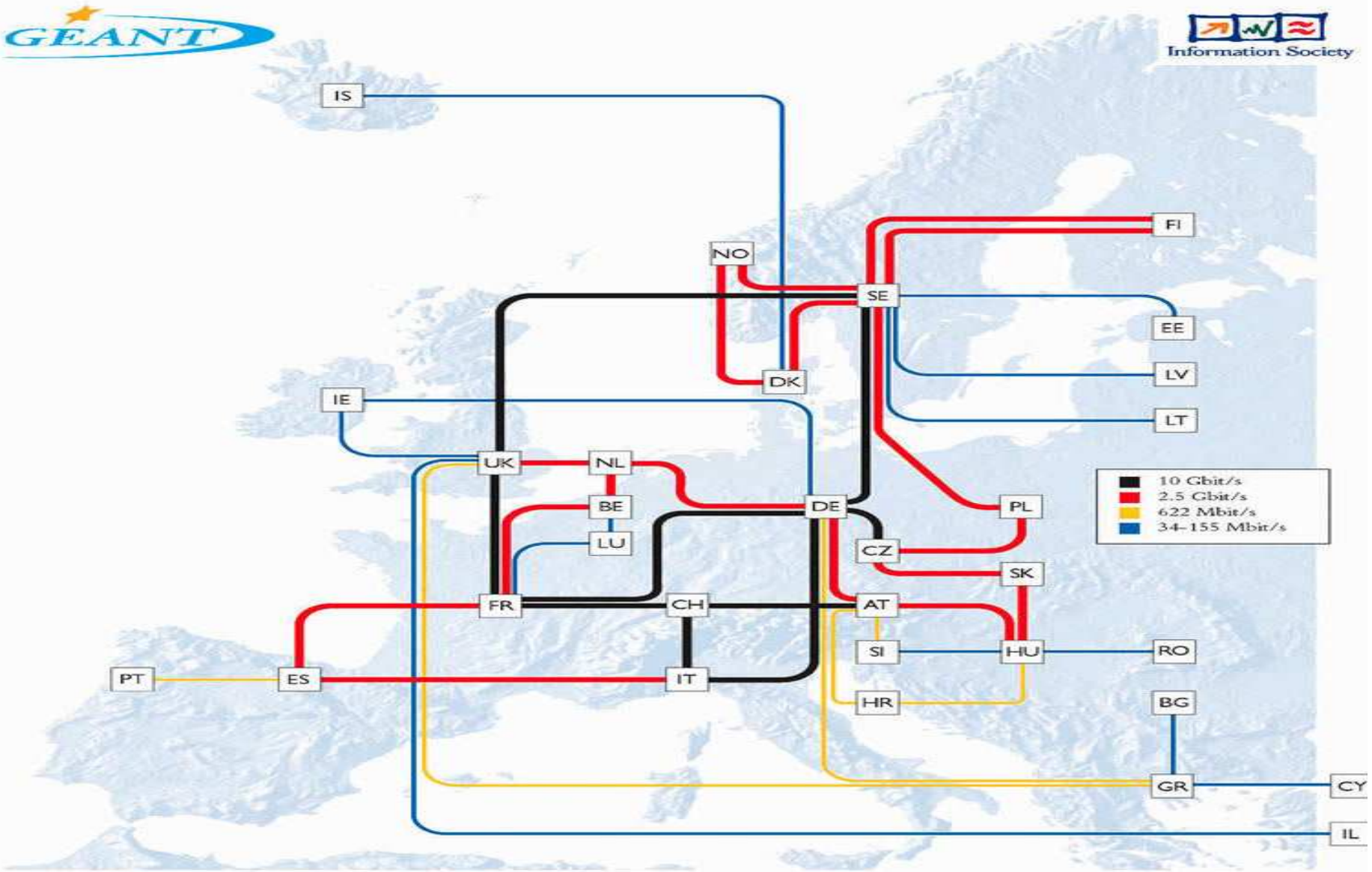


Some orders of magnitude

- Amount of data LHC will provide to us:
  - ~10^8 raw evt/year + same amount from required simulated data
  - Single event size : 1 MB (CMS), 25 MB (Alice)
  - 1 Simulated event : 2 MB (CMS), 2 GB (Alice)
- How many CPUs will we need to analyze all these data:
  - CMS estimate: Tier 0 : 455,000 SI95
  - around 3000 PCs in Tier 0
- How many data will we transfer on the network:
  - Hard to estimate : to transfer everything acquired and simulated in 1 year at Tier 0 by an experiment like CMS (= 3 - 4 PB)

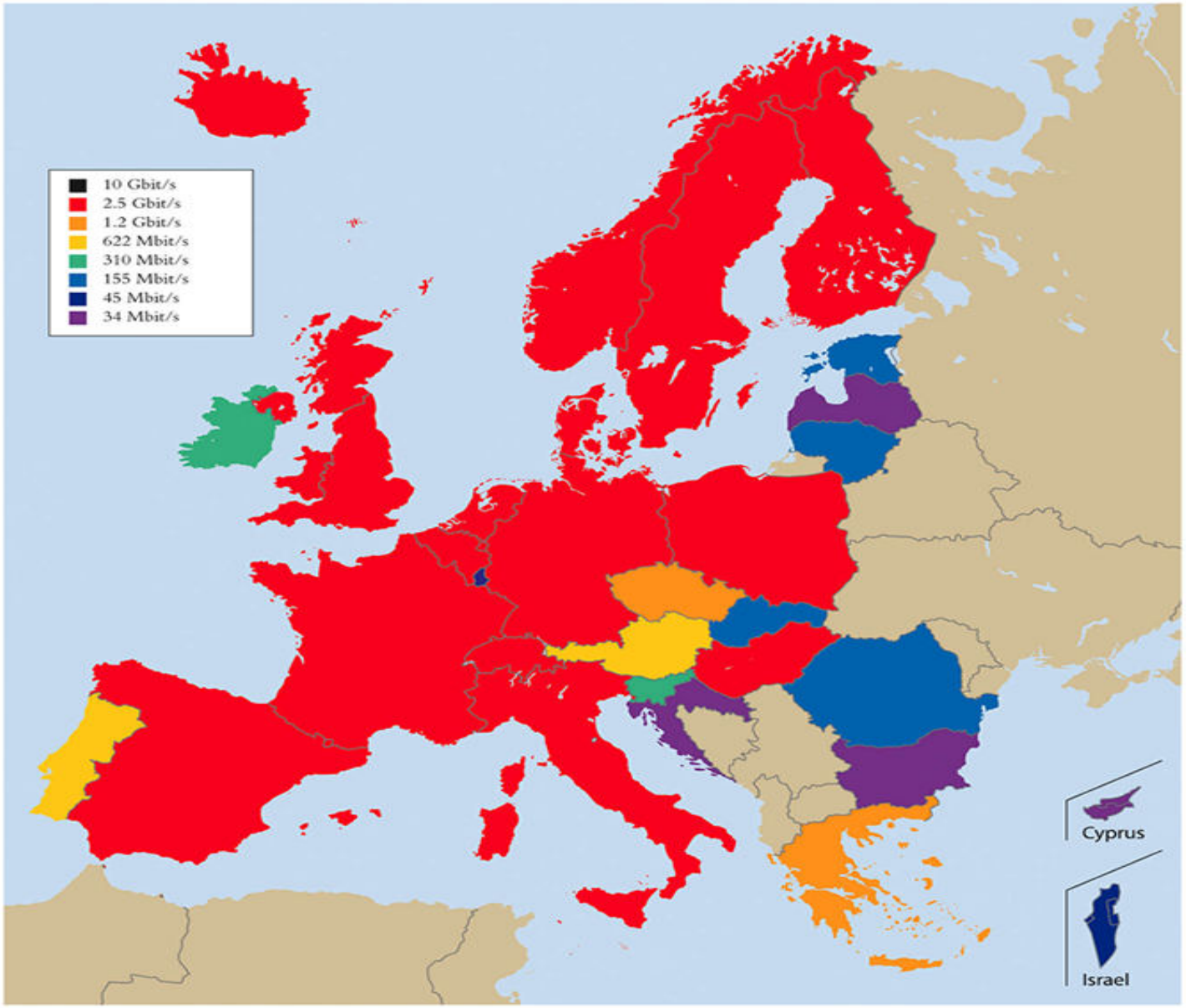
    at 2.5 Gbps one needs ~420 hours = ~ 18 days

    at 155 Mbps one needs ~ 6770 hours = ~ 290 days



AT	Austria	DE	Germany	FR	France	IL	Israel	LV	Latvia	RO	Romania
BE	Belgium	DK	Denmark*	GR	Greece	IS	Iceland*	NL	Netherlands	SE	Sweden*
BG	Bulgaria†	EE	Estonia	HR	Croatia†	IT	Italy	NO	Norway*	SI	Slovenia
CH	Switzerland	ES	Spain	HU	Hungary	LT	Lithuania	PL	Poland	SK	Slovakia
CY	Cyprus	FI	Finland*	IE	Ireland	LU	Luxembourg	PT	Portugal	UK	United Kingdom
CZ	Czech Republic										

† Planned connection * Connections between these countries are part of NORDUnet (the Nordic regional network)


10 Gbit/s
2.5 Gbit/s
1.2 Gbit/s
622 Mbit/s
310 Mbit/s
155 Mbit/s
45 Mbit/s
34 Mbit/s
Cyprus
Israel

What have HEP experiments already done on the EDG testbed

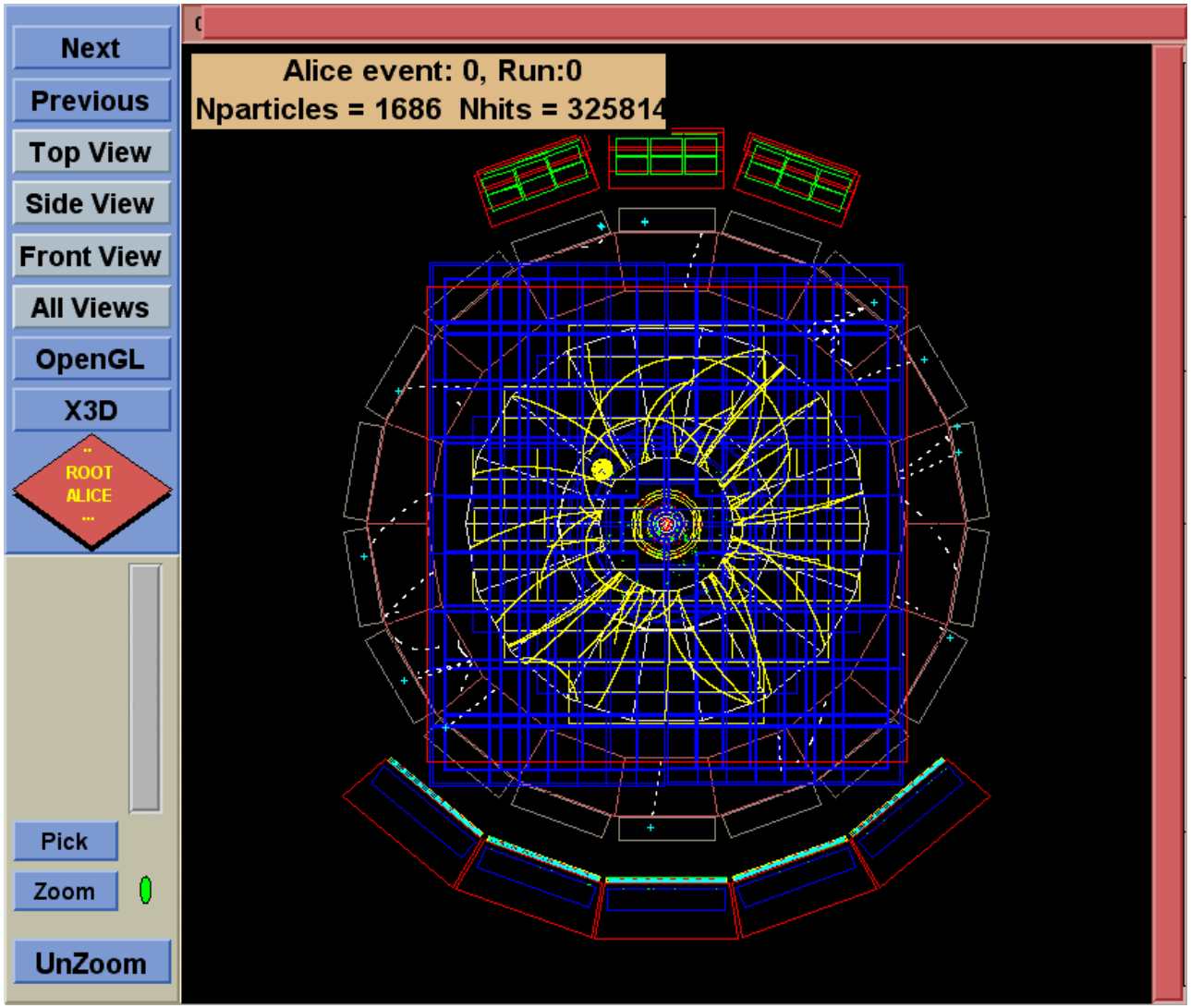
- The EDG User Community has actively contributed to the validation of the first EDG testbed (nov 2001 – feb 2002)
- All four LHC experiments have ran their software (although in some cases in a preliminary version) to perform the basics operations supported by the testbed 1 features provided by the EDG middleware
- Validation included job submission (JDL), output retrieval, job status query, basic data management operations (file replication, register into replica catalogs), check of possible s/w dependencies or incompatibility (e.g. missing libs, rpms) problems
- Everything has been reported in

  "testbed 1 assessment by HEP applications" (D8.2)

  DataGrid-08-D8.2-0111-3-1

(http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=334920&version=3.1)

The first ALICE simulated event on the testbed (january 2002)



An (incomplete) list of the HEP-related executables

- Alroot : generate, Display ALICE events
- DICE : generate ATLAS events
- Phythia, CMSIM, ORCA: generate CMS events
- Brunel,GAUDI,SICBMC: generate LHCb events
- PAW, PATCHY, CERNlibs: use CERN common lib analysis programs
- ROOT : object oriented framework for data analysis and data access, storage
- Objectivity : OODBMS
- GEANT3 : event reconstruction for simulated data

HEP use cases for EDG GRID

- The HEP community is making a big effort to study and catalogue Use Cases to describe its typical way of working in a distributed computing model architecture
- In EDG a set of preliminary interviews with experiment representatives have been carried out in within EDG WP8 to compile a detailed preliminary list of HEP experiments use cases, in view of the possible implementation of a HEP common application layer.
- Common use cases for the 4 LHC collaborations have been reported in the document **DataGrid-08-TEN-0201-1-14** available from EDMS (EDMS id 341682)

  (http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=341682&version=1)

  "common use cases for a HEP common application layer"
- UML modelling used to define classes and methods in an object oriented analysis (OOA) approach

What do HEP experiments want to do on the GRID in the long term ?

- **Production:**
  - Simulation (Monte Carlo generators).
  - Reconstruction (including detector geometry …).
  - Event Mixing (bit wise superposition of Signal and Backgrounds).
  - Reprocessing (Refinement, improved reconstruction data production).
  - Production (production of AODs and ESDs starting from Raw data).
    - *Very organized activity, generally centrally managed by prod teams*
- **Physics analysis**:
  - Searches for specific event signatures or particle types.
    (data access can be very sparse, perhaps on the order of one event out of each million).
  - Measurement of inclusive and exclusive cross sections for a given physics channel – Measurement of relevant kinematical quantities
    - *I/O not feasible to organize the input data in a convenient fashion unless one constructs new files containing the selected events .*
    - *the activities are also uncoordinated (not planned in advance) and (often) iterative.*

An example : fully simulated events production

- **MC Generation** : MC simulation of the simulated event (JETSET, HERWIG, ARIADNE, Phythia, ….) : all tracks, their flavour, energy and momentum is known at the origin of the event
- **Reconstruction** : Includes **particle-matter interaction** simulation : simulate flow inside the detector layers and take into account detector geometry
- **Digitization** and **hit-reconstruction** : simulate the detector electric response to the particle crossing it : includes detailed modelling of detector's electric channels response
- **Tracking** : Use data tracking algorithms to reconstruct momentum and energy of particles
- **Resolution estimate** : Compare original data with reconstructed measured values to have estimates of the detector's performances (resolution, efficiencies, purities)

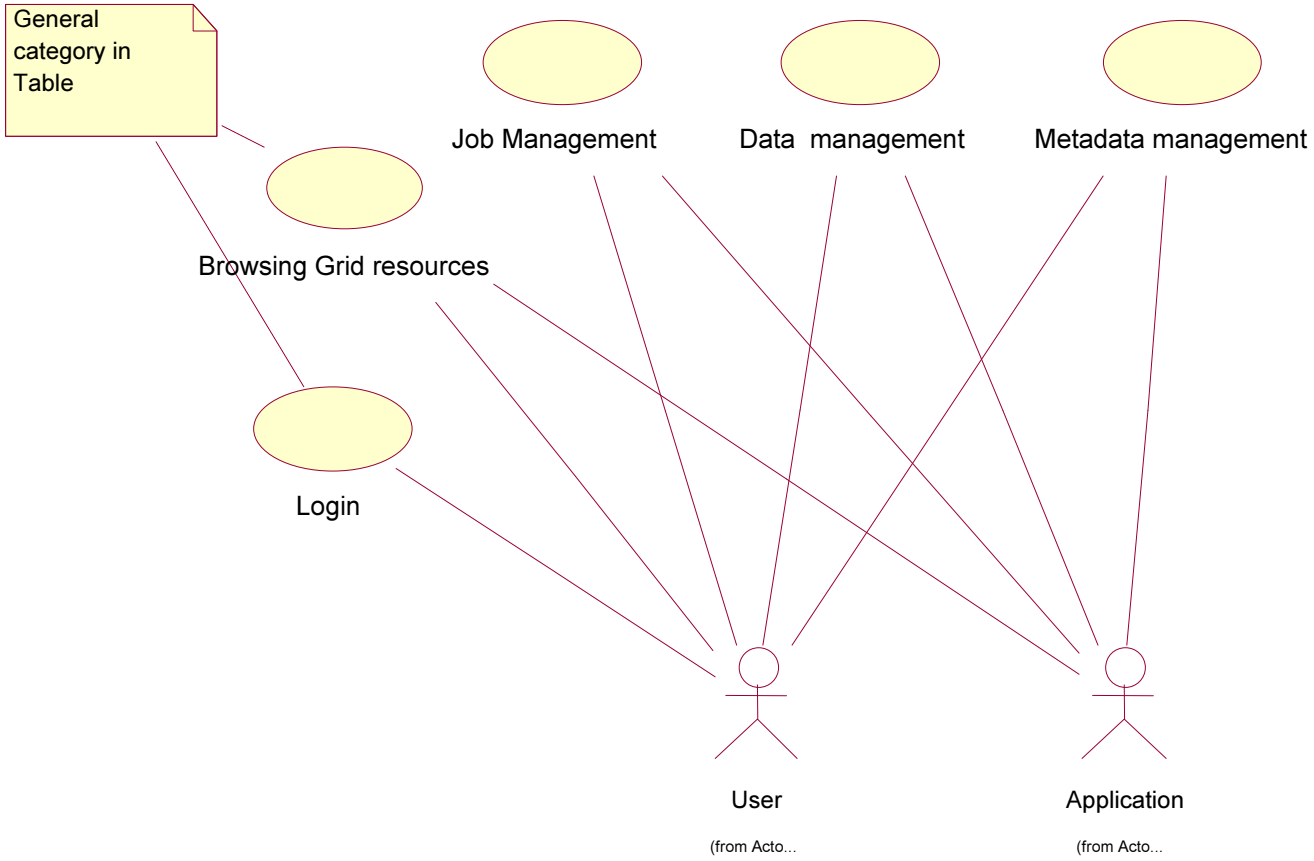
Classification of EDG GRID Use Cases for HEP

Use Case Categories

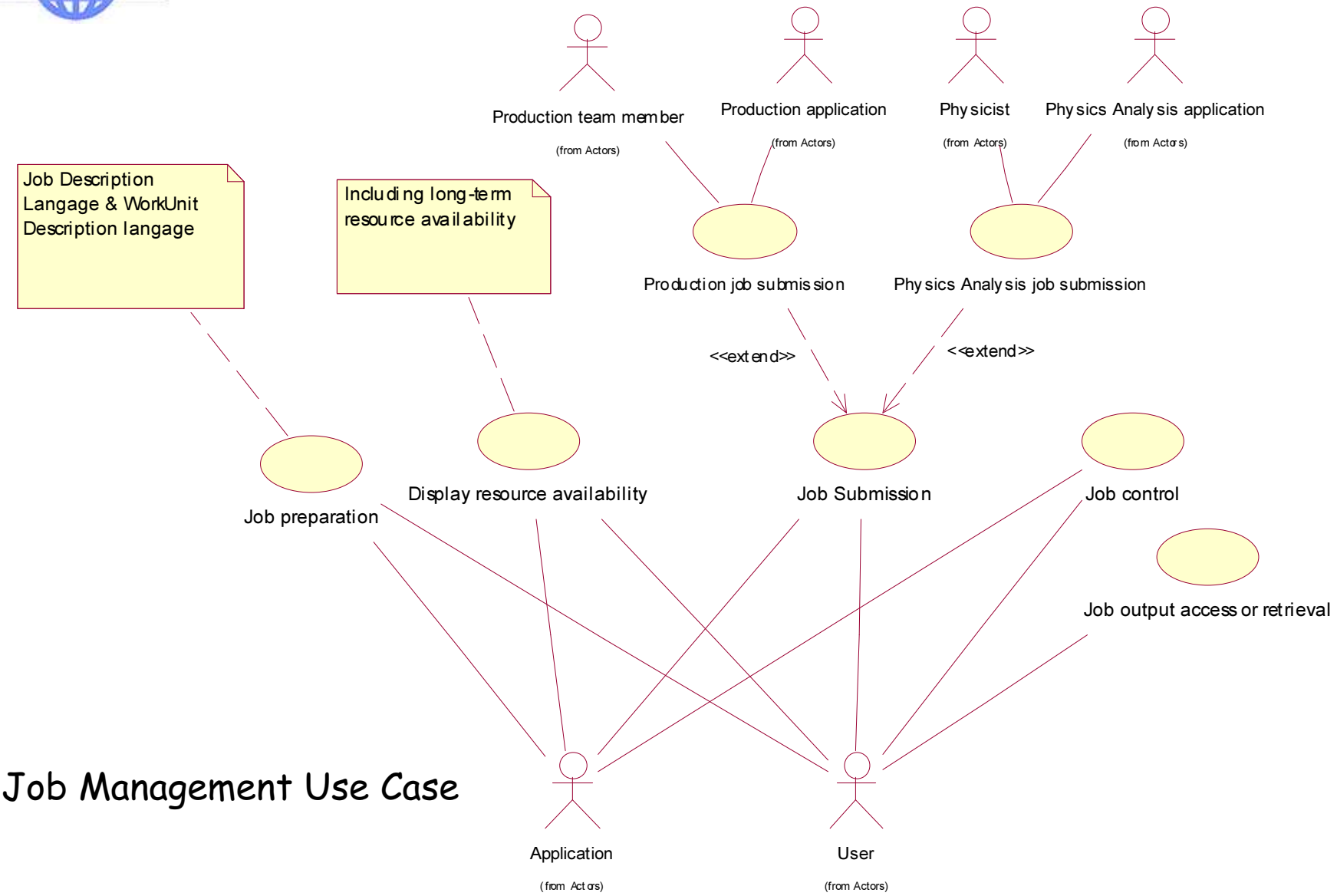
HEPCAL Component	General	Job Management	Data Management	Metadata Management
Grid Usage				
Common	G	CJ	CD	CMD
PROD		PJ	PD	PMD
ANA		AJ	AD	AMD

A deeper view inside HEP use cases (1/2)

- The Global Use Case diagram



A deeper view inside HEP use cases (2/2)



Job Management Use Case

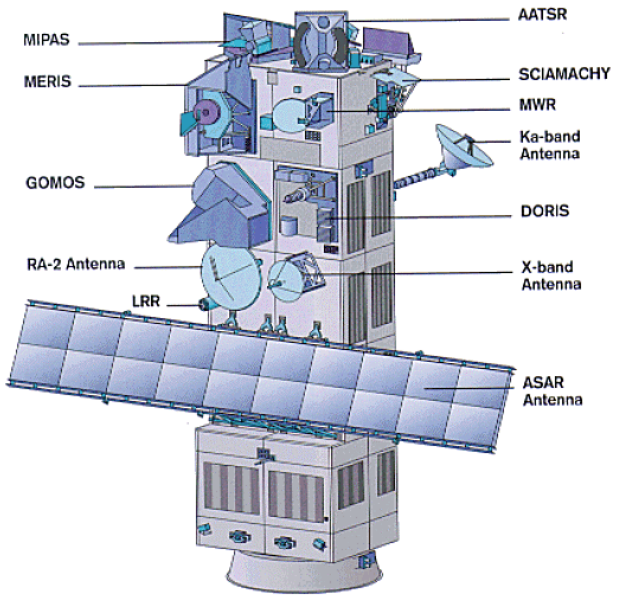
Summary of HEP use cases activities in EDG and open issues

- Results of the preliminary work on HEP use cases by EDG
  - Global "low resolution" classification of Use Cases compiled (large variety of different user's expectations from the GRID encountered)
  - Most difficult issues (Analysis data management, Object-to-File mapping) identified and boundaried
  - Basis for the definition of a HEP CAL architecture stated
  - Priority for EDG is currently demonstrate production data challenges with EDG 1.2, starting from ATLAS.
- On going work on Use Cases and open issues
  - The results of the EDG UCs document have been the starting point for the LCG GRID RTAG on HEP Use Cases, reported in the final report of the RTAG to the SSC. (http://lhcgrid.web.cern.ch/LHCgrid/SC2/RTAG4/finalreport.doc)
  - Work still on-going within Experiments and LCG (LHC Computing GRID) , especially on some topic like the object-to-file mapping in a common approach to interface the GRDI for all LHC experiments

Earth Observation science applications
EO mission and plans

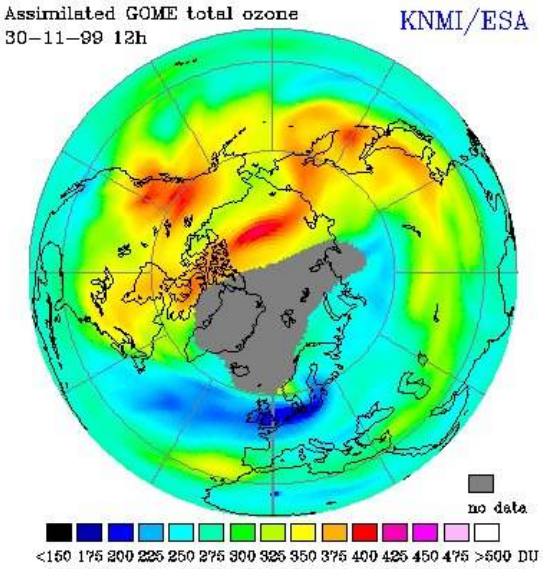
- The EO (ESA,KNMI,IPSL) mission is to **exploit the GRID to execute Earth Observation applications** in a distributed, heterogeneous (and possibly multi-platform) computing environment.
- EO has to deal with **huge amounts of remote sensing satellites data** (ERS-1,2, LANDSAT, ENVISAT) usually on distributed storage elements, whose analysis requires large amounts of CPU resources
- EO scientists has actively been involved in the demonstration of EDG testbed 1 with EO reference applications
- EO has defined its detailed Earth Observation use cases document
- Future:
  going towards a **web services based, multi-tiered integrated architecture based on the data layer, the application server and distributed web clients** to perform data analysis "on-the-fly" on demand to distributed customers (for carrying out processing, storage and retrieval of data products using the Grid infrastructure)

Earth Observation



ESA missions:

- about 100 Gbytes of data per day (ERS 1/2)
- 500 Gbytes, for the next ENVISAT mission (2002).



DataGrid contribute to EO:

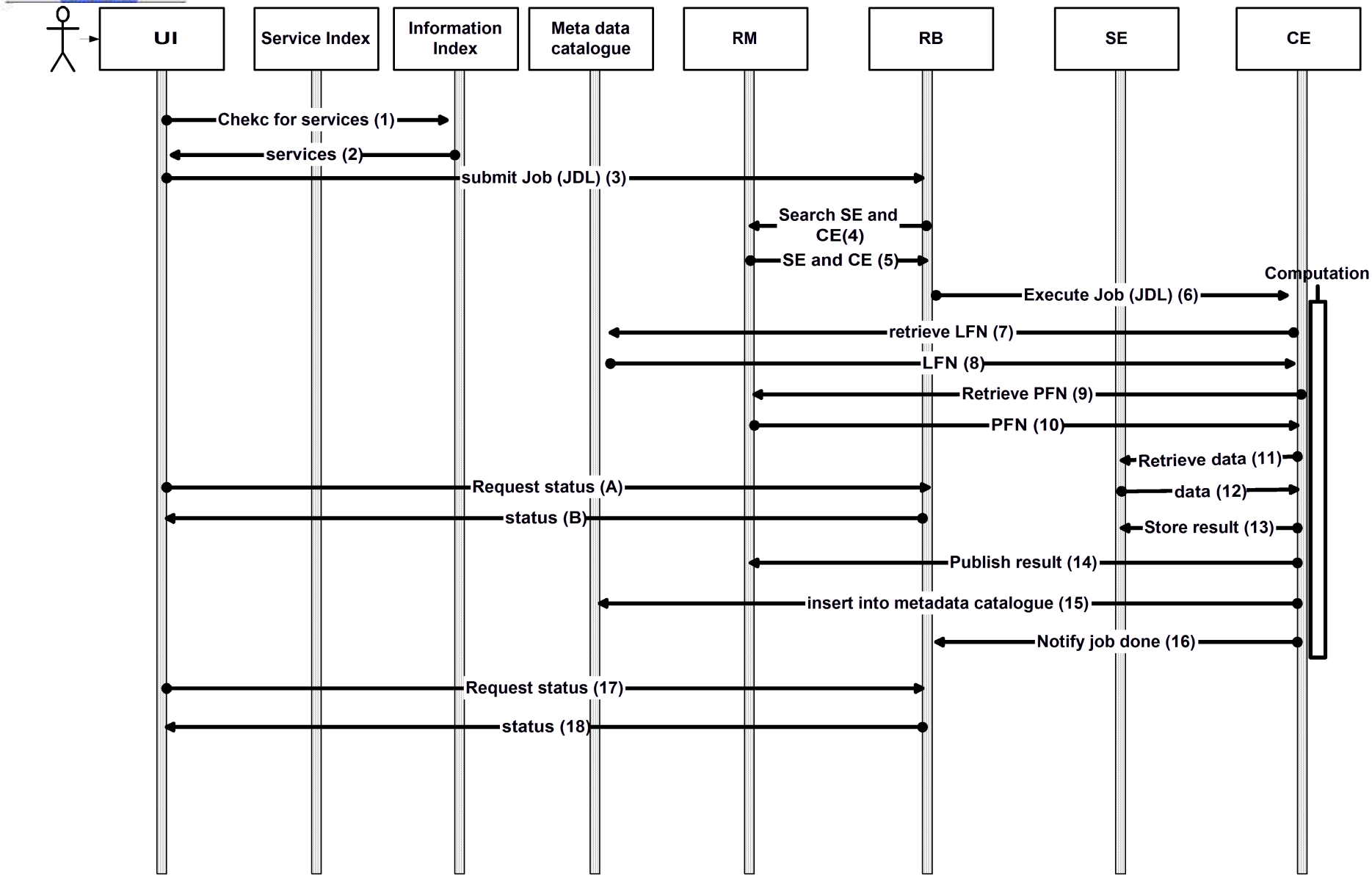
- enhance the ability to access high level products
- allow reprocessing of large historical archives
- improve Earth science complex applications (data fusion, data mining, modelling …)

Source: L. Fusco, June 2001

What do EO applications typically do ?

- EO workpackage used 3 main executables on EDG testbed 1:
  - NNO (ESA) written in IDL Level 1 → Level2 processing
  - OPERA (KNMI) written in C++ Level 1 → Level 2 processing
  - L2-validation-executable (IPSL) written in FORTRAN (L2 valid)
- Input files are Ozone profiles measurement data (*level 1* products: a 15 Mb file contains the measurements taken during a full orbit of the satellite/sensor)
- Output files are O_3 profiles to be analysed by earth scientists to monitor the layer of Ozone in the athmosphere (*level 2* products: a 10-12 kb file containing the results of the L1 data analysis : actual physical quantities for the ozone gas concentration at different pressure levels within a column of atmoshpere at a given location (lat, lon) above the Earths surface)

GOME EO tb1 validation – Sequence Diagram

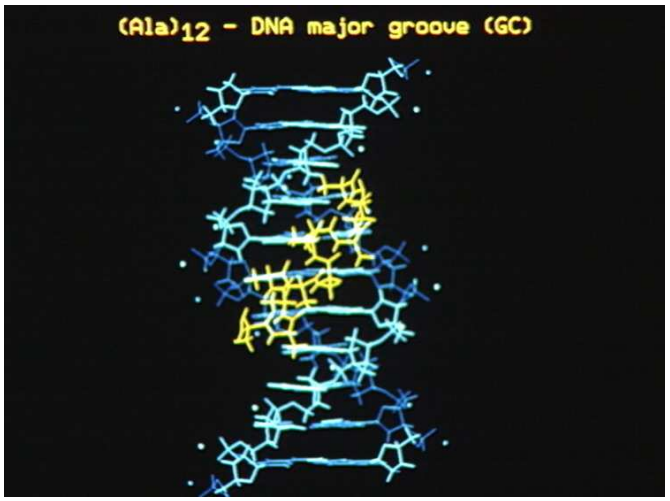


Biology and Bio-informatics applications

- The international community of Biologists has a keen interest in using of bio-informatic algorithms to perform research on the mapping of the human genomic code
- Biologist make use of large, geographically distributed databases with already mapped, identified sequences of proteins belonging to sequences of human genetic code (DNA sequencies)
- Typical goal of these algorithms is to analyse different databases, related to different samplings, to identify similarities or common parts
- dgBLAST (Basic Local Alignment Search Tool) is an example of such an application seeking particular sequences of proteins or DNA in the genomic code

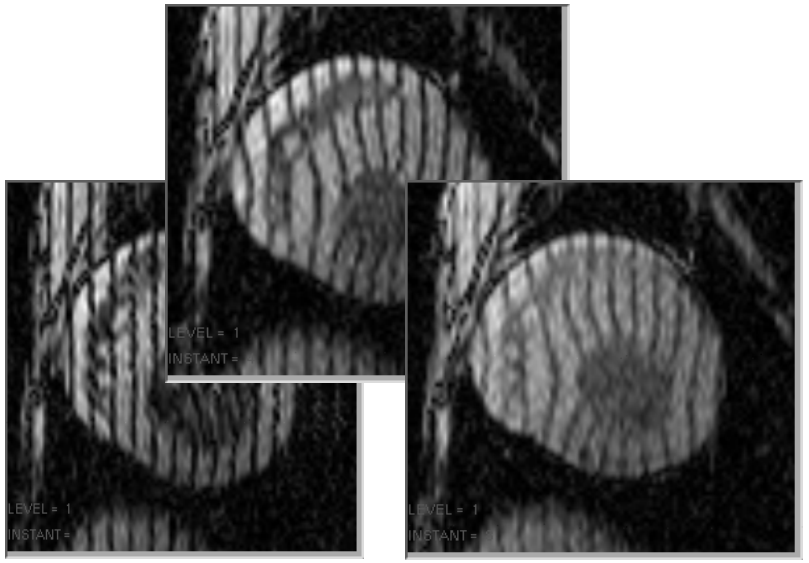
Biomedical Applications

Genomics, post-genomics, and proteomics



Explore strategies that facilitate the sharing of genomic databases and test grid-aware algorithms for comparative genomics

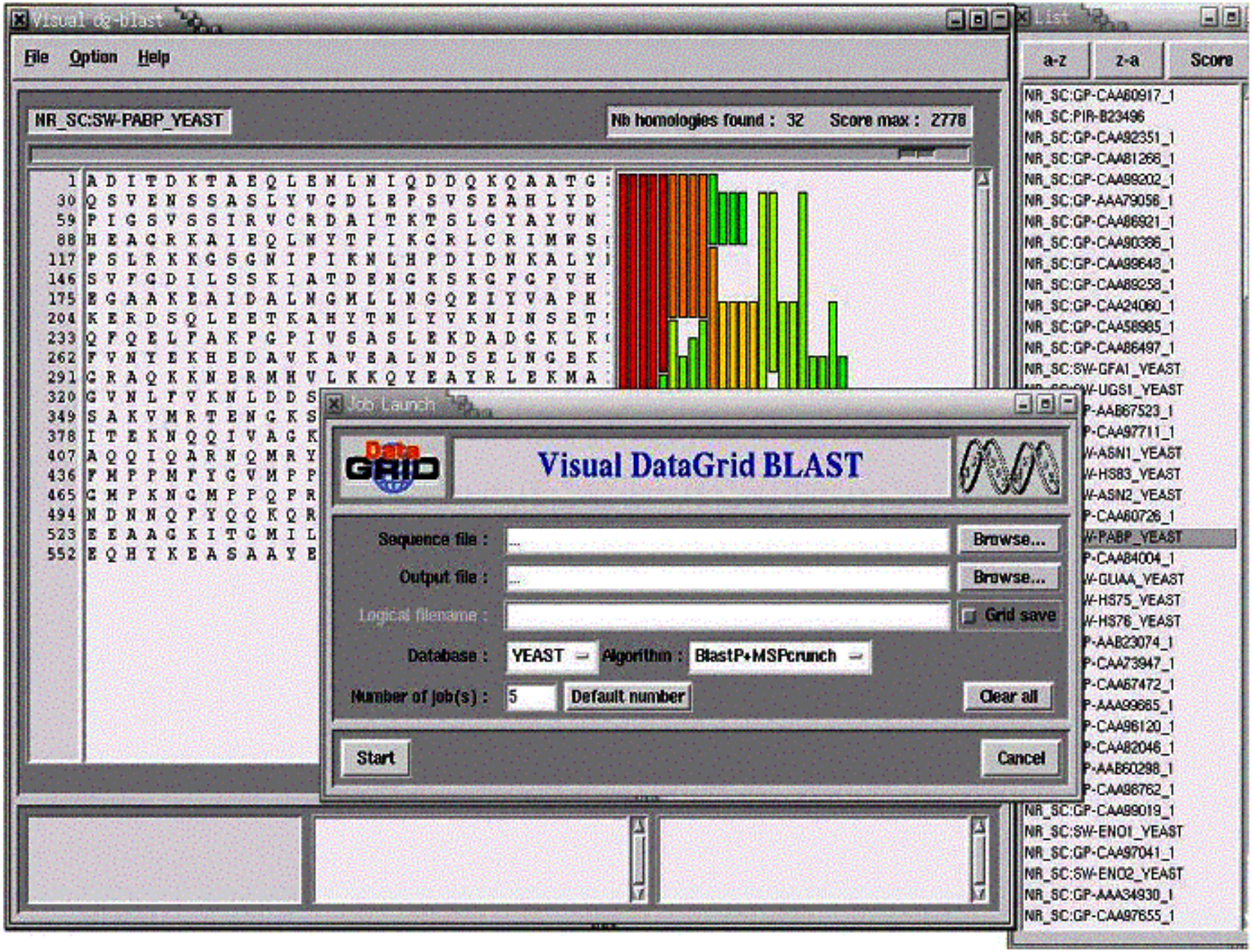
Medical images analysis



Process the huge amount of data produced by digital imagers in hospitals.

Example GRID application for Biology: dgBLAST

- dgBLAST requires as input a given sequence (protein or DNA) to be searched and a pointer to the set of databases to be queried.
- Designed for high speed (trade off vs sensitivity)
- A score is assigned to every candidate sequence found and the results graphically presented
- uses an heuristic algorithm
- detects relationships among sequences which share only isolated regions of similarity.
- **Blastn**: compares a nucleotide query sequence against a nucleotide sequence database
- **Blastp**: compares an amino acids query sequence against a protein sequence


Visual cg-blast
File
Option
Help
NR_SC:SW-PABP_YEAST
Nb homologies found : 32
Score max : 2778
Job Launch
Visual DataGrid BLAST
Sequence file :
Output file :
Logical filename :
Browse...
Browse...
Grid save
Database :
YEAST
Algorithm :
BlastP+MSPcrunch
Number of job(s) :
5
Default number
Clear all
Start
Cancel
List
a-z
z-a
Score
NR_SC:GP-CAA60917_1
NR_SC:PIR-B23496
NR_SC:GP-CAA92351_1
NR_SC:GP-CAA81266_1
NR_SC:GP-CAA99202_1
NR_SC:GP-AAA79056_1
NR_SC:GP-CAA86921_1
NR_SC:GP-CAA90386_1
NR_SC:GP-CAA99648_1
NR_SC:GP-CAA89258_1
NR_SC:GP-CAA24060_1
NR_SC:GP-CAA58985_1
NR_SC:GP-CAA86497_1
NR_SC:SW-GFAI_YEAST
NR_SC:SW-PABP_YEAST
NR_SC:GP-CAA84004_1
NR_SC:GP-CAA97711_1
NR_SC:GP-CAA99019_1
NR_SC:SW-ENO1_YEAST
NR_SC:GP-CAA97041_1
NR_SC:SW-ENO2_YEAST
NR_SC:GP-AAA34930_1
NR_SC:GP-CAA97655_1

Summary

- HEP, EO and Biology users have deep interest in the deployment and the actual availability of the GRID, boosting their computer power and data storage capacities in an unprecedented way.

- EDG has demonstrated the feasibility of the GRID by means of the distributed EDG testbed, to allow effective GRID computing to users belonging to three big families of target Virtual Organizations

- Many challanging issues are facing us :

  - demonstrate effective massive productions on the EDG testbed

  - keep up the pace with next generation grid computing evolutions, implementing or interfacing them to EDG

- further develop middleware components for all EDG workpackages to address growing user's demands