The EU DataGrid project: status and perspective

Fabrizio Gagliardi

Project Leader

CERN

Fabrizio.Gagliardi@cern.ch

Outline

- Background
- Grid computing: the vision
- The EU DataGrid project
- First project results
- Conclusion

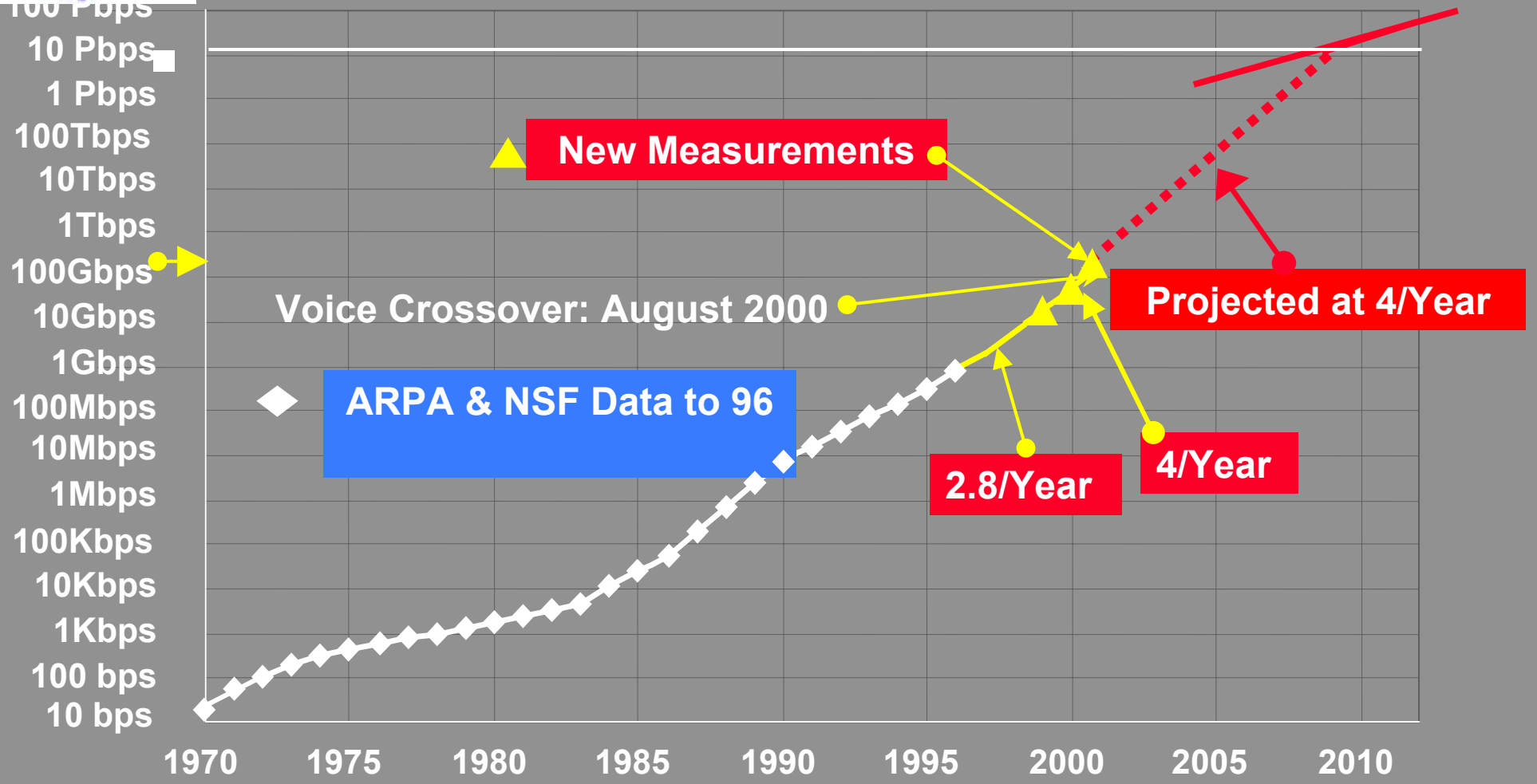
Background

- Grid computing is not a new invention
- Natural evolution of distributed computing, metacomputing…
  - To increase overall computing capacity and data distribution
- Driven by technology push and application pull

Technology push

- Networks: QoS, availability, cost
  - Cost and QoS of high performance networks evolution
  - Internet exponential grow (traffic doubles every 8-12 months)
- Processors:
  - The Moore's law: CPU power double every 18 months
- Traditional supercomputers evolving towards commodity computer clusters

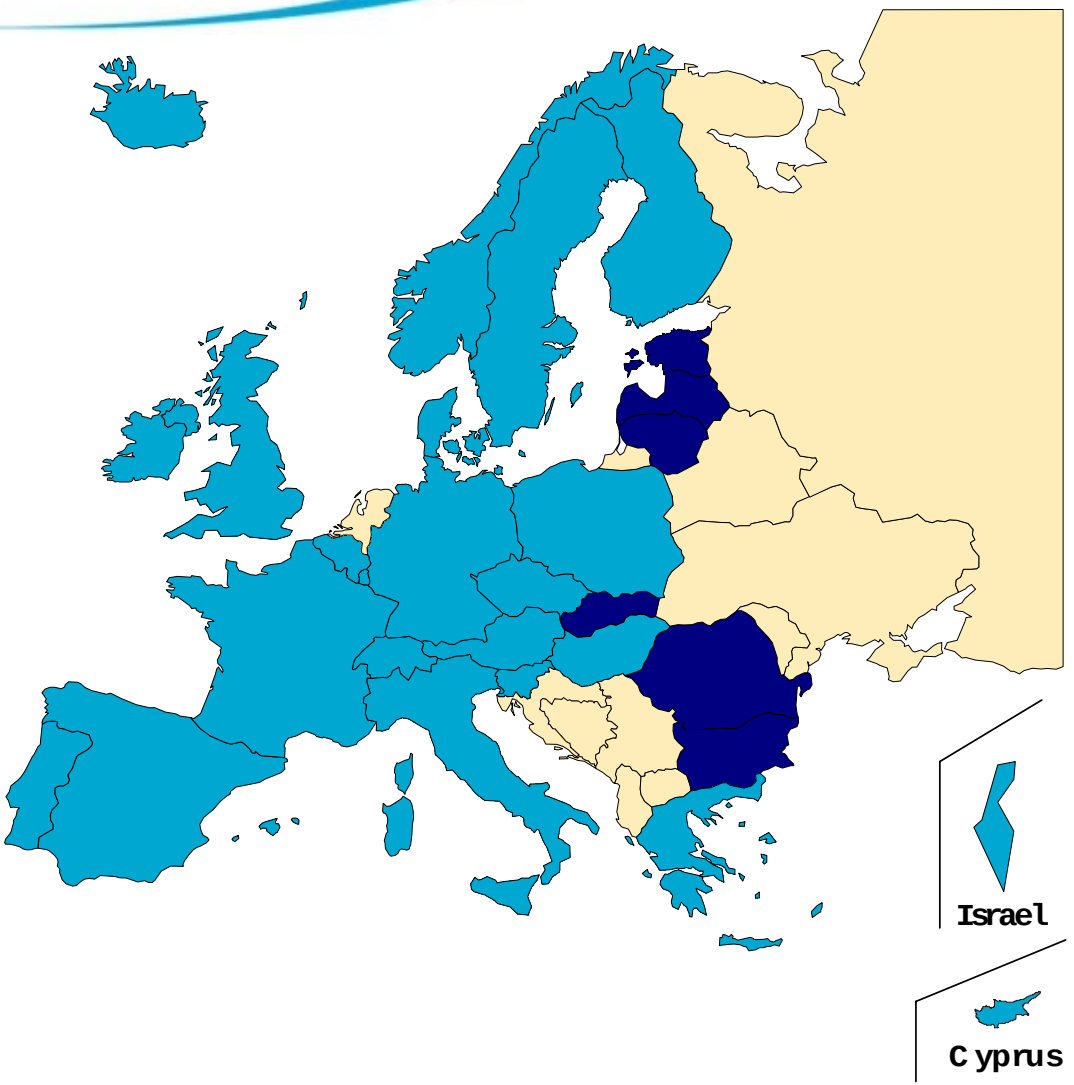
Total U.S. Internet Traffic



U.S. Internet Traffic

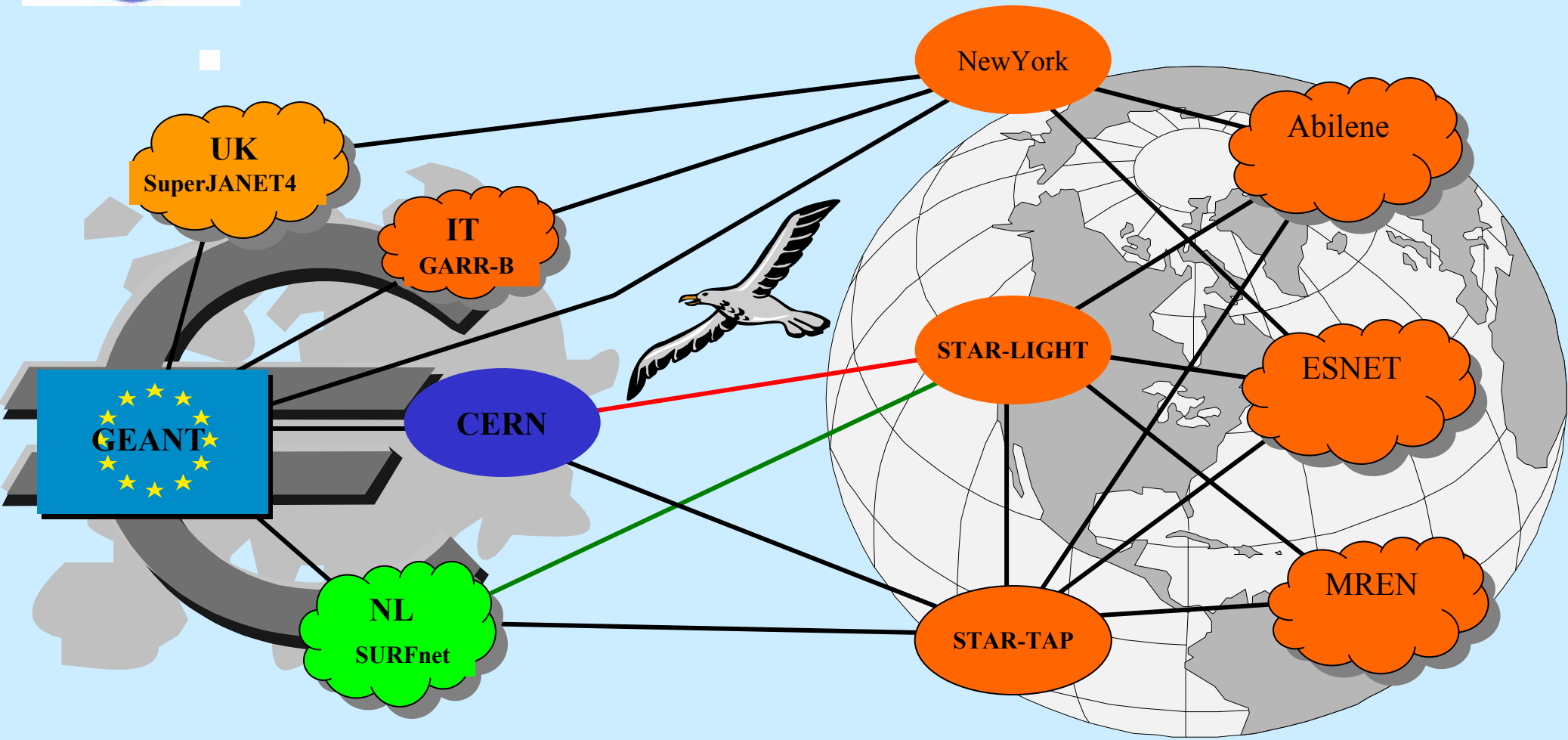
Source: Roberts et al., 2001

Participants and Geography



Planned logical topology,
October 2001
EE
LV
LT
155
34
45
IL
45
UK
SE
PL
155
IE
NL
10G
2.5G
Other
BE
155
155
LU
155
622
FR
DE
CZ
T-Systems
COLT
Telia
ES
SK
622
PT
BG
34
GR
CH
IT
AT
HU
34
RO
34
34
622
155
CY
622
HR
SI

DataTAG project



Asian Pacific Grid

- Common Framework for Asia-Pacific Grid researchers
- Represent AP interests to GGF
- Collaborate with APAN/TransPAC
- *Voluntary framework: Not a project funded from single source*

CERN latest supercomputer

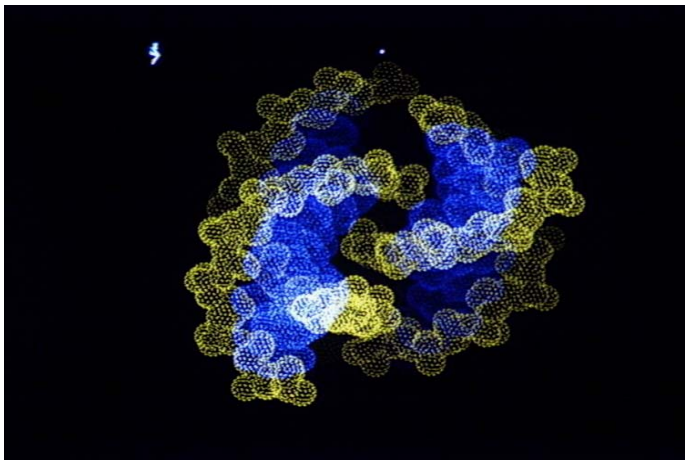
CERN

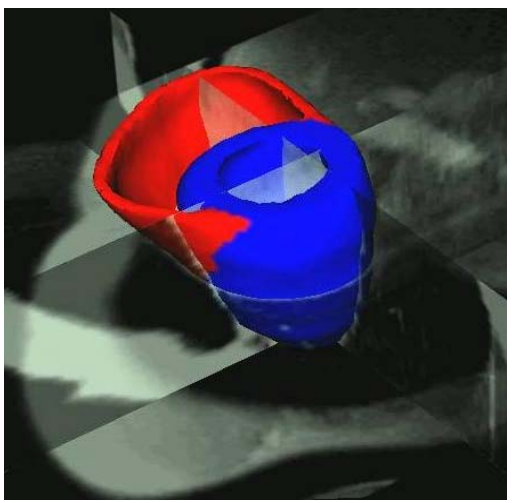
Application pull

- Most of scientific research becomes "Data Intensive"
- Emergence of the concept of Virtual Organisations:
  - Geographically dispersed sources of data and end-users
- Application pull for high performance and widely distributed intensive data processing

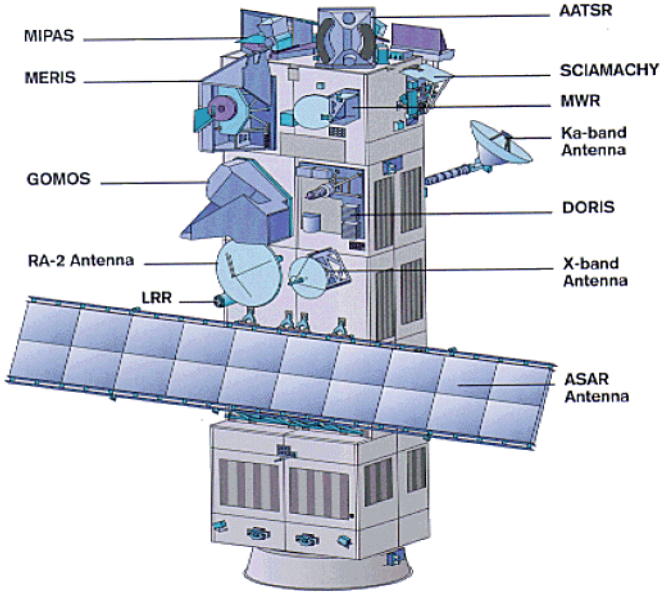
Biomedical applications

- Data mining on genomic databases (exponential growth)
- Indexing of medical databases (Tb/hospital/year)
- Collaborative framework for large scale experiments (e.g. epidemiological studies)
- Parallel processing for
  - Databases analysis
  - Complex 3D modelling



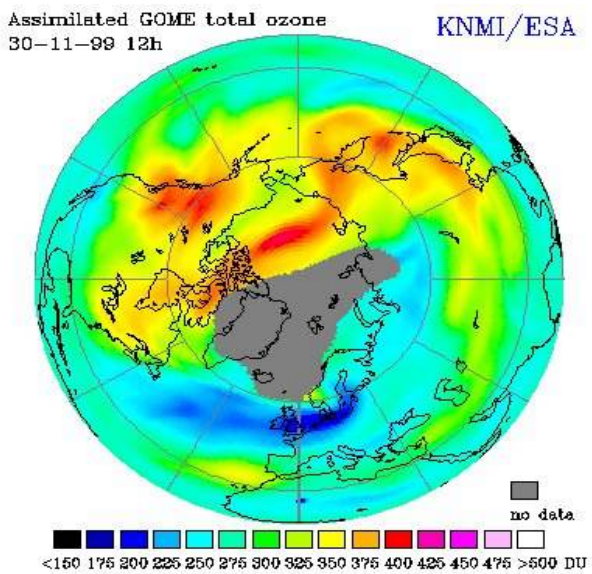


Earth Observations



ESA missions:

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

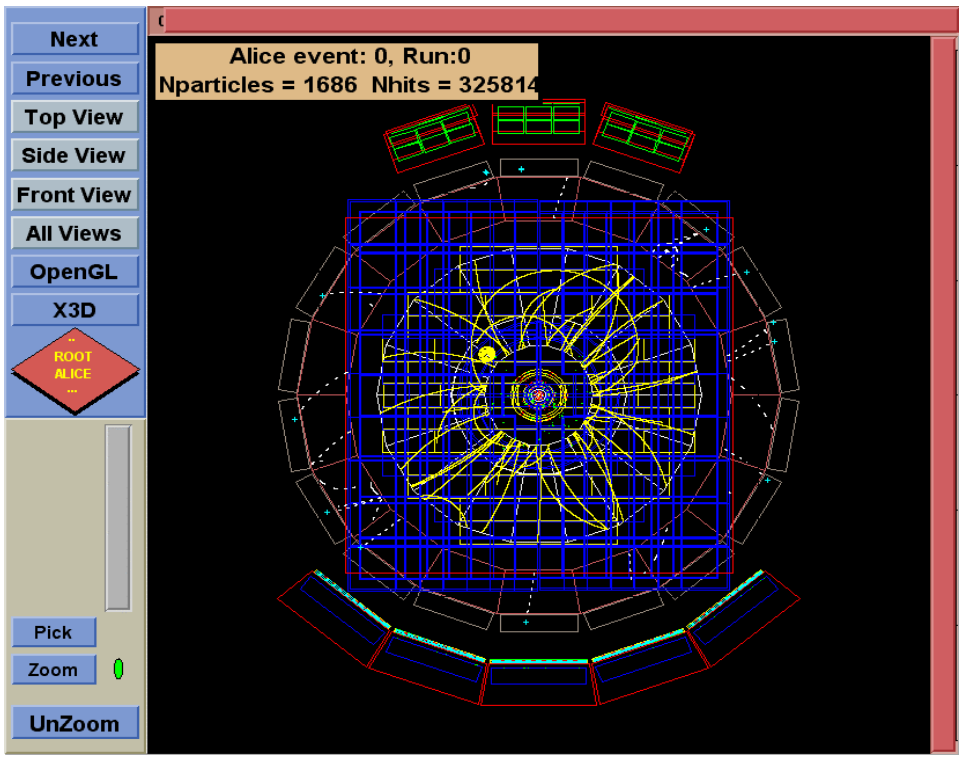


EO requirements for the Grid:

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

Particle Physics

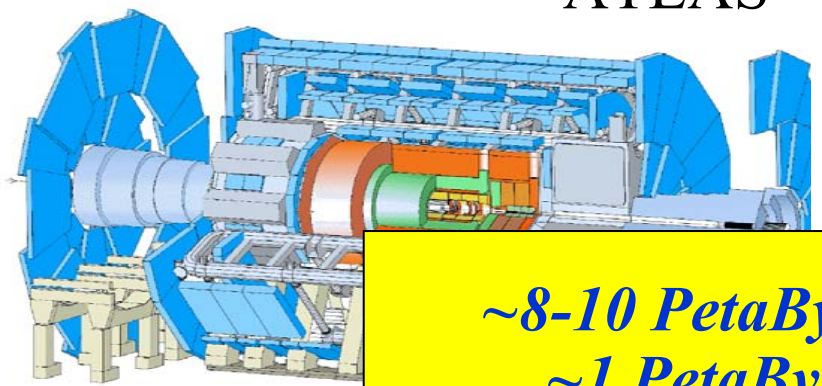
- Simulate and reconstruct complex physics phenomena millions of times



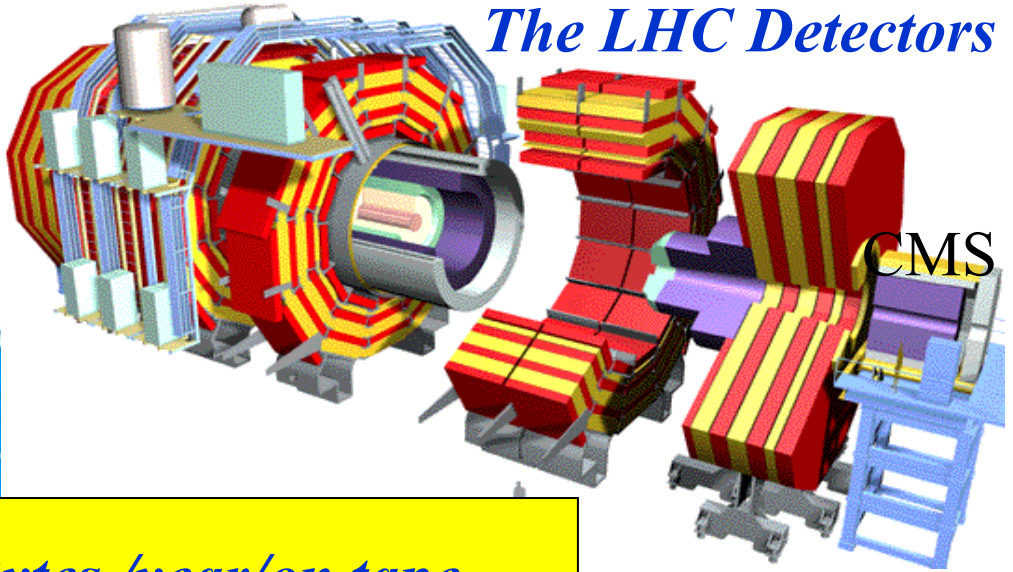
The LHC problem:
a typical Grid application

The LHC Detectors

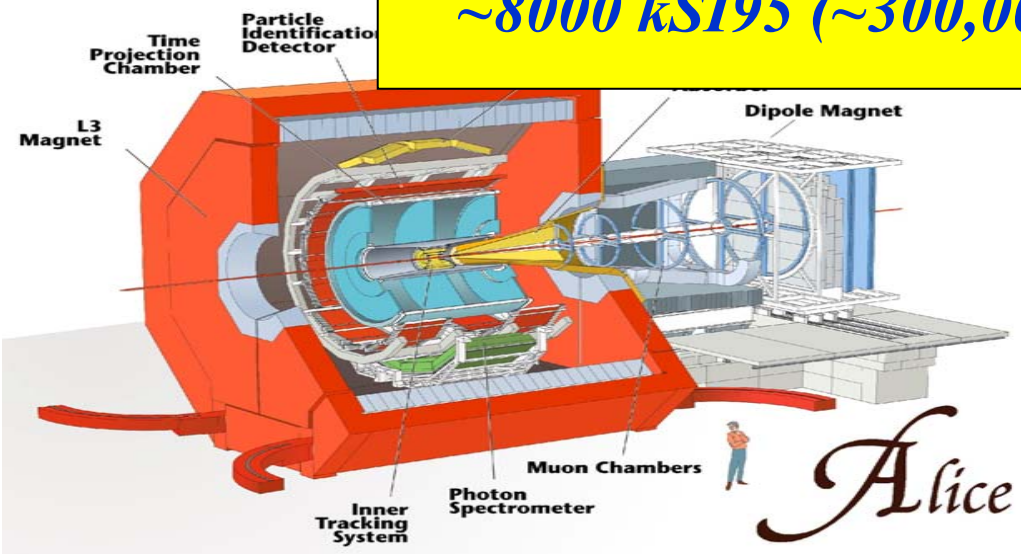
ATLAS

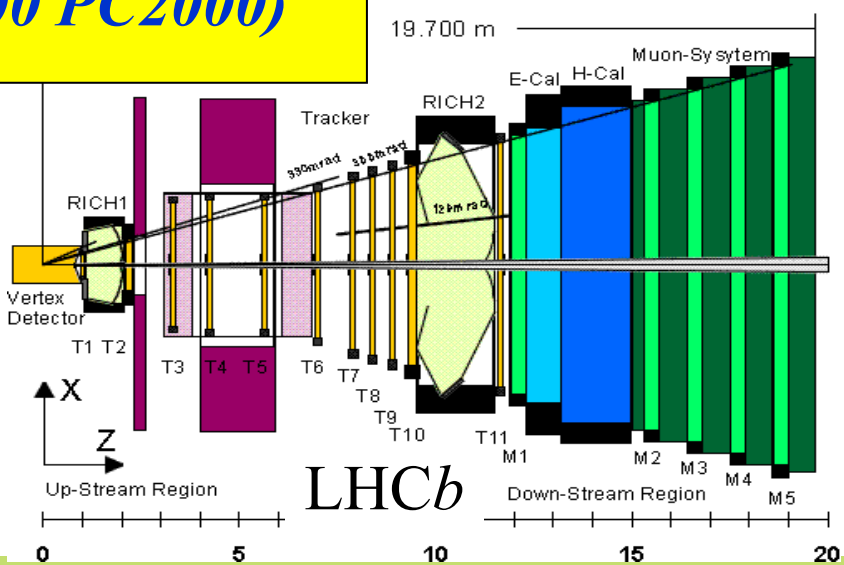


CMS



~8-10 PetaBytes /year/on tape
~1 PetaByte/year/on disk
~8000 kSI95 (~300,000 PC2000)





CERN's Network in the World

Europe: 267 institutes, 4603 users
Elsewhere: 208 institutes, 1632 users

EU DataGrid Project Objectives

- To build on the emerging Grid technology to develop a sustainable computing model for effective share of computing resources and data
- Specific project objectives:
  - Middleware for fabric & Grid management (mostly funded by the EU)
  - Large scale testbed (mostly funded by the partners)
  - Production quality demonstrations (partially funded by the EU)
- To collaborate with and complement other European and US projects
- Test and demonstrator of EU RN/Geant
- Contribute to Open Standards and international bodies:
  - Co-founder of Global GRID Forum and host of GGF1 and GGF3
  - Industry and Research Forum for dissemination of project results

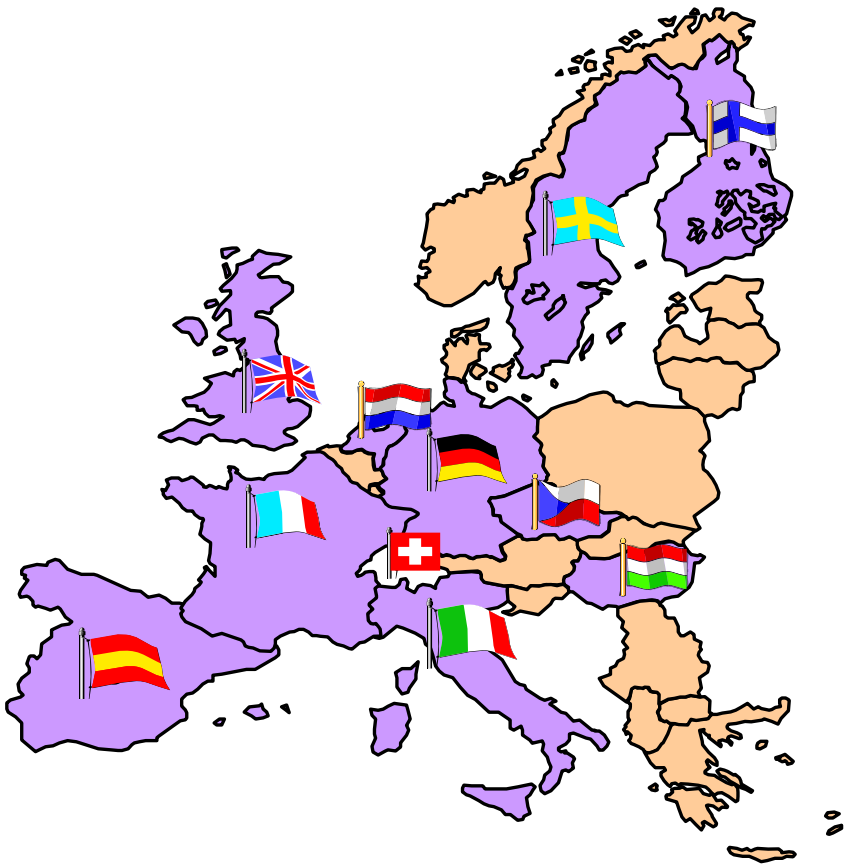
Main Partners

- CERN – International (Switzerland/France)
- CNRS - France
- ESA/ESRIN – International (Italy)
- INFN - Italy
- NIKHEF – The Netherlands
- PPARC - UK

Assistant Partners

Industrial Partners

- Datamat (Italy)
- IBM-UK (UK)
- CS-SI (France)



Research and Academic Institutes

- CESNET (Czech Republic)
- Commissariat à l'énergie atomique (CEA) – France
- Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI)
- Consiglio Nazionale delle Ricerche (Italy)
- Helsinki Institute of Physics – Finland
- Institut de Fisica d'Altes Energies (IFAE) - Spain
- Istituto Trentino di Cultura (IRST) – Italy
- Konrad-Zuse-Zentrum für Informationstechnik Berlin - Germany
- Royal Netherlands Meteorological Institute (KNMI)
- Ruprecht-Karls-Universität Heidelberg - Germany
- Stichting Academisch Rekencentrum Amsterdam (SARA) – Netherlands
- Swedish Research Council - Sweden

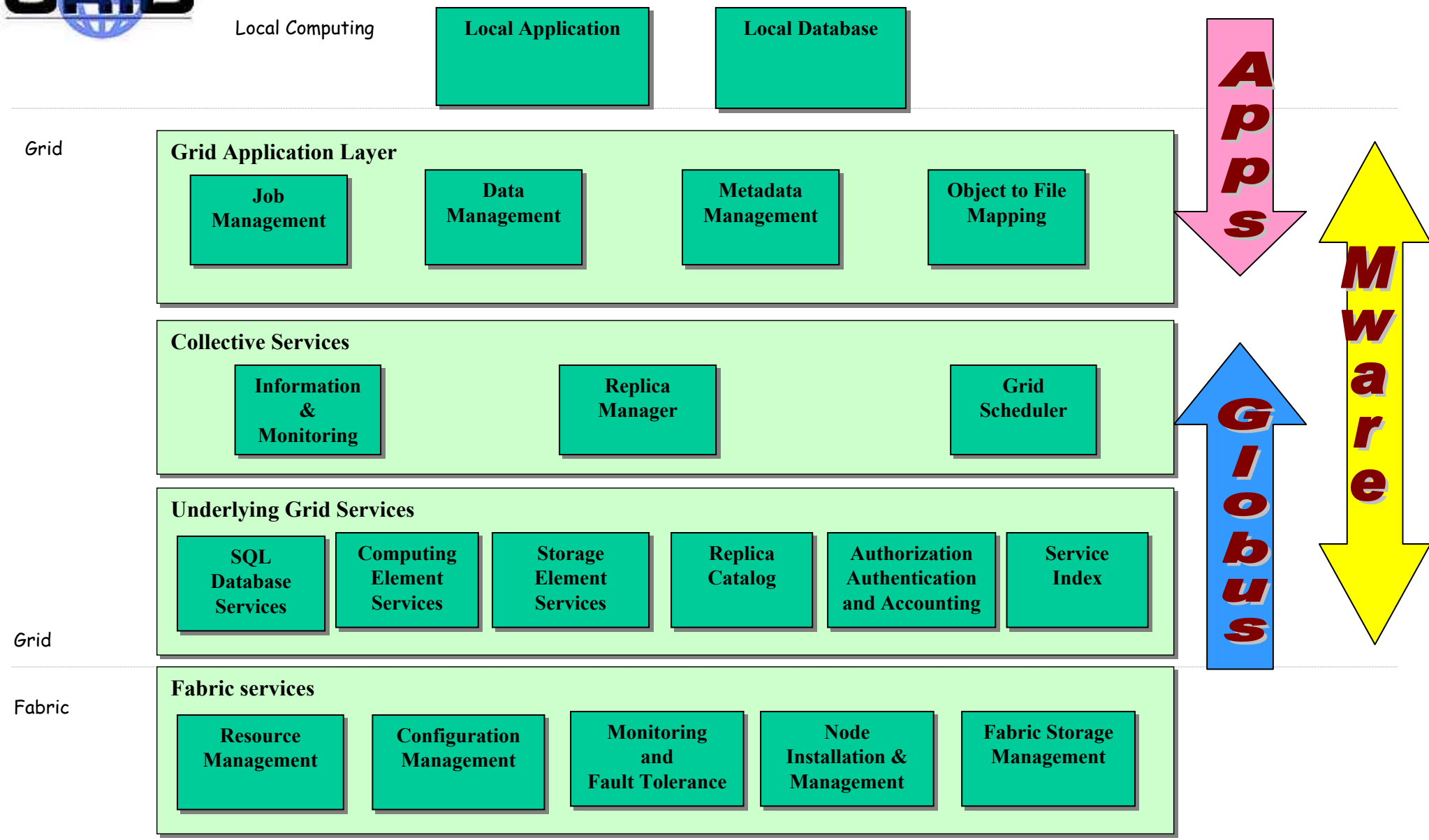
Project scope

- 9.8 M Euros EU funding over 3 years
- 90% for middleware and applications (HEP, EO and Biomedical)
- Three year phased developments & demos (2001-2003)
- Extensions (time and funds) on the basis of first successful results:
  - DataTAG (2002-2003)
  - CrossGrid (2002-2004)
  - GridStart (2002-2004)

Project Schedule

- Project started on 1/1/2001
- TestBed 0 (early 2001)
  - International test bed 0 infrastructure deployed
    - Globus 1 only - no EDG middleware
- TestBed 1 (now)
  - First release of EU DataGrid software to defined users within the project:
    - HEP experiments, Earth Observation, Biomedical applications
- **Project successfully reviewed by EU on March 1st 2002**
- TestBed 2 (September-October 2002)
  - Builds on TestBed 1 to extend facilities of DataGrid
- TestBed 3 (March 2003) & 4 (September 2003)
- Project end on 31-12-2003

EU DataGrid Architecture



TestBed 1 Sites Status

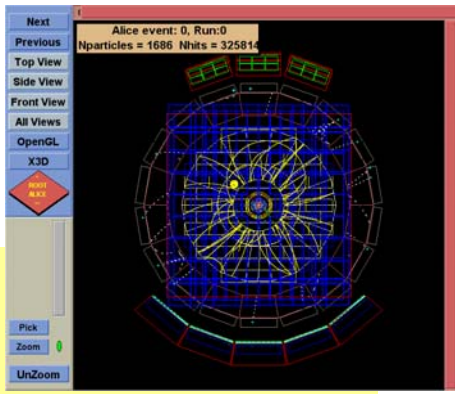
Web interface showing status of servers at testbed 1 sites

Min & Max Status = No Status Normal TCP failed Ping failed

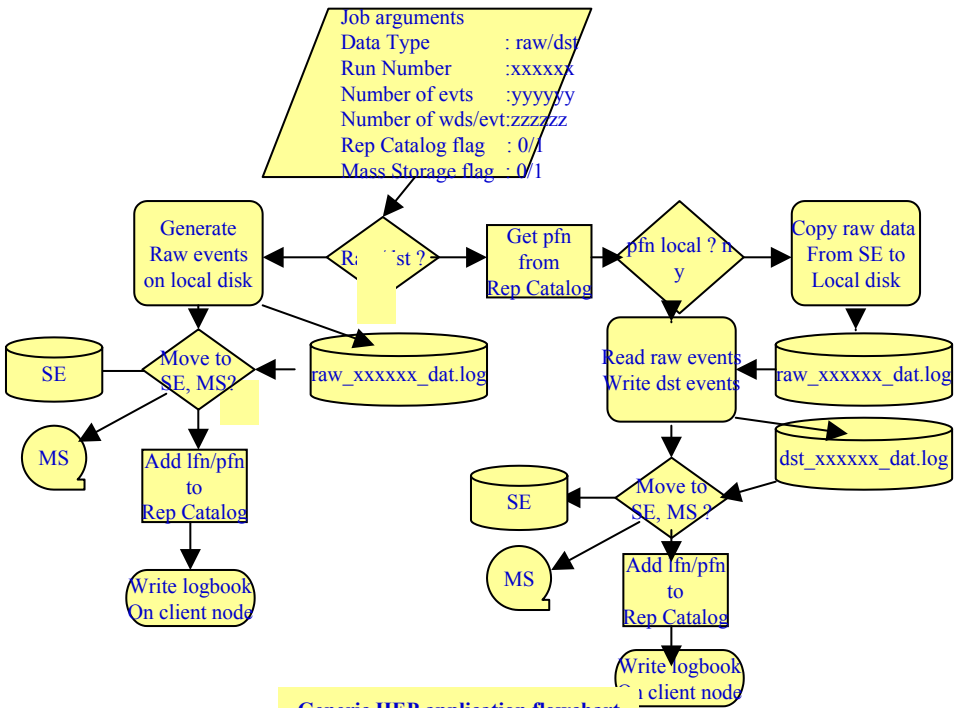
Initial testbed usage

Physicists from LHC experiments submit jobs with their application software that uses:

- User interface (job submission language etc.)
- Resource Broker & Job submission service
- Information Service & Monitoring
- Data Replication



First simulated ALICE event generated by using the DataGrid Job Submission Service

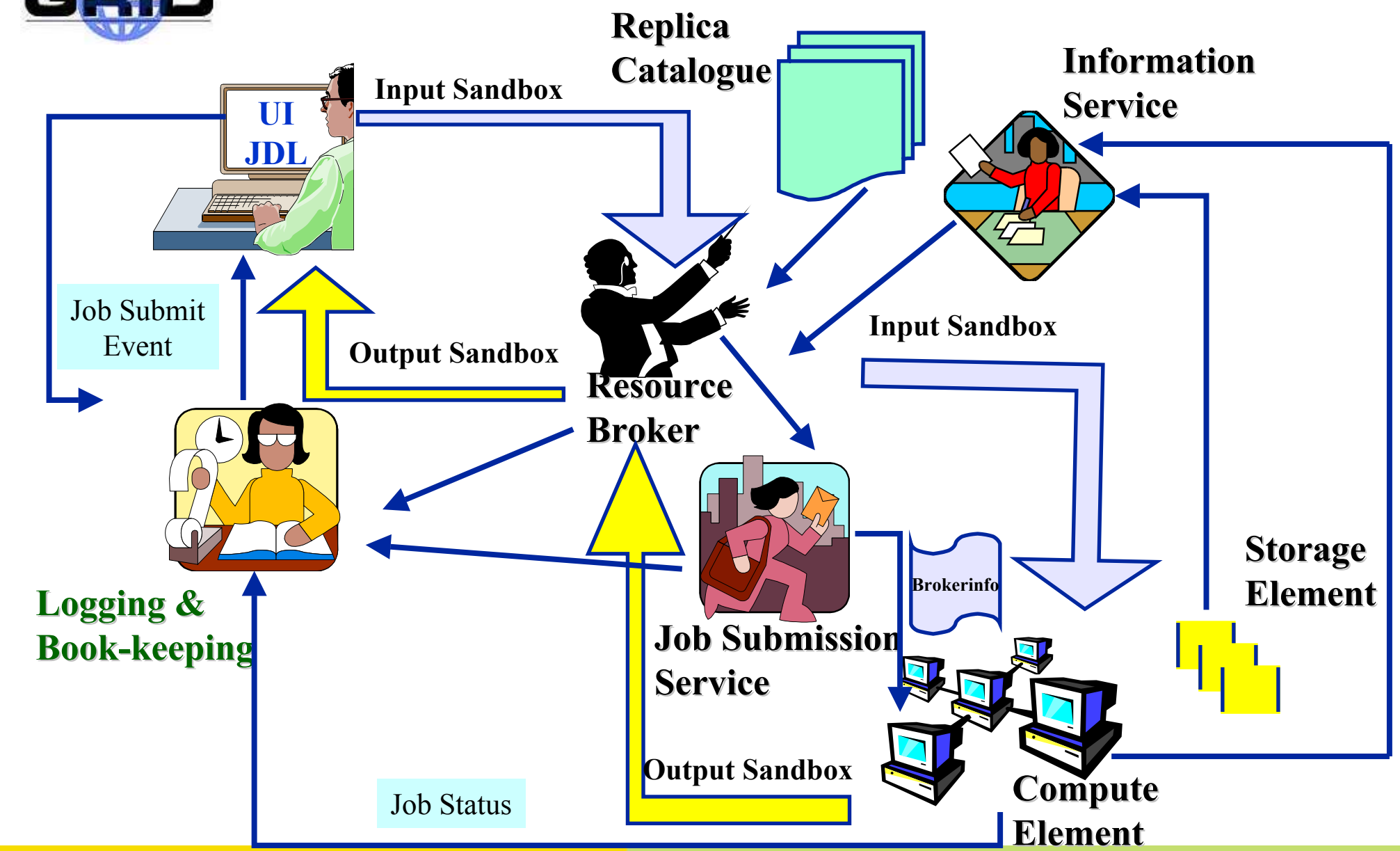


Generic HEP application flowchart

```
[reale@testbed006 JDL]$ dg-job-submit gridpawCNAF.jdl
Connecting to host testbed011.cern.ch, port 7771
Transferring InputSandbox files...done
Logging to host testbed011.cern.ch, port 1
5830
=========dg-job-submit Success ============
 The job has been successfully submitted to the Resource Broker.
 Use dg-job-status command to check job current status. Your job identifier (dg_jobId) is:

https://testbed011.cern.ch:7846/137.138.181.253/185337169921026?testbed011.cern.ch:7771
=======================================
[reale@testbed006 JDL]$ dg-job-get-output
        https://testbed011.cern.ch:7846/137.138.181.253/185337169921026?testbed011.cern.ch:777
        1
Retrieving OutputSandbox files...done
============ dg-get-job-output Success ============
 Output sandbox files for the job:
 - https://testbed011.cern.ch:7846/137.138.181.253/185337169921026?testbed011.cern.ch:7771
 have been successfully retrieved and stored in the directory:
 /sandbox/185337169921026
```

A Job Submission Example



Major achievements to date

- Core middleware group developed innovative S/W now considered by our US colleagues (GDMP and resource broker)
- Excellent collaboration with US Globus and Condor developments
- Close collaboration with similar US projects (PPDG, GriPhyN/iVDGL)
- Large community of enthusiastic, dedicated scientists
- End users involved in all stages of the project (requirements definition, architecture, S/W integration, deployment, validation and tests)
- Unfunded staff effort about twice the EU funded (voluntary participation from Portugal, Ireland, Russia & Denmark both in M/W and in the test bed) as a good measure of success for the project

more achievements, continued…

- Good relations to industry (through the I&R Forum)
- Seed funds for national Grid projects, coordinator and initiator of other projects (DataTAG, CrossGrid, GridSTART)
- Initiator and active participant in GGF, Intergrid, EIROForum Grid WG, OCDE interest to start a WG, exploratory work in Asian Pacific and South America
- Pioneering role (EU Grid flagship project): first opportunity to work on Grid for ESA with fostering effect of internal Grid activity
- Prototype use of national RNs for Grid deployment (building Grids of Grids)

Future Plans

- Expand and consolidate testbed operations
  - need to improve the distribution, maintenance and support process, also understand what operating a Grid means
- Evolve architecture and software on the basis of TestBed usage and feedback from users (consider converging to a common document with PPDG/GriPhyN, understand possible impact of OGSA)
- Prepare for second test bed in autumn 2002
- Enhance synergy with US via DataTAG-iVDGL and InterGrid
- Build a complete and solid collaboration plan with CrossGrid and other relevant projects (also using GridSTART)
- Promote early standards adoption with participation to GGF and other international bodies

Closing Remarks

- The project after just one year is up and running with 21 partners all contributing according to the contract
- First testbed deployed on 5 main sites (in France, Italy, NL, UK and CERN)
- Real applications from Biology and Medicine, Earth Observation and Particle Physics demonstrated on the test bed
- First review passed to the full satisfaction of the EU reviewers
  - Excellence of technical results and sound management commended
- EU Grid flagship role confirmed with increased visibility in the international bodies (GGF and others)
- Aggressive programme ahead to evolve towards more production quality testbeds for next two years, understand Globus migration to OGSA and prepare for the next EU FP6

Conclusions

- For the future extension and consolidation of the project test bed we rely on the progress of European Research Networks
- We require high QoS and be confident to be able to consume all bandwidth you will be able to provide us with!!!

Learn more about Grids & DataGrid

Programme includes
Grid Lectures by
Ian Foster
Carl Kesselman
Hands-on tutorial
DataGrid

CERN School of Computing 2002

Apply now via web
http://csc.web.cern.ch/CSC/
Places are limited

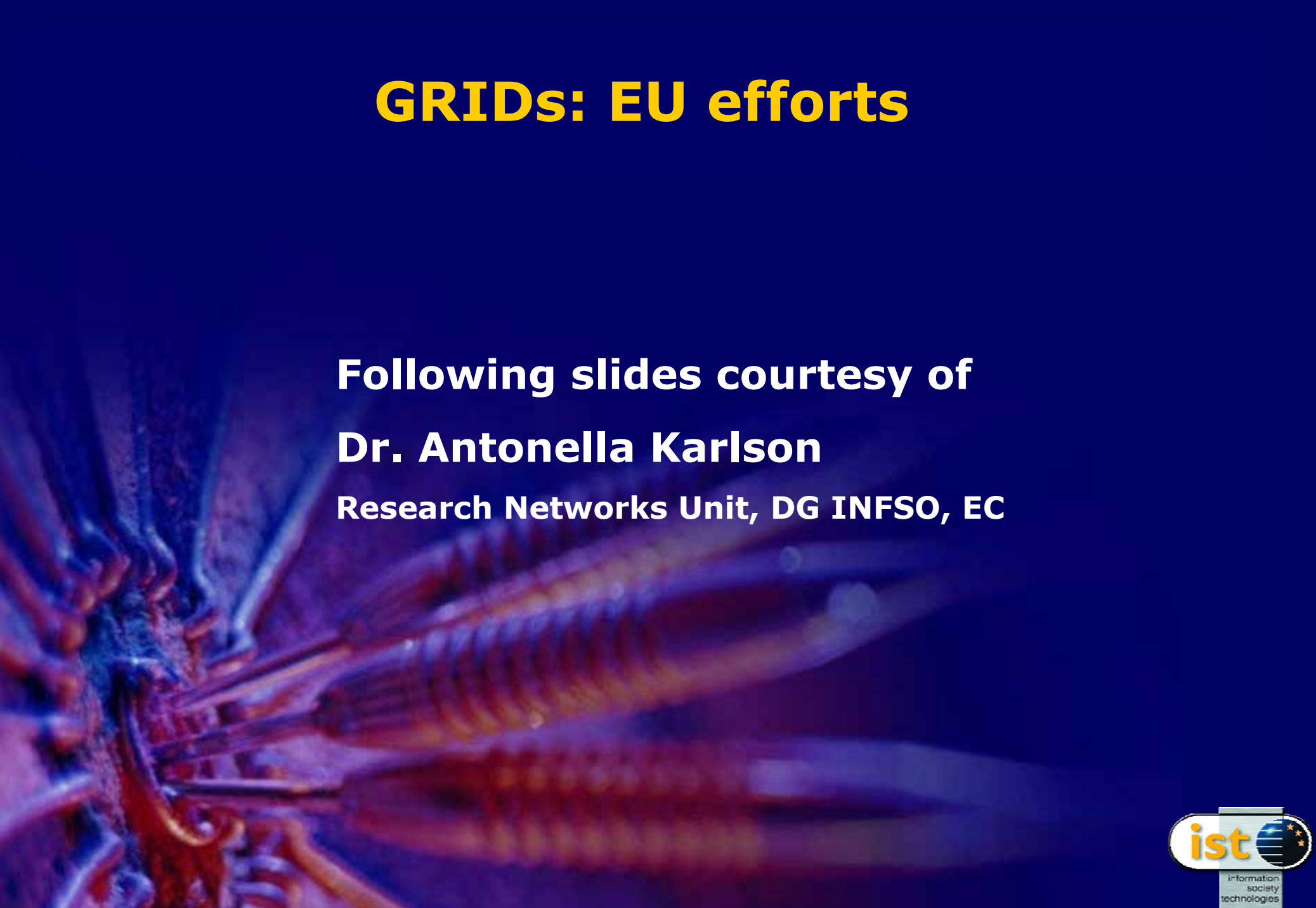
Vico Equense, Italy, 15-28 September 2002
The 2002 CERN School of Computing is organised by CERN, with the Institute of Composite and Biomedical Materials, National Research Council, Naples, Italy.

GRIDs: EU efforts

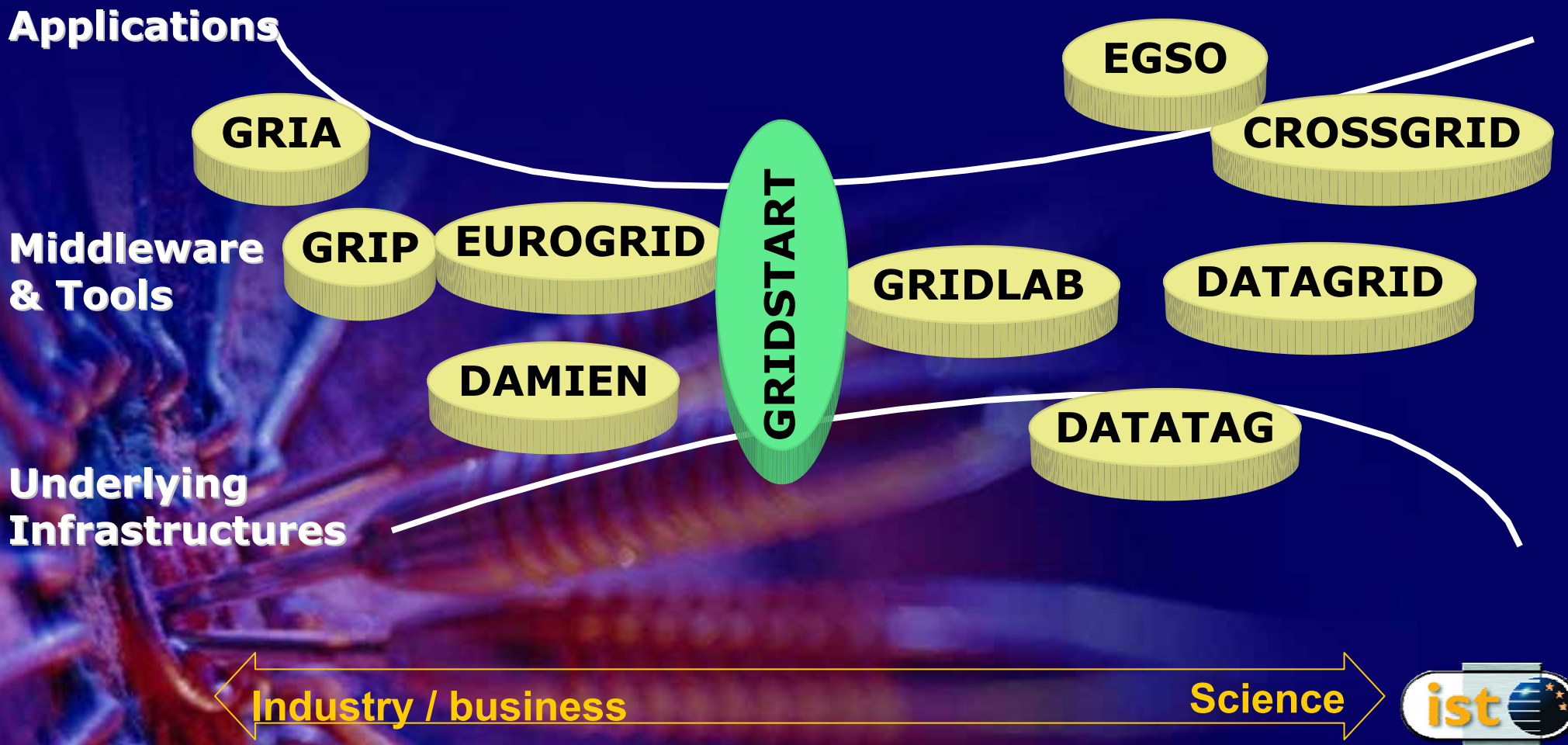
Following slides courtesy of

Dr. Antonella Karlson

Research Networks Unit, DG INFSO, EC



GRIDs - IST projects (~36m Euro)
An integrated approach



FP6
Research networks related work

Integrating European Research

- **Genomics** **2200 Meuro**
- **IST** **3600 Meuro**
- **Nanotechnologies, intellig. materials, new processes** **1300 Meuro**
- **Aeronautics and space** **1075 Meuro**
- **Food safety /health risks** **685 Meuro**
- **Sustainable development** **… 2120 Meuro**
- **Citizens in knowledge soc.** **225 Meuro**
- **Specific activities covering a wider field of research** **1320 Meuro**

FP6
Research networks related work

Structuring the ERA

Research and Innovation	300 Meuro
Human resources	1630 Meuro
Research Infrastructures	665 Meuro
Science/Society	60 MEuro

Strengthening the Foundations of ERA

Support to co-ordination	280 Meuro
Support policy developm	50 Meuro