



# Introduction to GRID computing and overview of the European Data Grid



The European DataGrid Project

<http://www.edg.org>

---



# Overview

- Introduction
  - What is GRID computing ?
  - What is a GRID ?
  - Why GRIDs ?
- GRID projects world-wide
- The European Data Grid (EDG)
  - Overview of EDG goals and organization
  - Overview of the EDG middleware components



## What is GRID computing :

- **coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations.** [ I.Foster]
- **A VO is a collection of users sharing similar needs and requirements in their access to processing, data and distributed resources and pursuing similar goals.**
- **Key concept :**
  - **ability to negotiate resource-sharing arrangements among a set of participating parties (providers and consumers) and then to use the resulting resource pool for some purpose** [I.Foster]



## What is a GRID:

- collaborative set of computing, data storage and network resources **belonging to different administrative domains** that has knowledge about the status of its components through active, distributed information services
- allows certified users belonging to multi-domain **Virtual Organizations** to access a large amount of resources via single log in. (sign on)
- Manage concurrent access by large numbers of dispersed users
- Provide a service that can cope with unavailability of distributed resources
- Has no single point of failure



# A Checklist for a GRID to be a GRID

(I. Foster)

- a GRID coordinates resources that are not subject to centralized control and live within **different control domains** and addresses the issues of security, policy, payment, membership... that arise in these settings. (i.e. it is not a local management system.)
- uses **standard, open, general-purpose** protocols and interfaces (i.e. it is not an application-specific system).
- a GRID allows its constituent resources to be used in a coordinated fashion to deliver various **qualities of service** ( response time, throughput, availability, and security, and/or co-allocation of multiple resource types ).
- the utility of the combined system is significantly greater than that of the sum of its parts to meet complex user demands.



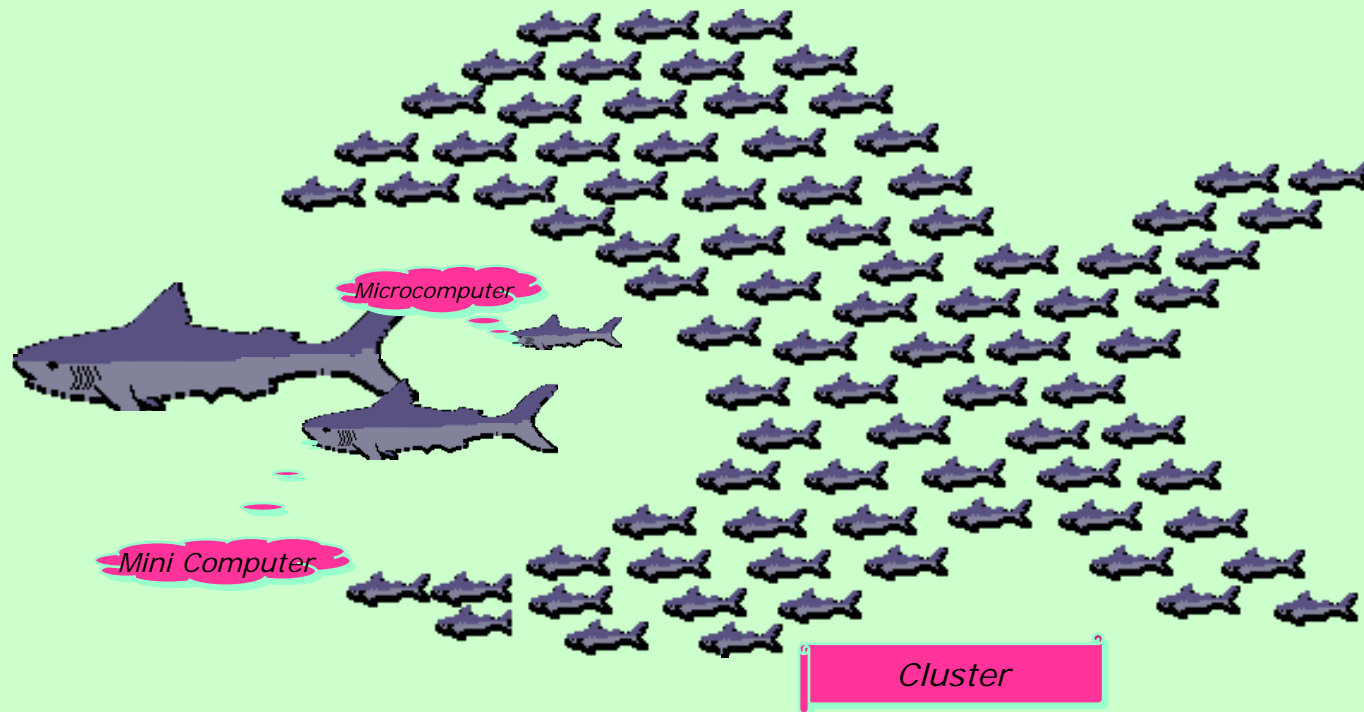
## Elements of the GRID problem

- Flexible, secure, coordinated resource sharing
  - Trust
  - Policy
  - Negotiation
  - Payment
- User communities able to share geographically distributed resources
- Absence of a central location, a centralized control
- Optimize the global efficiency in the usage of resources
  - status is not under our direct control
  - current status is uncertain to some degree



# The GRID distributed computing idea 1/2

Once upon a time.....

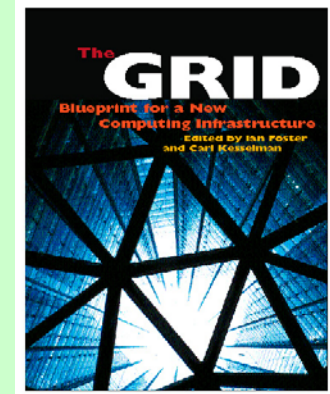
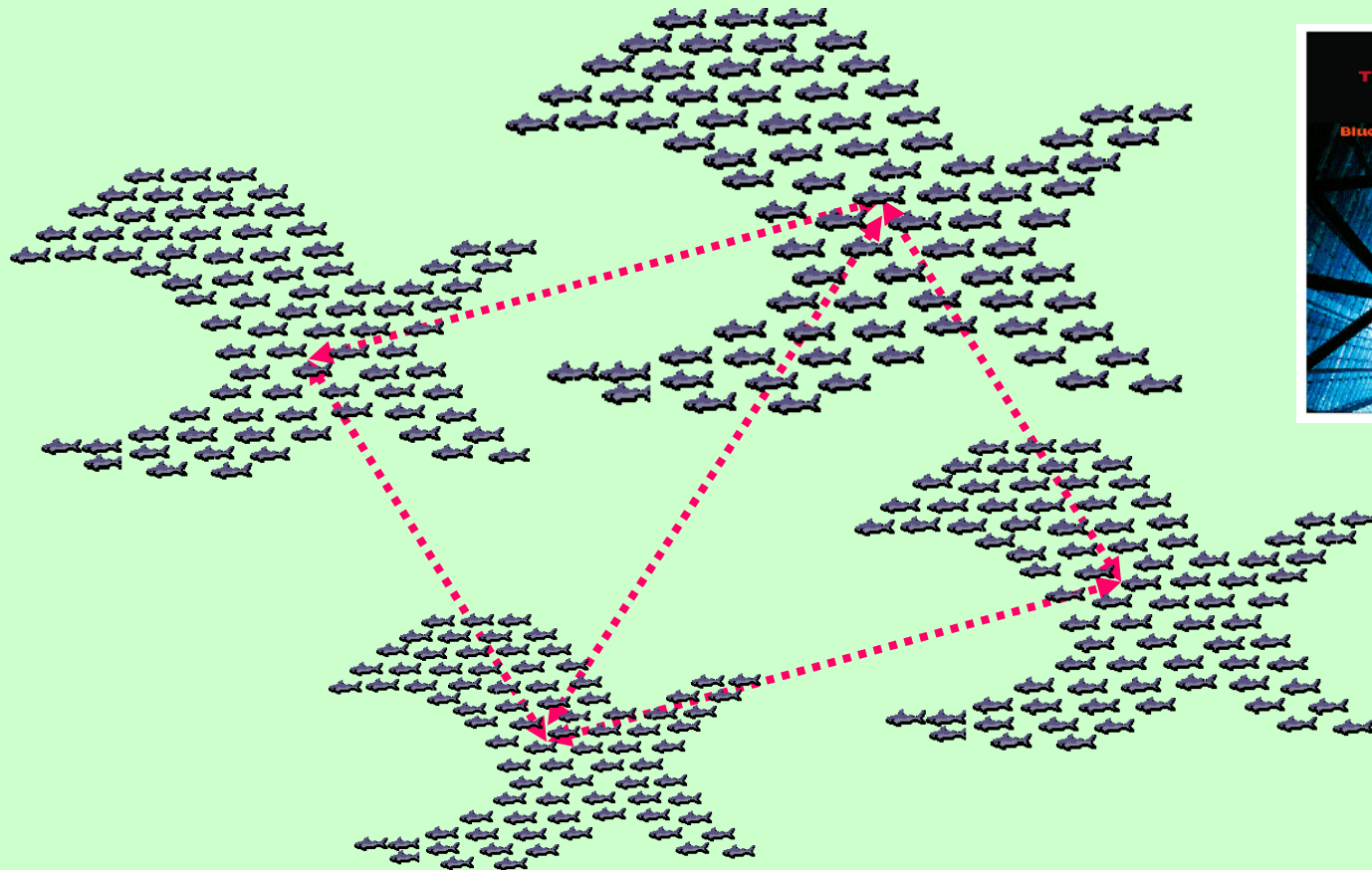


(by Christophe Jacquet)



# The GRID distributed computing idea 2/2

...and today



(by Christophe Jacquet)

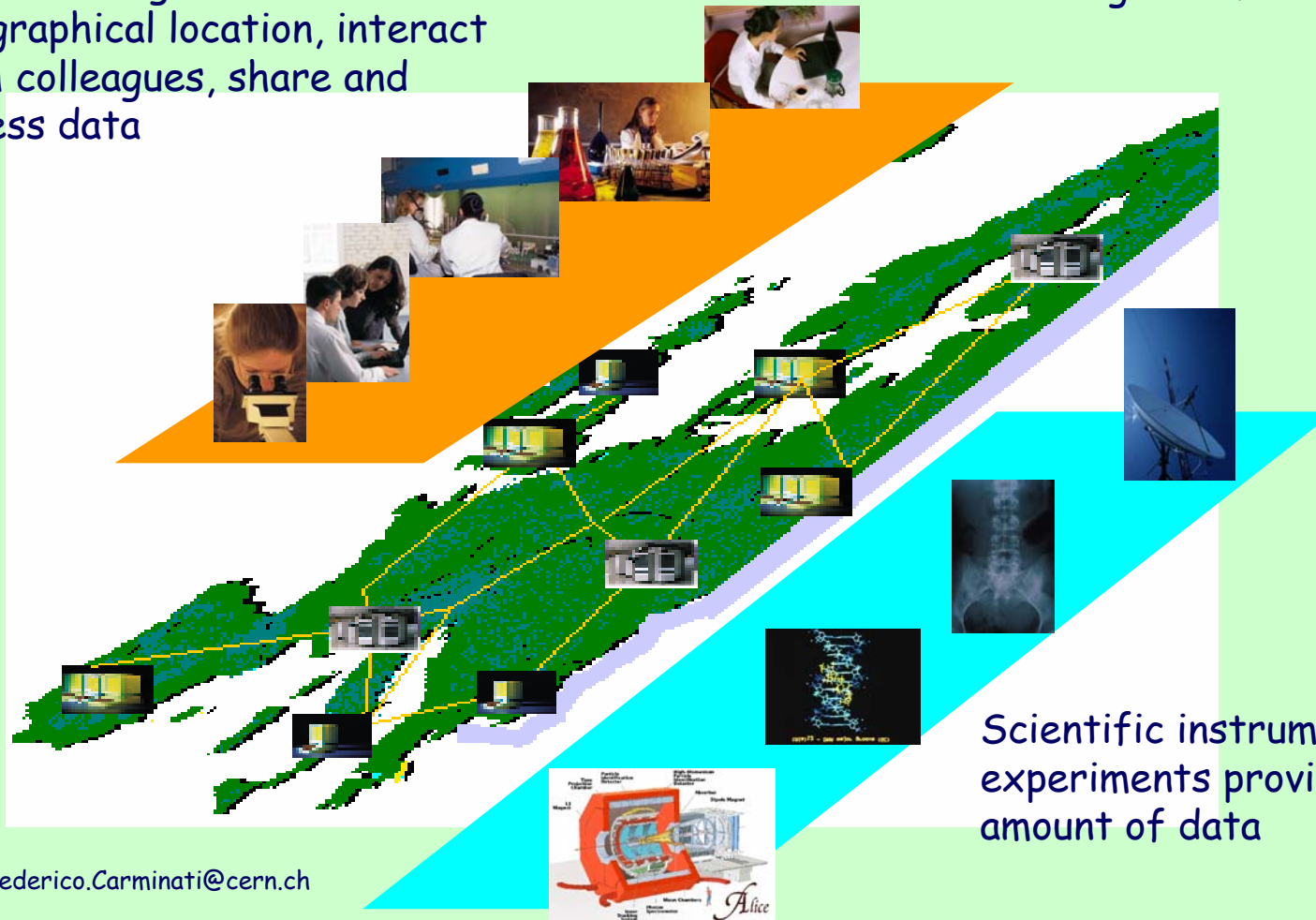




# The Grid Vision (1/2)

Researchers perform their activities regardless geographical location, interact with colleagues, share and access data

The GRID: networked data processing centres and "middleware" software as the "glue" of resources.



Scientific instruments and experiments provide huge amount of data

Federico.Carminati@cern.ch



## The Grid Vision (2/2) (I.Foster,G.Gilder)

- On-demand, ubiquitous access to computing, data, and services.
- New capabilities constructed dynamically and transparently from distributed services
- "When the network is as fast as the computer's internal links, the machine disintegrates across the net into a set of special purpose appliances"  
(George Gilder)

# Why GRIDs



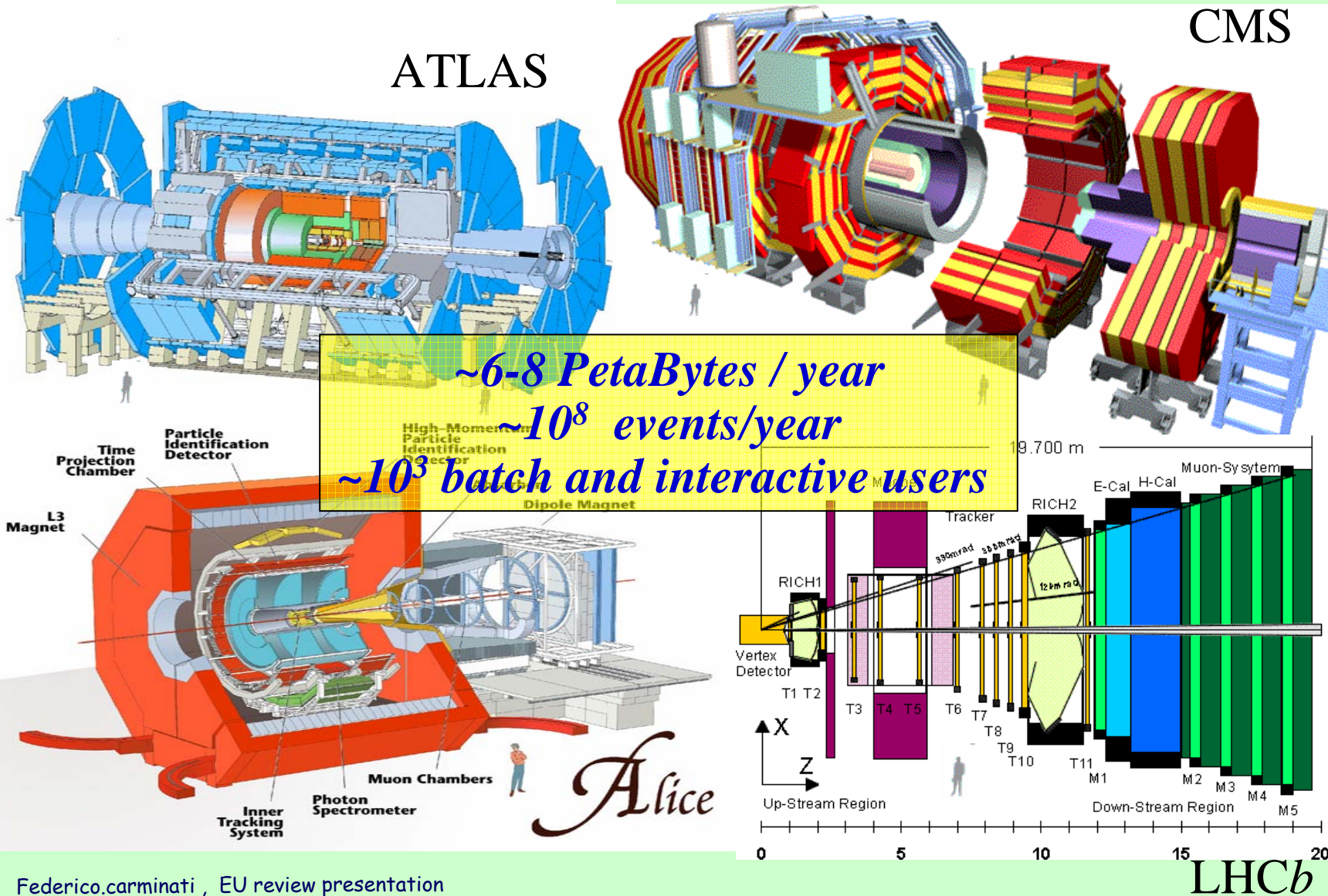
The scale of the problems *human beings* have to face to perform frontier research in many different fields is **constantly increasing**.

- Performing frontier research in these fields already today requires world-wide collaborations (i.e. multi domain access to distributed resources).
- GRIDs naturally address this need for collecting and sharing resources (CPUs, Data Storage, Data ), providing - thanks to always growing throughputs and QoS in the underlying networks - unprecedented possibilities to access **large data processing power** and **huge data storage** and **data access possibilities**.
- Large Community of possible GRID users :  
High Energy Physics, planet Earth's health studies (*Geology, Environmental studies, Earthquakes forecast, geologic and climate changes, ozone monitoring*), Biology, Genetics, Earth Observation, Astrophysics, New composite materials research, Astronautics



# High Energy Physics

## The LHC Detectors



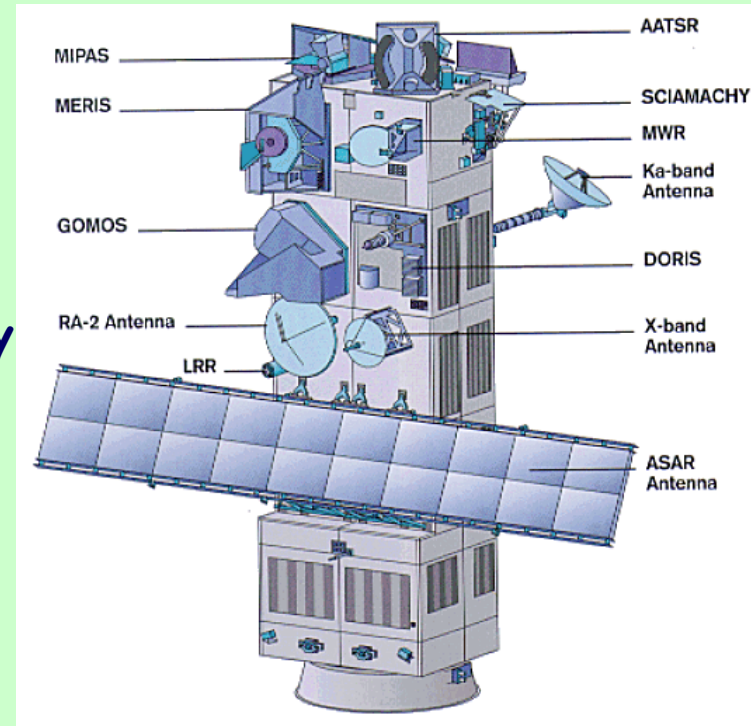
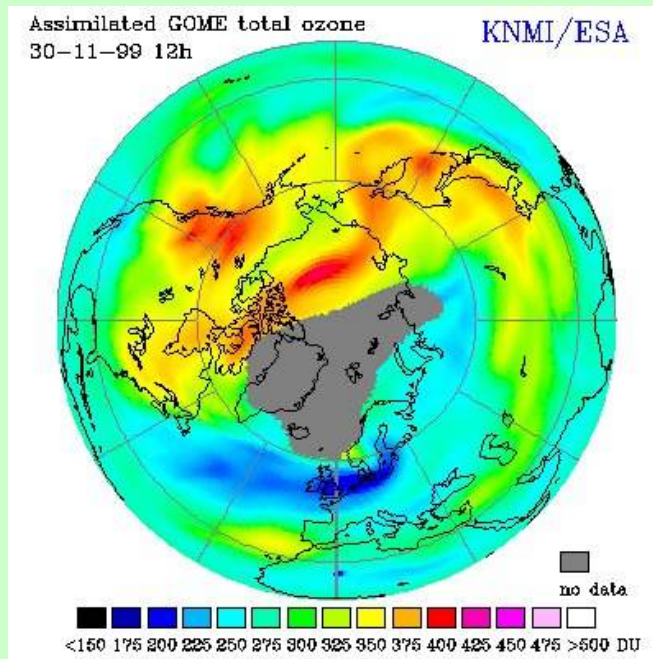
Federico.carminati , EU review presentation



# 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

The screenshot displays the Visual DataGrid BLAST application. The main window shows a sequence alignment for 'NR\_SC:SW-PABP\_YEAST' with 32 homologies found and a maximum score of 2778. A bar chart visualizes the alignment scores. A 'Job Launch' dialog box is open in the foreground, allowing users to configure the BLAST job. The dialog includes fields for 'Sequence file', 'Output file', and 'Logical filename', each with a 'Browse...' button. It also features a 'Grid save' checkbox, a 'Database' dropdown set to 'YEAST', an 'Algorithm' dropdown set to 'BlastP+MSPcrunch', and a 'Number of job(s)' field set to 5. The 'Start' button is highlighted.

**Visual DataGrid BLAST Job Launch Configuration:**

- Sequence file:  Browse...
- Output file:  Browse...
- Logical filename:   Grid save
- Database: YEAST Algorithm: BlastP+MSPcrunch
- Number of job(s): 5 Default number Clear all
- Start Cancel

**Sequence Alignment (NR\_SC:SW-PABP\_YEAST):**

```

1  A D I T D K T A E Q L E N L N I Q D D Q K Q A A T G :
30 Q S V E N S S A S L Y V G D L E P S V S E A H L Y D :
59 P I G S V S S I R V C R D A I T K T S L G Y A Y V N :
88 H E A C R K A I E Q L N Y T P I K C R L C R I M W S :
117 P S L R K K G S G N I P I K N L H P D I D N K A L Y :
146 S V F G D I L S S K I A T D E N G K S K G F G F V H :
175 E G A A K E A I D A L N G M L L N G Q E I Y V A P H :
204 K E R D S Q L E E T K A H Y T N L Y V K N I N S E T :
233 Q F Q E L F A K P G P I V S A S L E K D A D G K L K :
262 F V N Y E K H E D A V K A V S A L N D S E L N G E K :
291 C R A Q K K N E R M H V L K K Q Y E A Y R L E K M A :
320 G V N L F V K N L D D S :
349 S A K V M R T E N G K S :
378 I T E K N Q Q I V A G K :
407 A Q Q I Q A R N Q M R Y :
436 F H P P M F Y G V M P P :
465 C H P K N C M P P Q P R :
494 N D N N Q P Y Q Q K Q R :
523 E E A A G K I T C M I L :
552 E Q H Y K E A S A A Y E
  
```

**Search Results List:**

a-z	Z-a	Score
NR_SC:GP-CAA60917_1		
NR_SC:PIR-B23496		
NR_SC:GP-CAA82351_1		
NR_SC:GP-CAA81266_1		
NR_SC:GP-CAA99202_1		
NR_SC:GP-AAA79056_1		
NR_SC:GP-CAA86921_1		
NR_SC:GP-CAA90396_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-GFA1_YEAST		
NR_SC:SW-UGS1_YEAST		
P-AB67523_1		
P-CAA97711_1		
W-ASN1_YEAST		
W-HS83_YEAST		
W-ASN2_YEAST		
P-CAA80726_1		
W-PABP_YEAST		
P-CAA84004_1		
W-GLUA_YEAST		
W-HS75_YEAST		
W-HS76_YEAST		
P-AB23074_1		
P-CAA73947_1		
P-CAA67472_1		
P-AAA99685_1		
P-CAA96120_1		
P-CAA82046_1		
P-AB60298_1		
P-CAA96762_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		



# GRID computing and High Throughput Computing

- High Throughput Computing is the effective management and exploitation of all available computing resources.
- limited predictability of the actual availability of distributed, remote, multi-domain resources requires a way to cope with it.
- main challenge for HTC:  
**maximizing the amount of resources accessible** to its customers. Distributed ownership of computing resources is the major obstacle such an environment has to overcome in order to expand the pool of resources it can draw from.



## GRID Security

- user's identity has to be certified by (mutually recognized) national Certification Authorities (accessing resources belonging to different domains requires identity to be certified).
- secure access to resources is required (security framework to allow resources access only to certified, identified users (X.509 Public Key Infrastructure)).
- resources (node machines) have to be certified by CAs
- temporary delegation from users to processes to be executed "in user's name" ( proxy certificates ).
- Common agreed policies for accessing resource and handling user's rights across different domains in within the same Virtual Organization a user belongs to.





## GRID projects world wide

### ➤ EU

- EDG (EU-IST) - R&D EU GRID project [ [www.edg.org](http://www.edg.org) ]
- CrossGRID - QoS - Real Time apps. [ [www.crossgrid.org](http://www.crossgrid.org) ]
- DataTAG - GLUE (EU-USA) [ [www.datatag.org](http://www.datatag.org) ]
- LCG - The LHC Computing GRID - Deployment [ [cern.ch/lcg](http://cern.ch/lcg) ]
- The new 16,2 B Euro EU VI Framework Prog. GEANT based GRID projects

### ➤ USA

- GriPhyN [ [www.griphyn.org](http://www.griphyn.org) ]
- iVDGL-VDTv1 [ [www.idvgl.org](http://www.idvgl.org) ]
- PPDG (NSF, DoE) [ [www.ppdg.org](http://www.ppdg.org) ]

### ➤ Asia

- ApGrid [ [www.apgrid.org](http://www.apgrid.org) ]
- Pragma (USA-Asia) [ <http://pragma.ucsd.edu> ]



# The European Data Grid

- EDG is a project funded by the European Union to exploit and build the next generation computing infrastructure providing intensive computation and analysis of shared large-scale databases.
- Enable data intensive sciences by providing world wide Grid test beds to large distributed scientific organisations.
- |                                  |                    |
|----------------------------------|--------------------|
| Start ( Kick off ) : Jan 1, 2001 | End : Dec 31, 2003 |
|----------------------------------|--------------------|
- Applications/End User Communities : HEP, Earth Observation, Biology.
- Specific Project Objectives:
  - Middleware for Jobs (Workload) and Data Management, Information Systems, Fabric & GRID management, Network Monitoring
  - Large scale testbed
  - Production quality demonstrations
  - Contribute to Open Standards and international bodies  
( GGF, Industry & Research forum)



# The EDG Main Partners

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





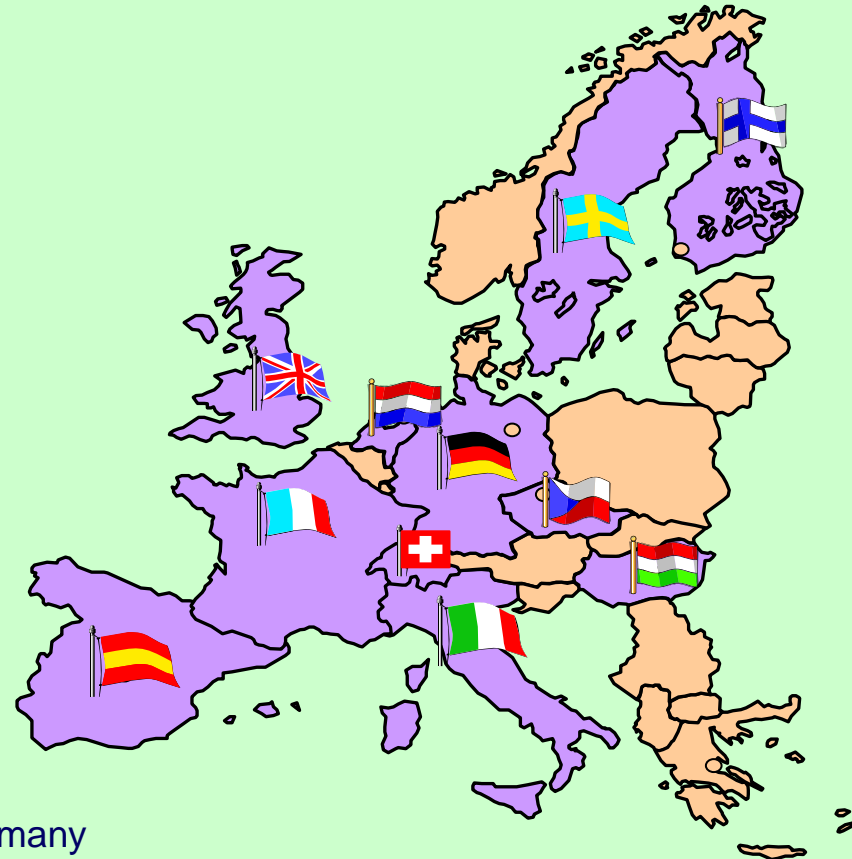
# EDG Assistant Partners

## 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





## EDG overview: Middleware release schedule

- Release schedule
  - **Release 1.4:** December 2002
  - **Release 2.0:** May 2003
- Each release includes
  - feedback from use of previous release by application groups
  - planned improvements/extension by middle-ware groups
- High Energy Physics experiments and **Data Challenges:**
  - ATLAS production data challenge demonstration on EDG currently On-going ( main EDG production demo effort - mid September )
  - CMS, LHCb, ALICE, Earth Obs. & Bio-Info. will follow ATLAS in demonstrating productions



## EDG overview : current project status

- EDG currently provides set of middleware services
  - Job & Data Management
  - GRID & Network monitoring
  - Security, Authentication & Authorization tools
  - Fabric Management
- Runs on Linux Red Hat 6.2 platform
  - Site install & config tools and set of common services available
    - ( Resource Brokers, VO-LDAP servers for Authentication, VO-based Replica Catalogs, VO-management services )
- 5 principle EDG 1.2.0 sites currently belonging to the EDG-Testbed
  - CERN(CH), RAL(UK), NIKHEF(NL), CNAF(I), CC-Lyon(F),
    - being deployed on other EDG testbed sites (~10)
- Intense middleware development continuously going on, concerning:
  - New features for job partitioning and check-pointing, billing and accounting
  - New tools for Data Management and Information Systems.
  - Integration of network monitoring information inside the brokering polices



## EDG structure : work packages

➤ The EDG collaboration is structured in 12 Work Packages:

- WP1: Work Load Management System
- WP2: Data Management
- WP3: Grid Monitoring / Grid Information Systems
- WP4: Fabric Management
- WP5: Storage Element
- *WP6: Testbed and demonstrators*
- WP7: Network Monitoring
- **WP8: High Energy Physics Applications**
- **WP9: Earth Observation**
- **WP10: Biology**
- **WP11: Dissemination**
- **WP12: Management**



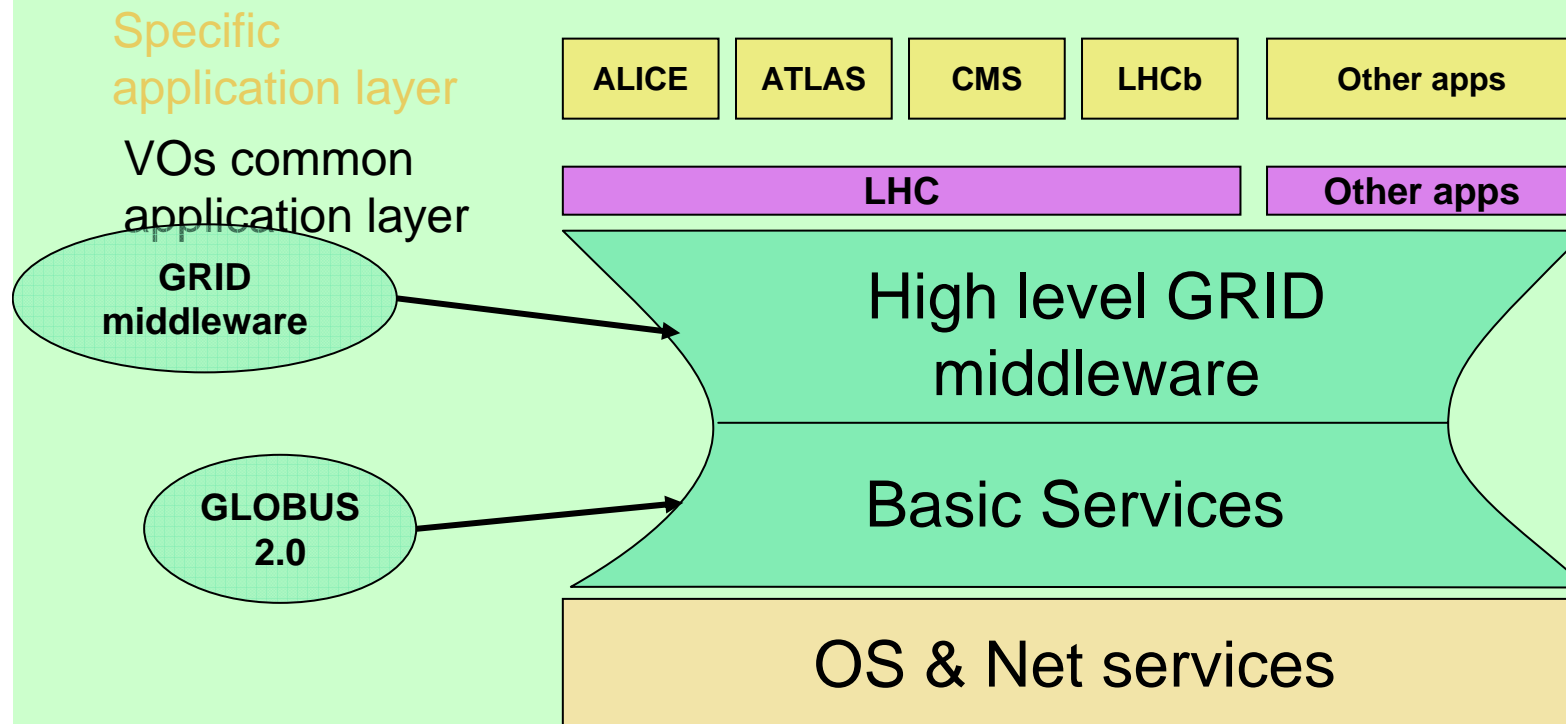
**Applications**



# EDG Globus-based middleware architecture

## ➤ Current EDG architectural functional blocks:

- Basic Services ( authentication, authorization, Replica Catalog , secure file transfer,Info Providers) rely on Globus 2.0 (G 2.0 release 21)  
(GSI, GRIS/GIIS,GRAM, MDS)
- Higher level EDG middleware
- Top level applications (HEP,BIO,EO)







# EDG middleware GRID architecture

APPLICATIONS

Local Computing

Local Application

Local Database

Grid

Grid Application Layer

Job Management

Data Management

Metadata Management

Collective Services

Grid Scheduler

Replica Manager

Information & Monitoring

Underlying Grid Services

SQL Database Services

Computing Element Services

Storage Element Services

Replica Catalog

Authorization Authentication and Accounting

Service Index

Grid

M / W

Fabric

Fabric services

Resource Management

Configuration Management

Monitoring and Fault Tolerance

Node Installation & Management

Fabric Storage Management

GLOBUS



## EDG : reference web sites

- EDG web site
  - <http://www.edg.org>
- Source for all required software :
  - <http://datagrid.in2p3.fr>
- EDG testbed web site
  - <http://marianne.in2p3.fr>
- EDG Users' Guide and other documentation
  - <http://marianne.in2p3.fr/datagrid/documentation/>
- EDG tutorials web site (username: griduser passwd: tutorials123)
  - <http://cern.ch/edg-tutorials>
- EDG production testbed current real time updated set up
  - <http://testbed007.cern.ch/>