An Introduction to Grid Computing and the EGEE Project

Mike Mineter
Training Outreach and Education
National e-Science Centre, UK

mjm@nesc.ac.uk

With thanks to EGEE colleagues for many of these slides

Contents

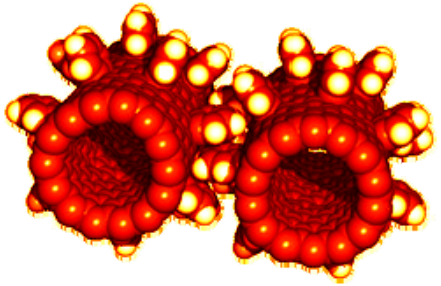
- **Introduction to**
  - e-Research and e-Science
  - Grid Computing
  - e-Infrastructure
- **Some examples**
- **Grid concepts**
- **Grids - Where are we now?**
- **More about the EGEE project**

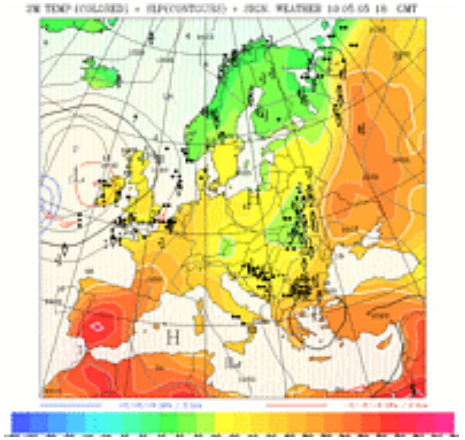
Computing intensive science

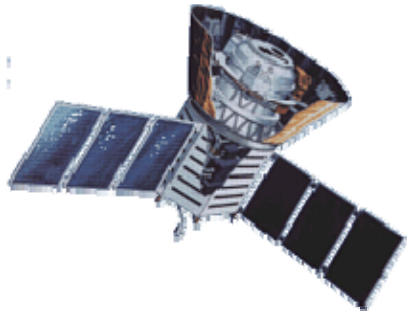
- **Many vital challenges require community effort**
  - Fundamental properties of matter
  - Genomics
  - Climate change
  - Medical diagnostics
- **Research is increasingly digital, with increasing amounts of data**
- **Computation ever more demanding**

e.g.: experimental science uses ever more sophisticated sensors

  - Huge amounts of data
  - Serves user communities around the world
  - International collaborations







"e-Science" and "e-Research"

- **Collaborative research that is made possible by the sharing across the Internet of resources (data, instruments, computation, people's expertise...)**
  - Crosses organisational boundaries
  - Often very compute intensive
  - Often very data intensive
  - Sometimes large-scale collaboration

- **Early examples were in science: "e-science"**

- **Relevance of "e-science technologies" to new user communities (social science, arts, humanities…) led to the term "e-research"**

e-Science: the invitation

Collaborative
"virtual computing"

Sharing data, computers, software
Enabled by Grids – two main types
- specific to a project
- supporting many collaborations

Improvised cooperation

Email
File exchange
ssh access to run programs
Enabled by networks:
national, regional and International: GEANT

People with shared goals

e-Infrastructure

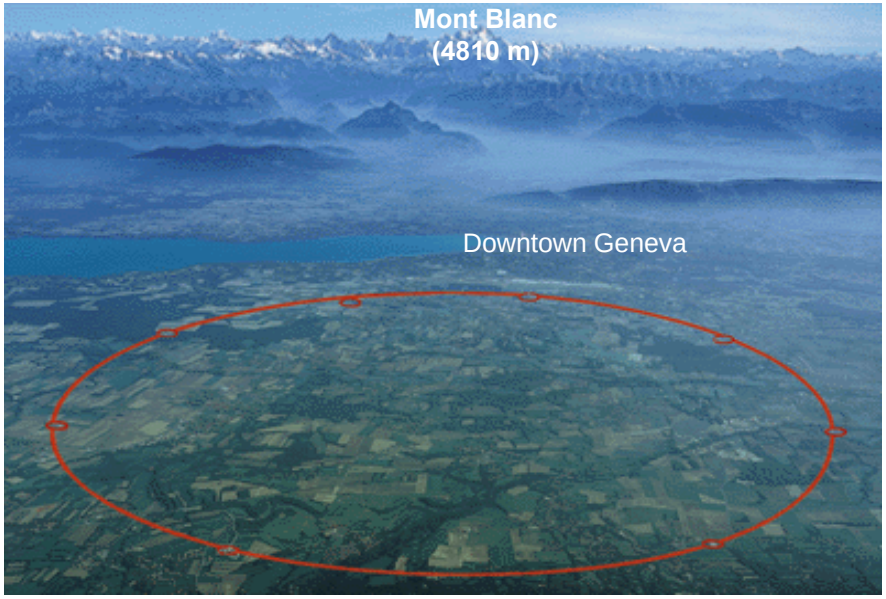
- ***Networks + Grids***
  - *Networks connect resources*
  - *Grids enable "virtual computing" - resource sharing across administrative domains*
    - *"admin. domain": institute, country where resource is; system management processes;…*
- ***+ Operations, Support, Training…***
- ***+ Data centres, archives,…***

Some examples of e-science

Particle Physics

- **Large amount of data**
- **Large worldwide organized collaborations**
- **Computing and data management resources distributed world-wide owned and managed by many different entities**

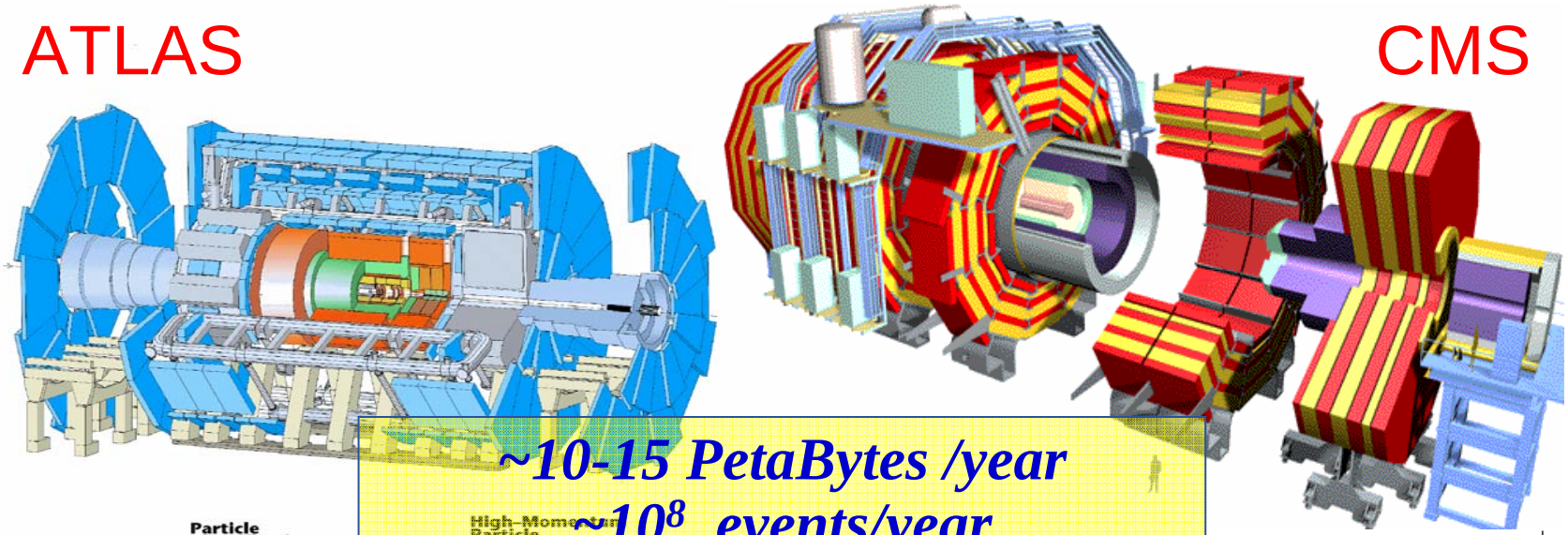
- **Large Hadron Collider (LHC) at CERN in Geneva Switzerland:**
  - One of the most powerful instruments ever built to investigate matter



The LHC Experiments

ATLAS

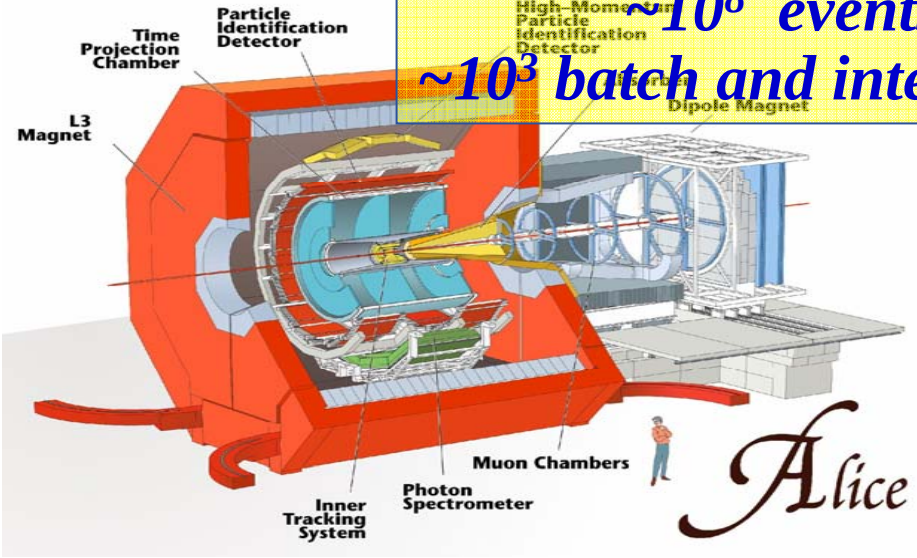
CMS

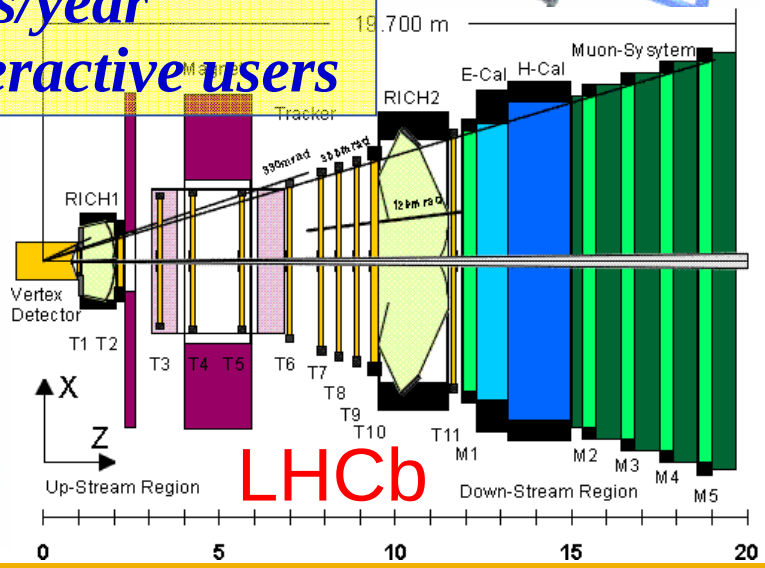


~10-15 PetaBytes /year

~10^8 events/year

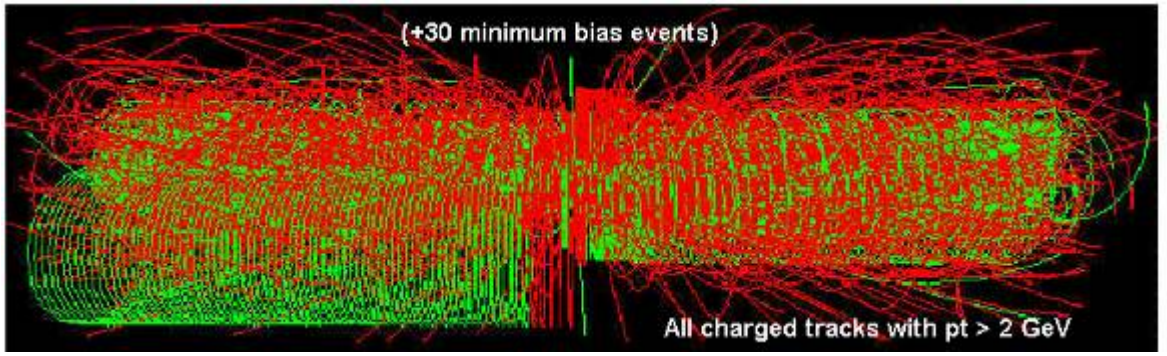
~10^3 batch and interactive users



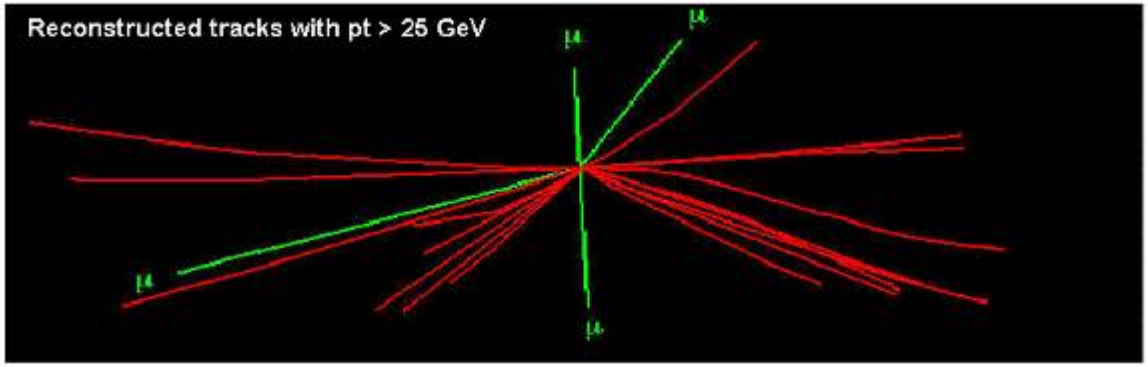


The LHC Data Challenge

Starting from this event



Looking for this "signature"

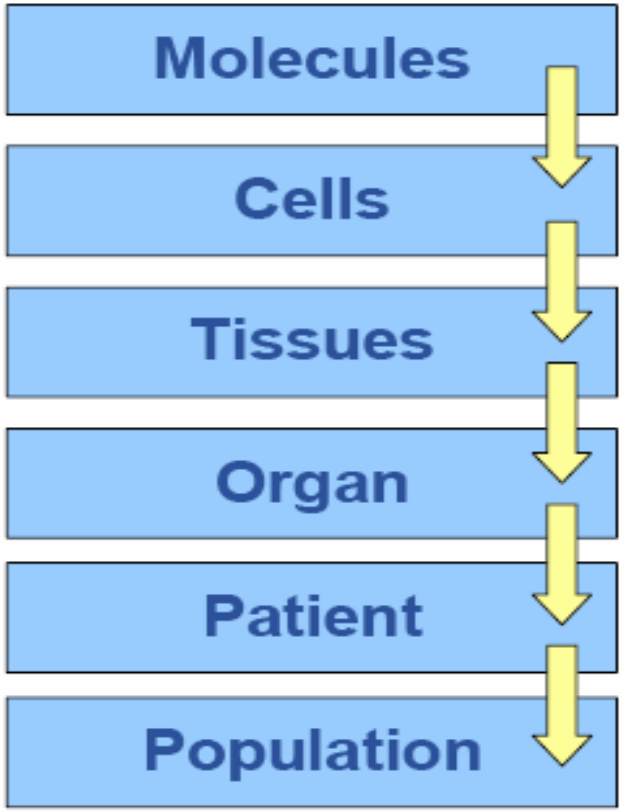


→ **Selectivity: 1 in 10^{13}**

(Like looking for a needle in 20 million haystacks)

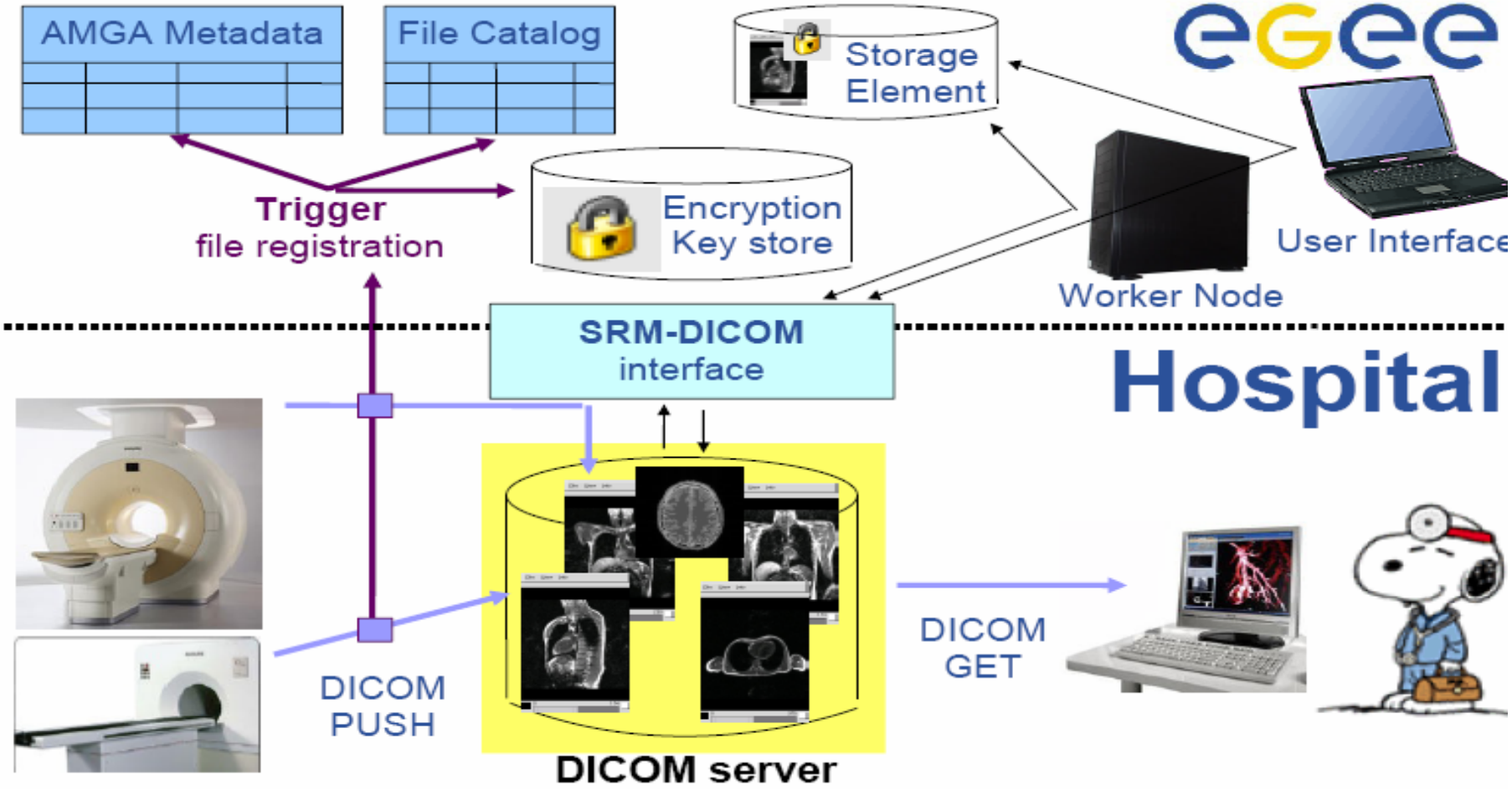
Biomedical applications

- **Bioinformatics**
  - Genomics
  - Proteomics
  - Phylogeny…
- **Medical imaging**
  - Medical imaging
  - Computer Aided Diagnosis
  - Therapy planning
  - Simulation…
- **Life sciences**
  - Drug discovery
  - Epidemiology
  - …



Biomedical community and the Grid, EGEE User Forum, March 1st 2006, I. Magnin

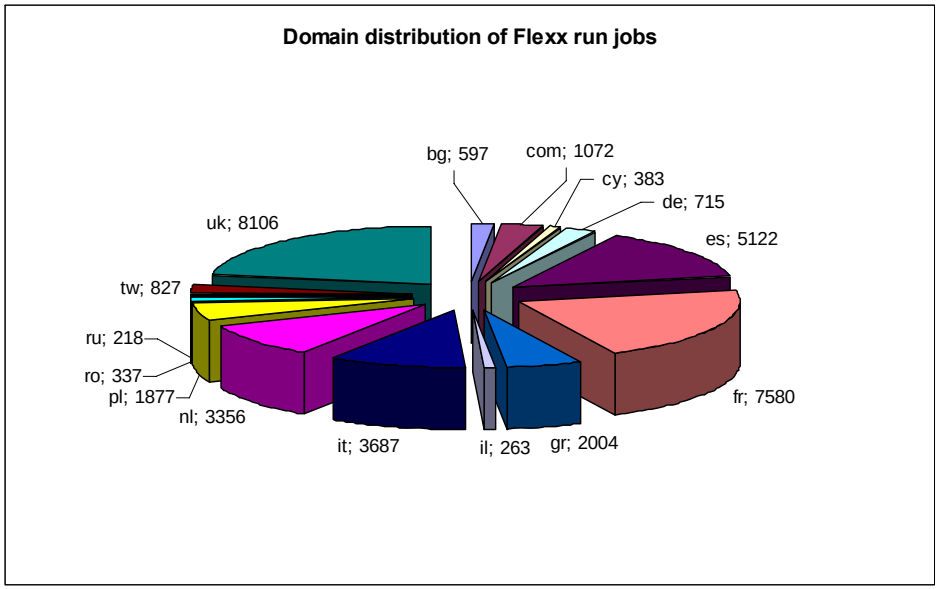
Data management – medical images



Biomedical community and the Grid, EGEE User Forum, March 1st 2006, I. Magnin

First biomedical data challenge: World-wide In Silico Docking On Malaria (WISDOM)

- **Significant biological parameters**
  - two different molecular docking applications (Autodock and FlexX)
  - about one million virtual ligands selected
  - target proteins from the parasite responsible for malaria
- **Significant numbers**
  - Total of about 46 million ligands docked in 6 weeks
  - 1TB of data produced
  - Up to 1000 computers in 15 countries used simultaneously for a total of about 80 CPU years
- **Significant results**
  - Best hits to be re-ranked using Molecular Dynamics

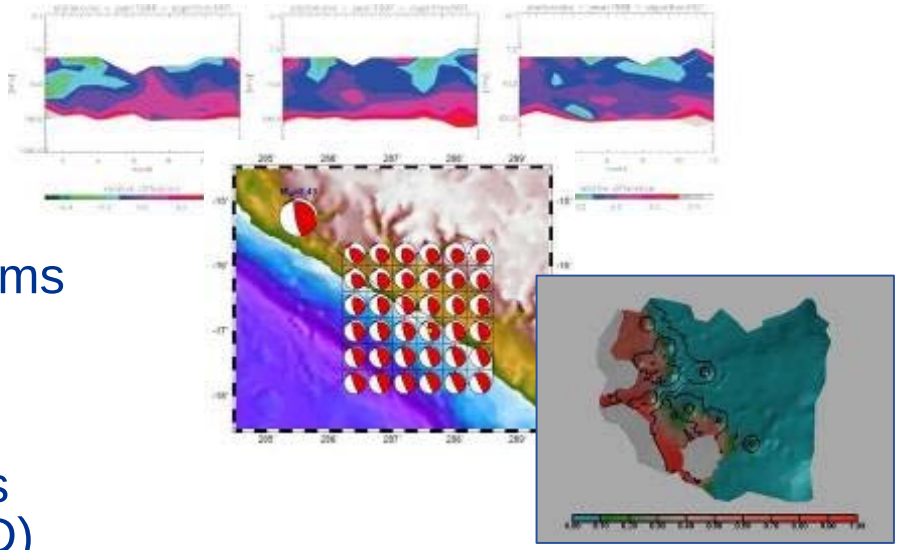


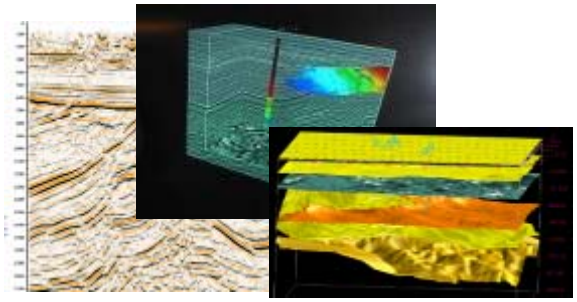
New data challenge in the fall of 2006
New malaria targets
Focus on other neglected diseases
Enlarged collaboration
(possibly including related projects)

Roberto Barbera, 1st EGEE User Forum, CERN, 1st March 2006

Earth sciences applications

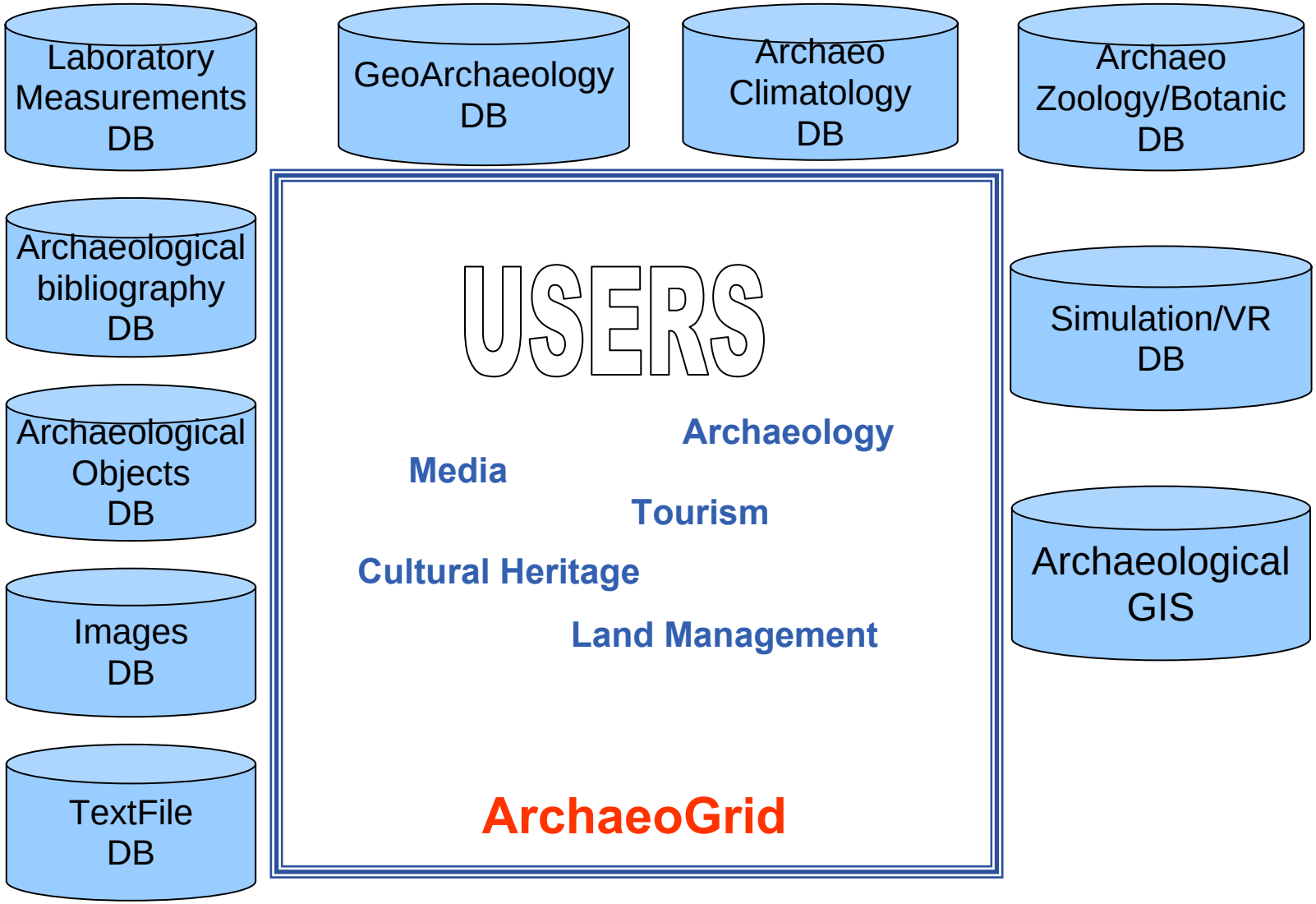
- **Earth Observations by Satellite**
  - Ozone profiles
- **Solid Earth Physics**
  - Fast Determination of mechanisms of important earthquakes
- **Hydrology**
  - Management of water resources in Mediterranean area (SWIMED)
- **Geology**
  - Geocluster: R&D initiative of the Compagnie Générale de Géophysique





➢ **A large variety of applications ported on EGEE**

The newest EGEE application: Archaeology

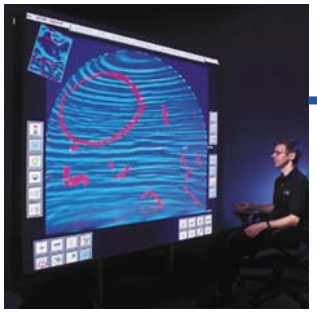


Grid concepts

The Grid Metaphor

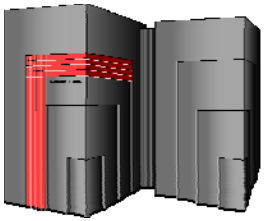
Mobile Access

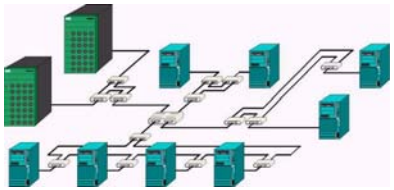
Workstation



Visualising

GRID MIDDLEWARE





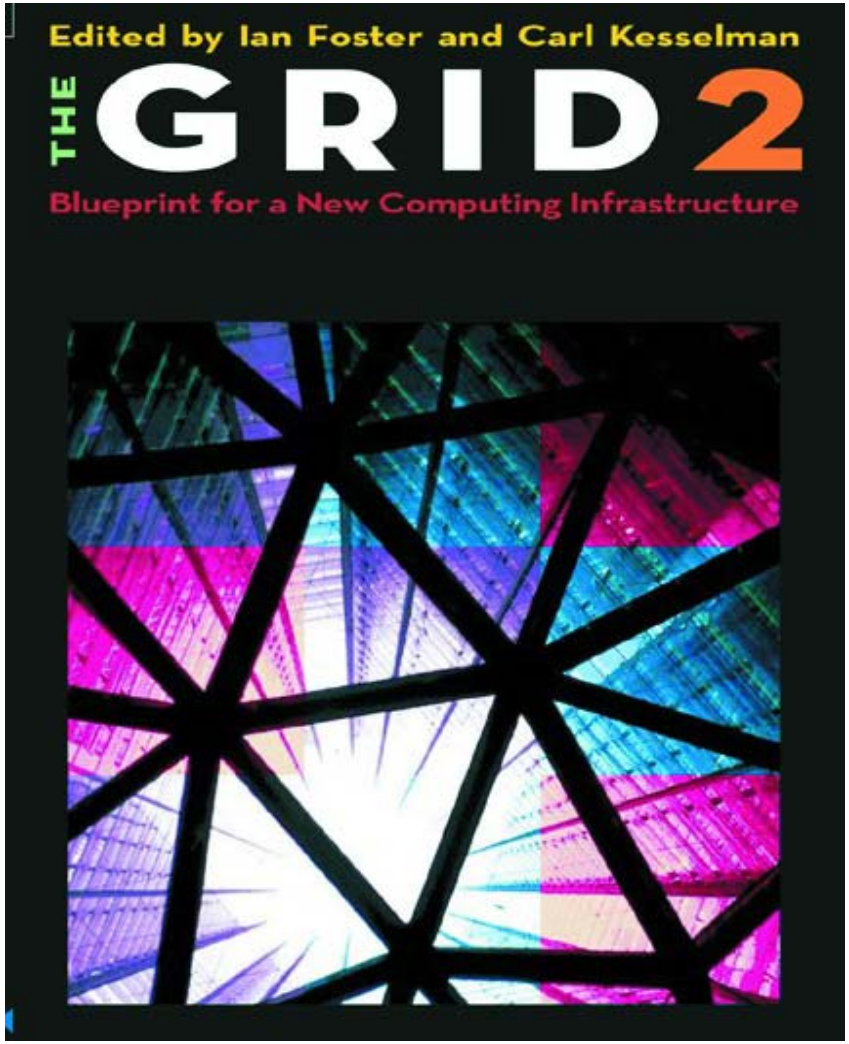
Supercomputer, PC-Cluster

Data-storage, Sensors, Experiments

Internet, networks

What is Grid Computing?

- **The grid vision is of "Virtual computing" (+ information services to locate computation, storage resources)**
  - Compare: The web: "virtual documents" (+ search engine to locate them)
- **MOTIVATION: collaboration through sharing resources (and expertise) to expand horizons of**
  - Research
  - Commerce – engineering, …
  - Public service – health, environment,…



Grids: a foundation for e-Research

- Enabling a whole-system approach
- A challenge to the imagination
- Effect > Σparts

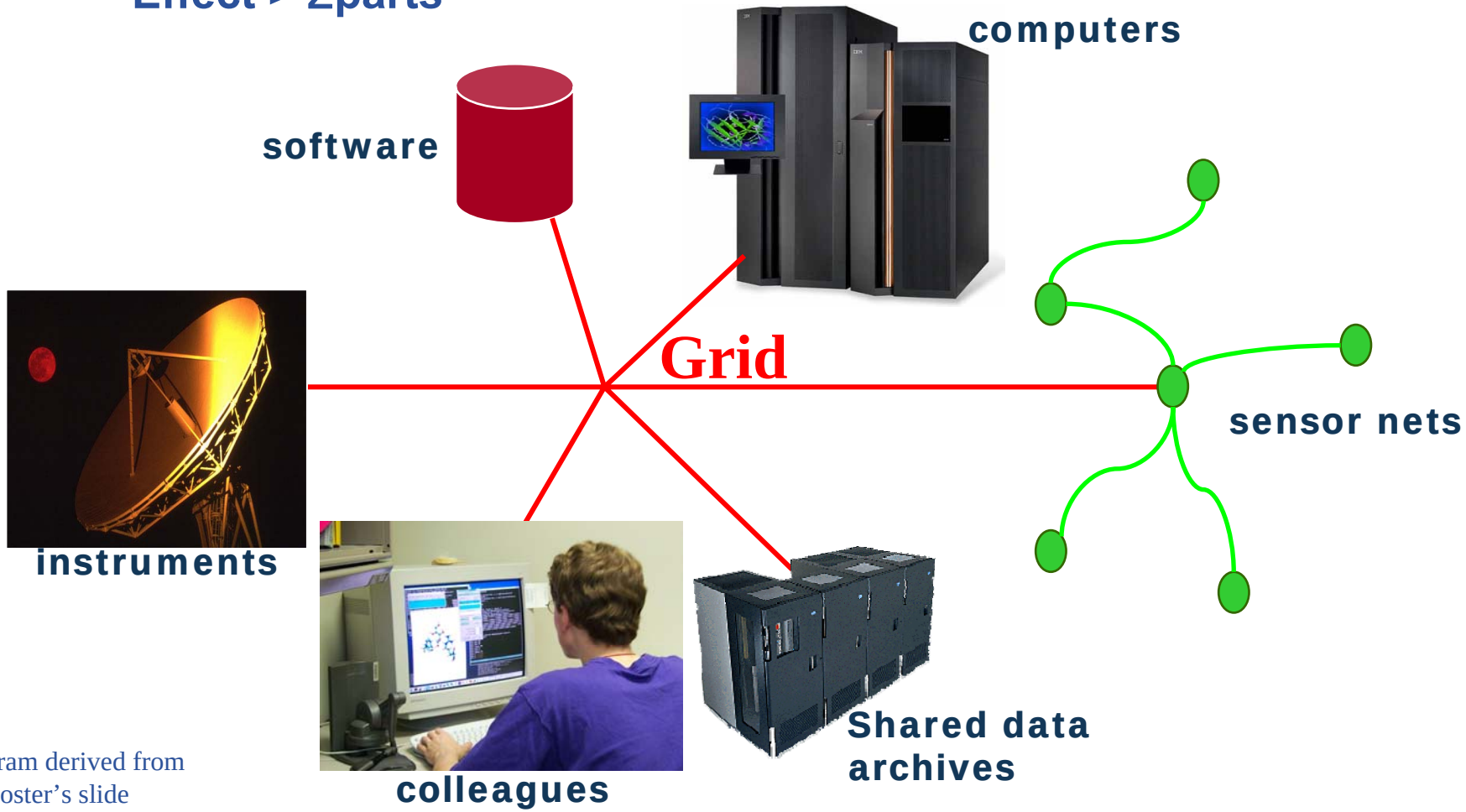


Diagram derived from Ian Foster's slide

"Effect > Σparts"

- **Flexible, simplified orchestration of resources available to a collaboration**
  - Across administrative domains
  - Abstractions hide detail of individual resources
    - Conform to Grid's procedures to gain benefit
  - Operations services (people and software)

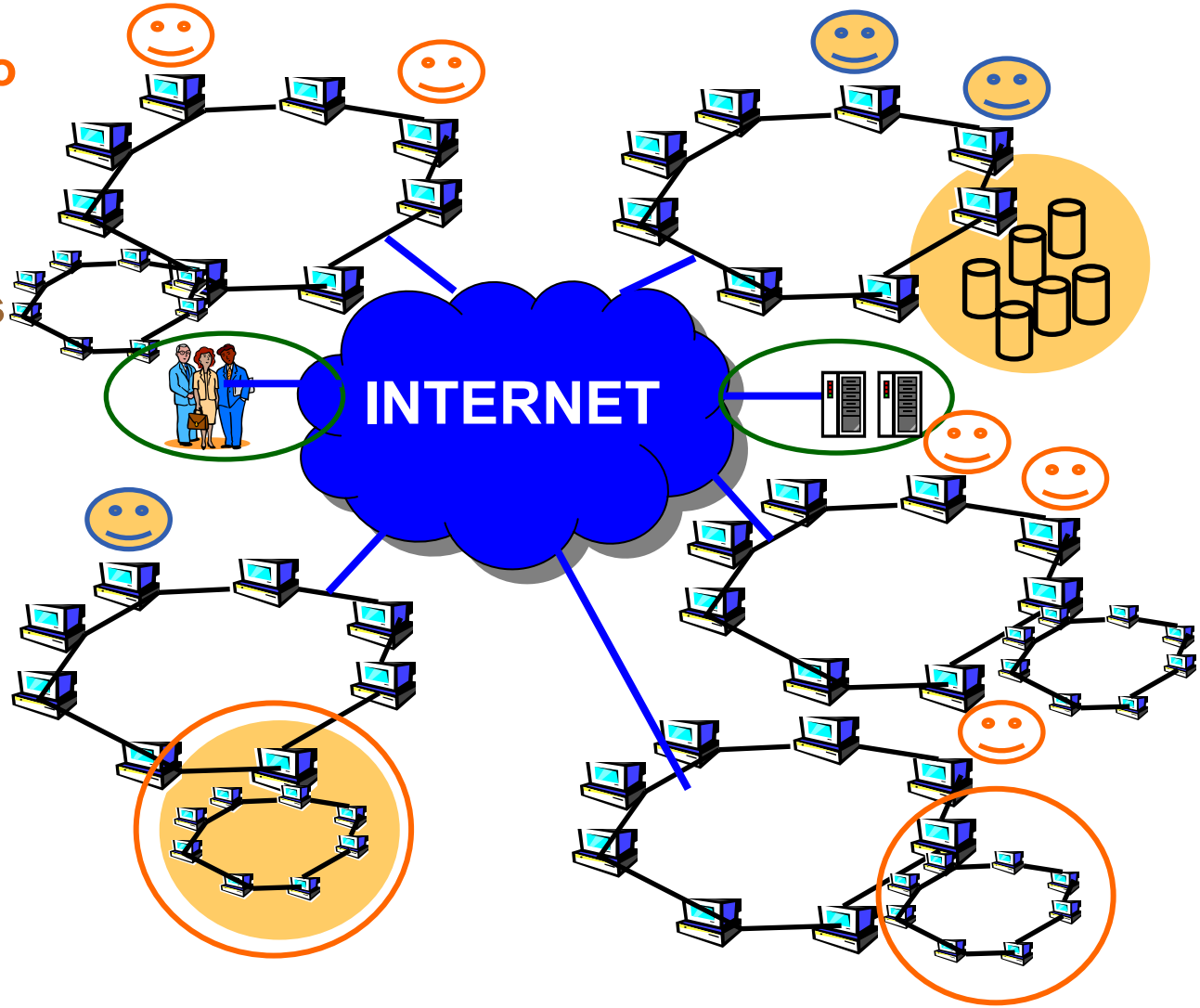
- **Increased utilisation**
  - A collaboration shares its resources building on Grid services
  - Collaborations share resources
    - Each contributes average requirements (cpus, storage)
    - Each can benefit from
      - *Heterogeneity*
      - *Scale*

Virtual organisations and grids

- **What is a Virtual Organisation?**
  - People in different organisations seeking to cooperate and share resources across their organisational boundaries
  - E.g. A research collaboration
- **Each grid is an infrastructure enabling one or more "virtual organisations" to share and access resources**
- **Each resource is exposed to the grid through an abstraction that masks heterogeneity, e.g.**
  - Multiple diverse computational platforms
  - Multiple data resources
- **Resources are usually owned by VO members. Negotiations lead to VOs sharing resources**

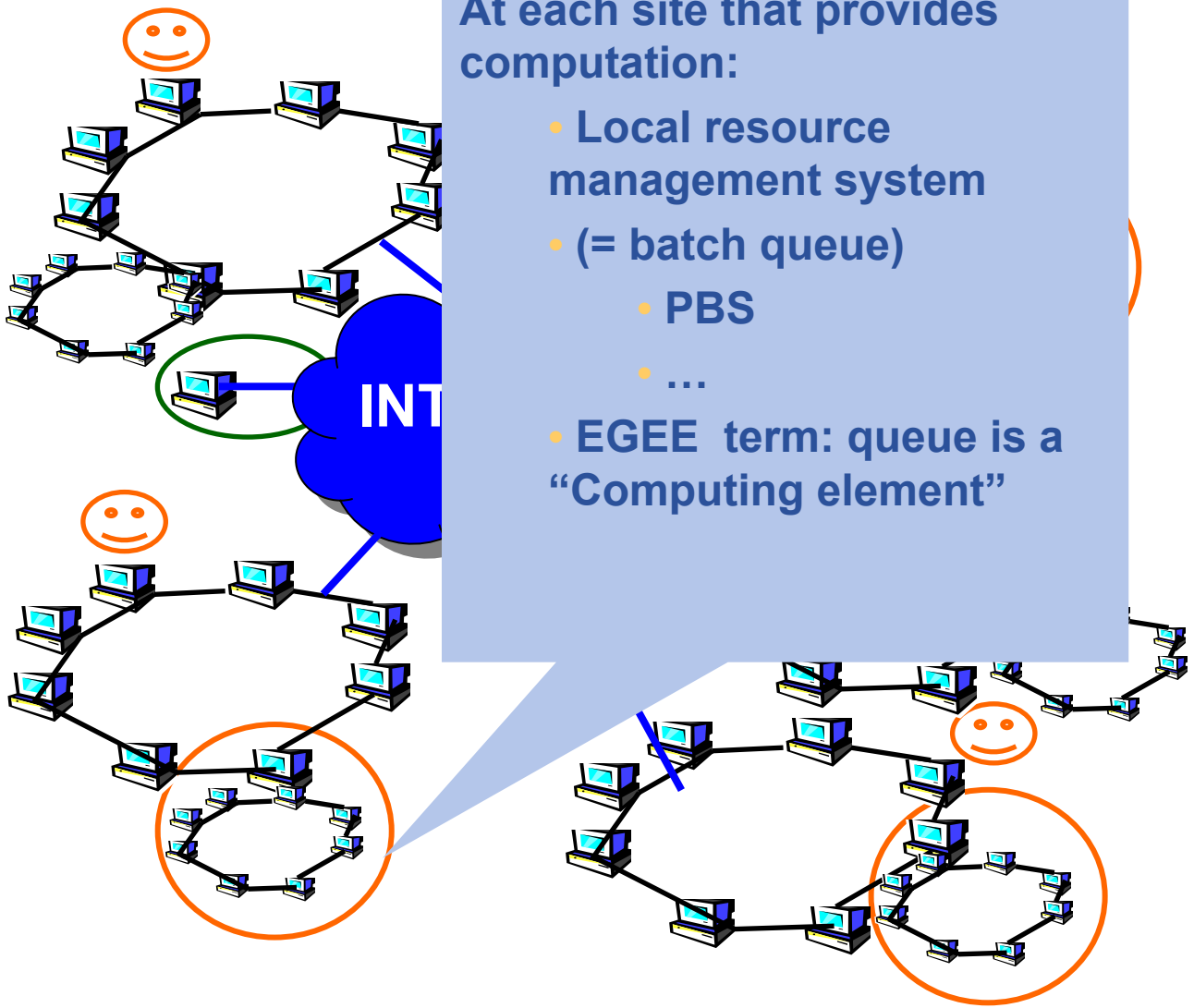
Typical current grid

- **Virtual organisations negotiate with sites to agree access to resources**
- **Grid middleware runs on each shared resource to provide**
  - Data services
  - Computation services
  - Single sign-on
- **Distributed services (both people and middleware) enable the grid**



Typical current grid

- **Grid middleware runs on each shared resource**
  - Data storage
  - (Usually) batch queues on pools of processors
- **Users join VO's**
- **Virtual organisation negotiates with sites to agree access to resources**
- **Distributed services (both people and middleware) enable the grid, allow single sign-on**



At each site that provides computation:

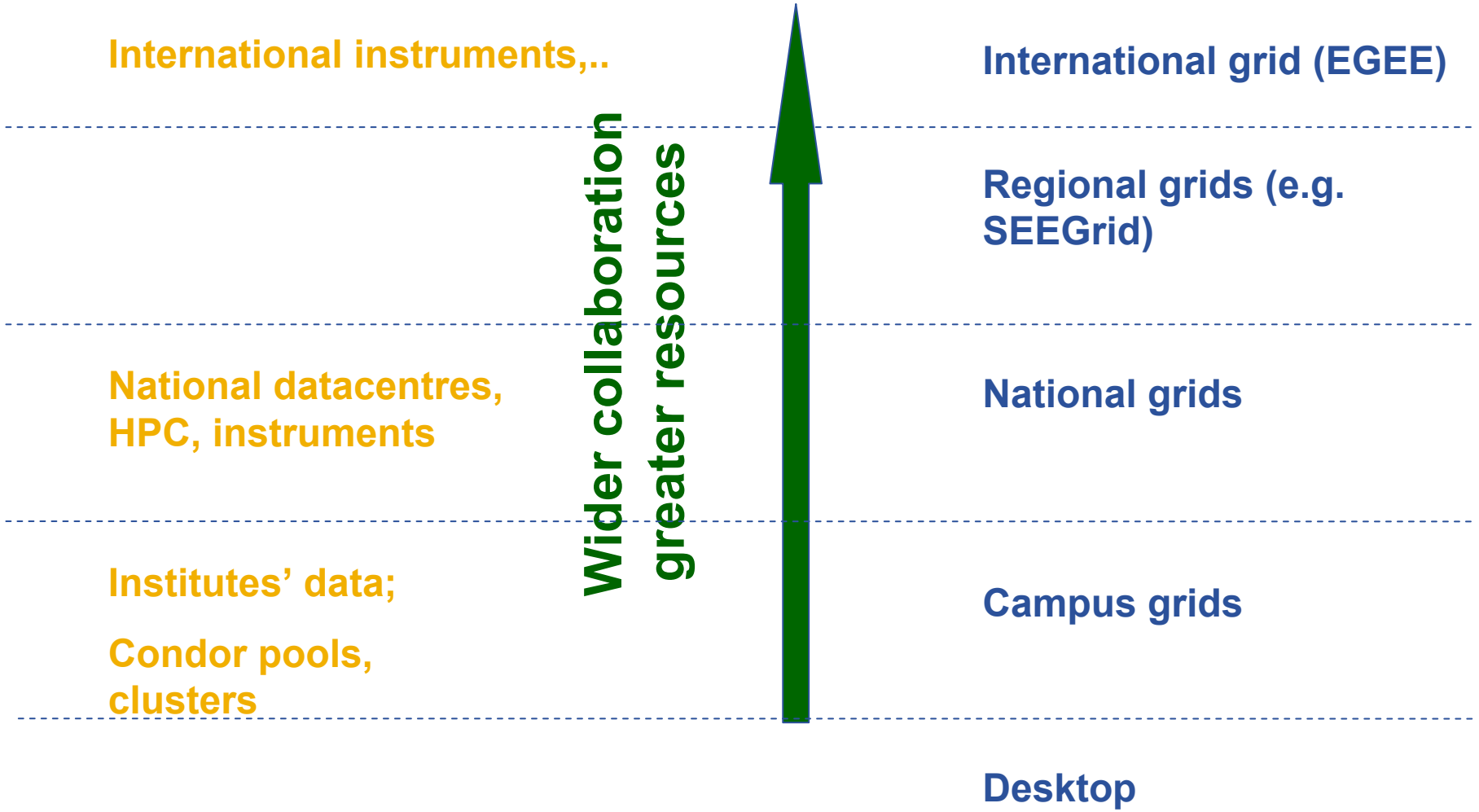
- **Local resource management system**
- **(= batch queue)**
  - **PBS**
  - **…**
- **EGEE term: queue is a "Computing element"**

Grid Middleware

- **When using a PC or workstation you**
  - Login with a username and password (“Authentication”)
  - Use rights given to you (“Authorisation”)
  - Run jobs
  - Manage files: create them, read/write, list directories
- **Components are linked by a bus**
- **Operating system**
- **One admin. domain**

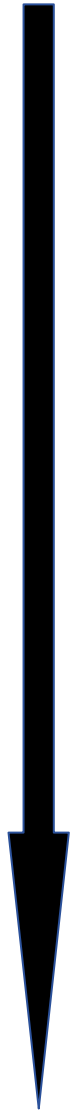
- **When using a Grid you**
  - Login with digital credentials – single sign-on (“Authentication”)
  - Use rights given you (“Authorisation”)
  - Run jobs
  - Manage files: create them, read/write, list directories
- **Services are linked by the Internet**
- **Middleware**
- **Many admin. domains**

The many scales of grids



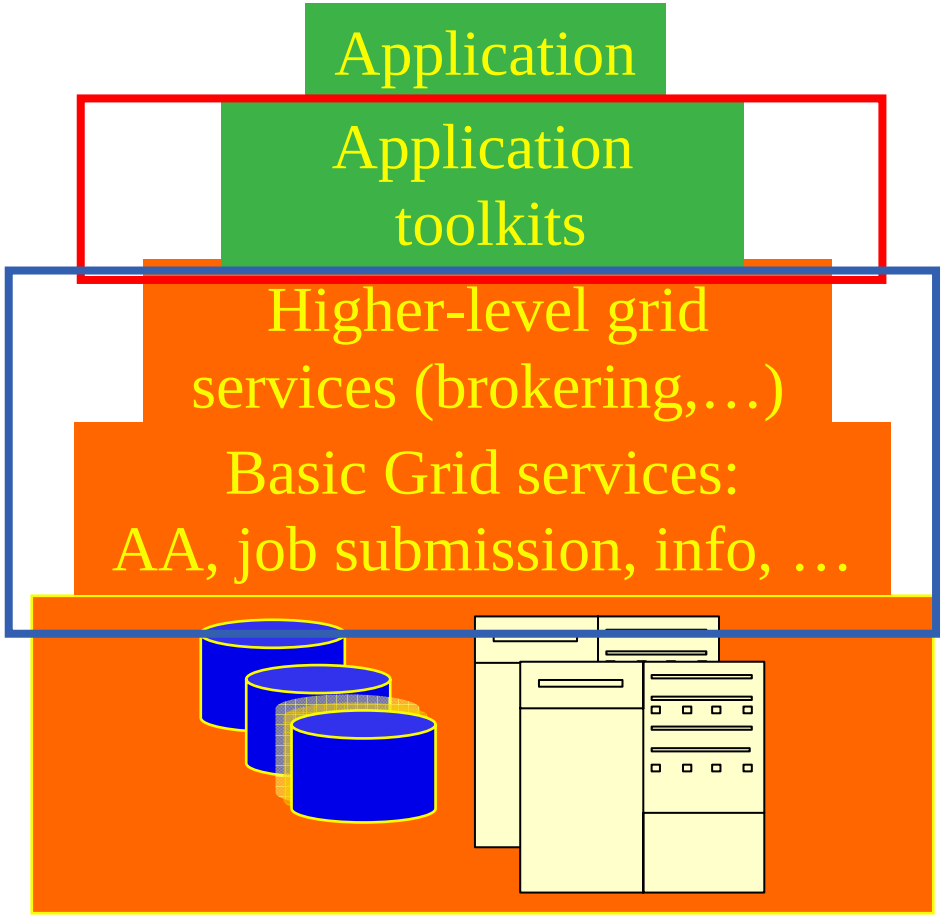
Different motivations for researchers

- **I need resources for my research**
  - I need richer functionality
    - MPI, parametric sweeps,…
    - Data and compute services together…
- **I provide an application for (y)our research**
  - How!?
    - Pre-install executables ?
    - Hosting environment?
    - Share data
    - Use it via portal?
- ***We* provide applications for (y)our research**
  - Also need:
    - Coordination of development
    - Standards
    - …



Engineering challenges increasing

Empowering VO's



Where computer science meets the application communities!

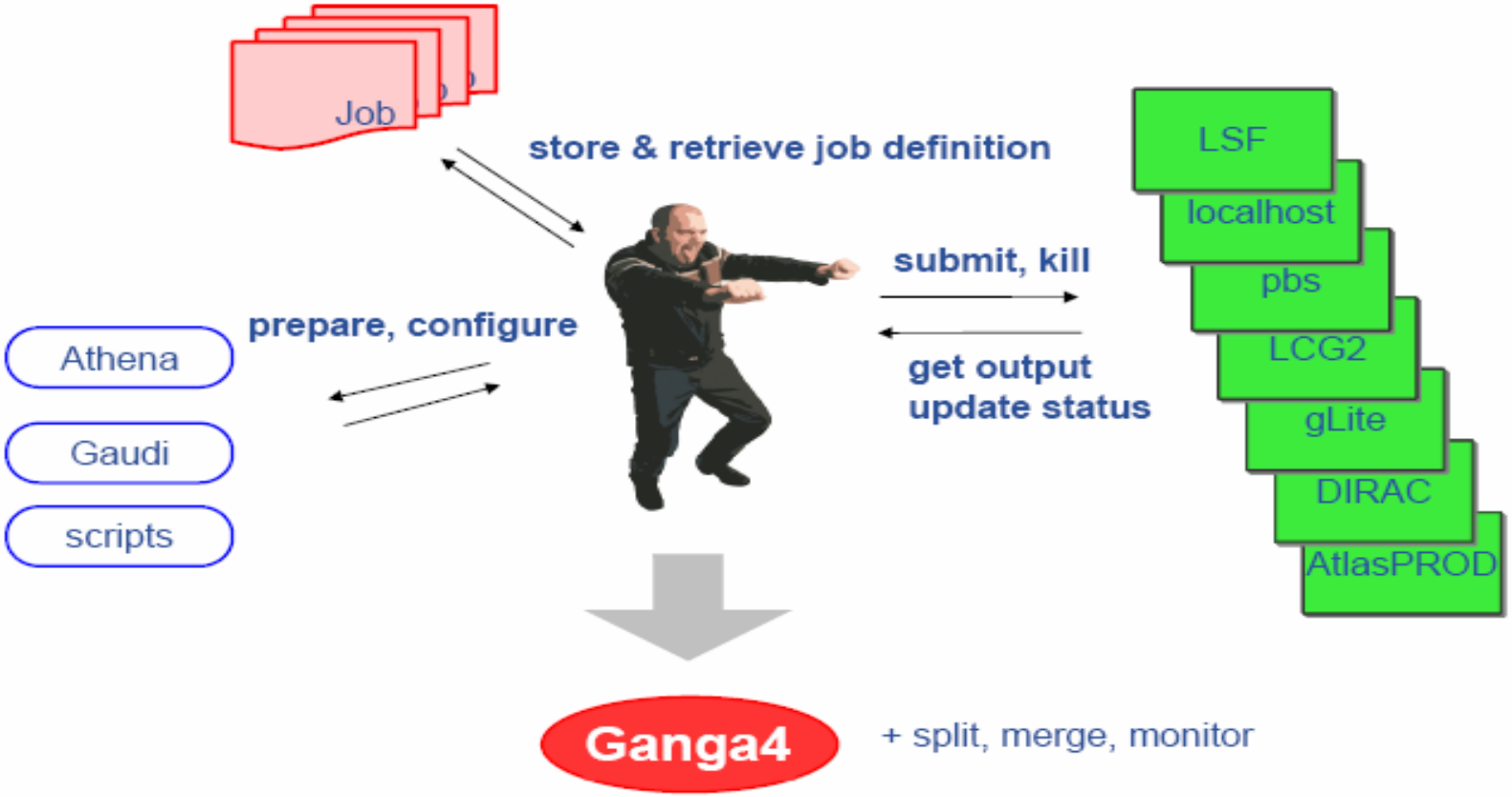
High level tools and

VO-specific developments:

- Portals
- Virtual Research Environments
- Semantics, ontologies
- Workflow
- Registries of VO services

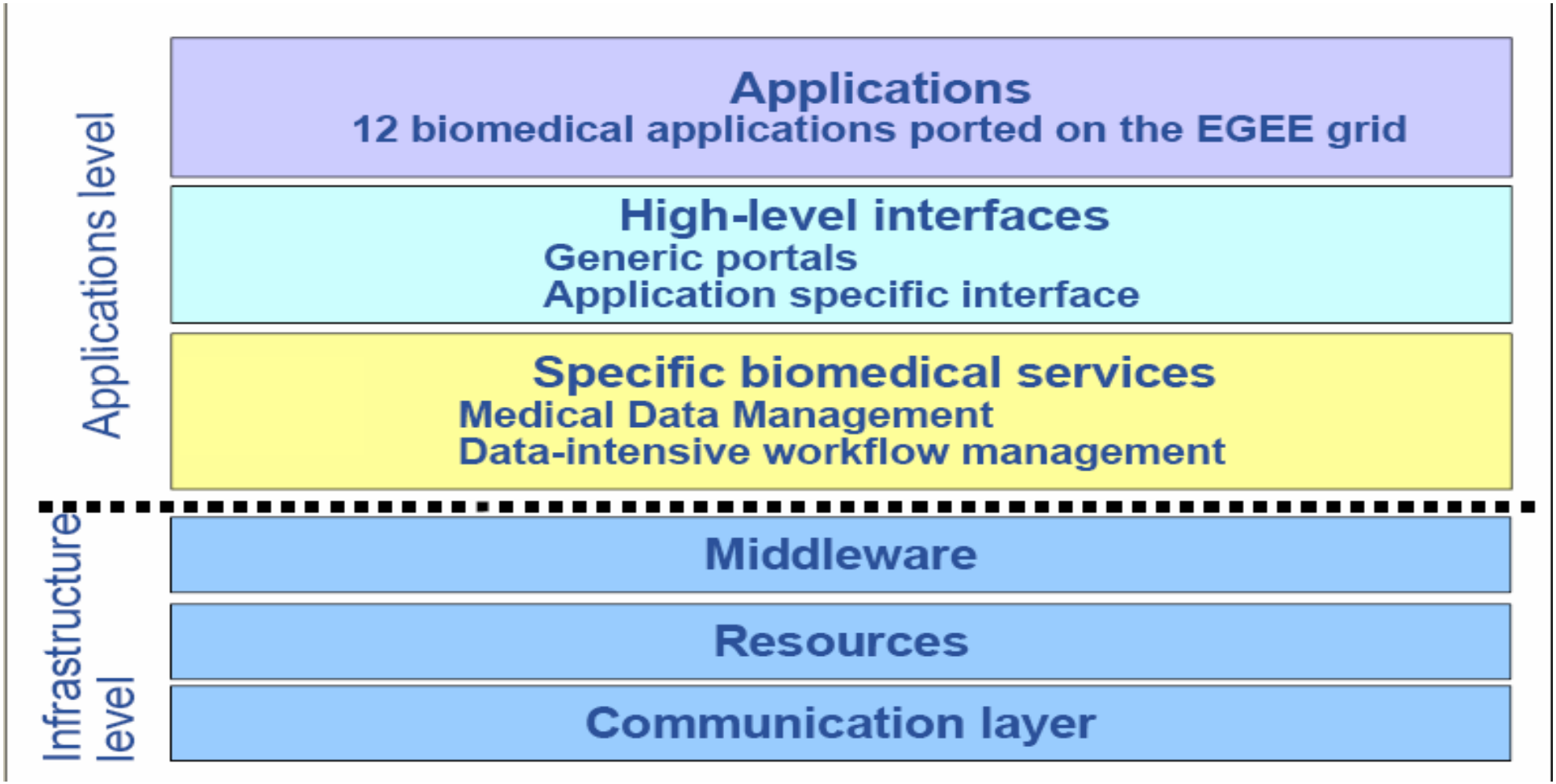
Production grids provide these services.

Example of higher-level service -1: GANGA



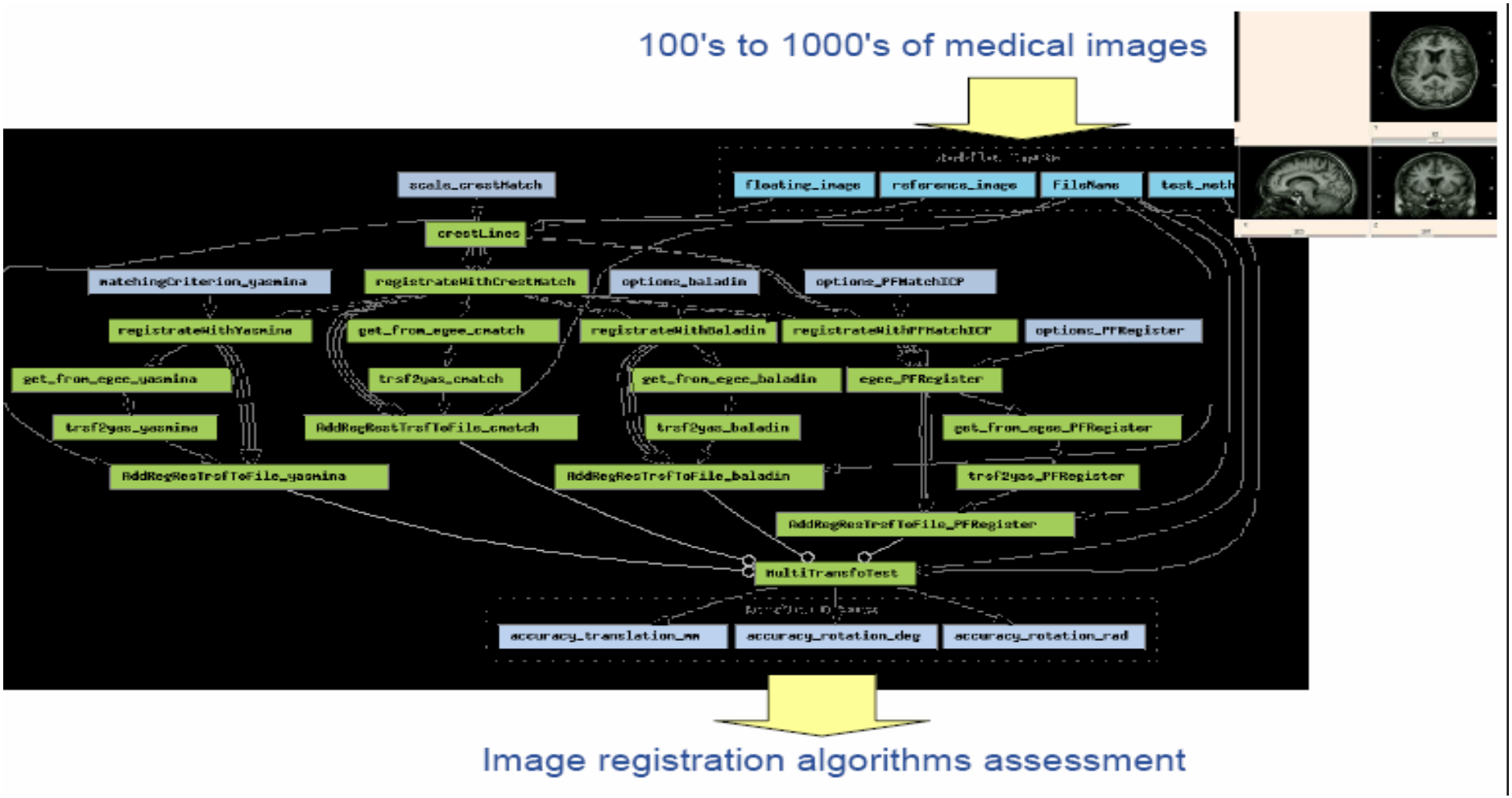
- **Ganga is a lightweight user tool**
  ganga.web.cern.ch/
- **But also: Ganga is a developer framework**

Example – Biomedical applications

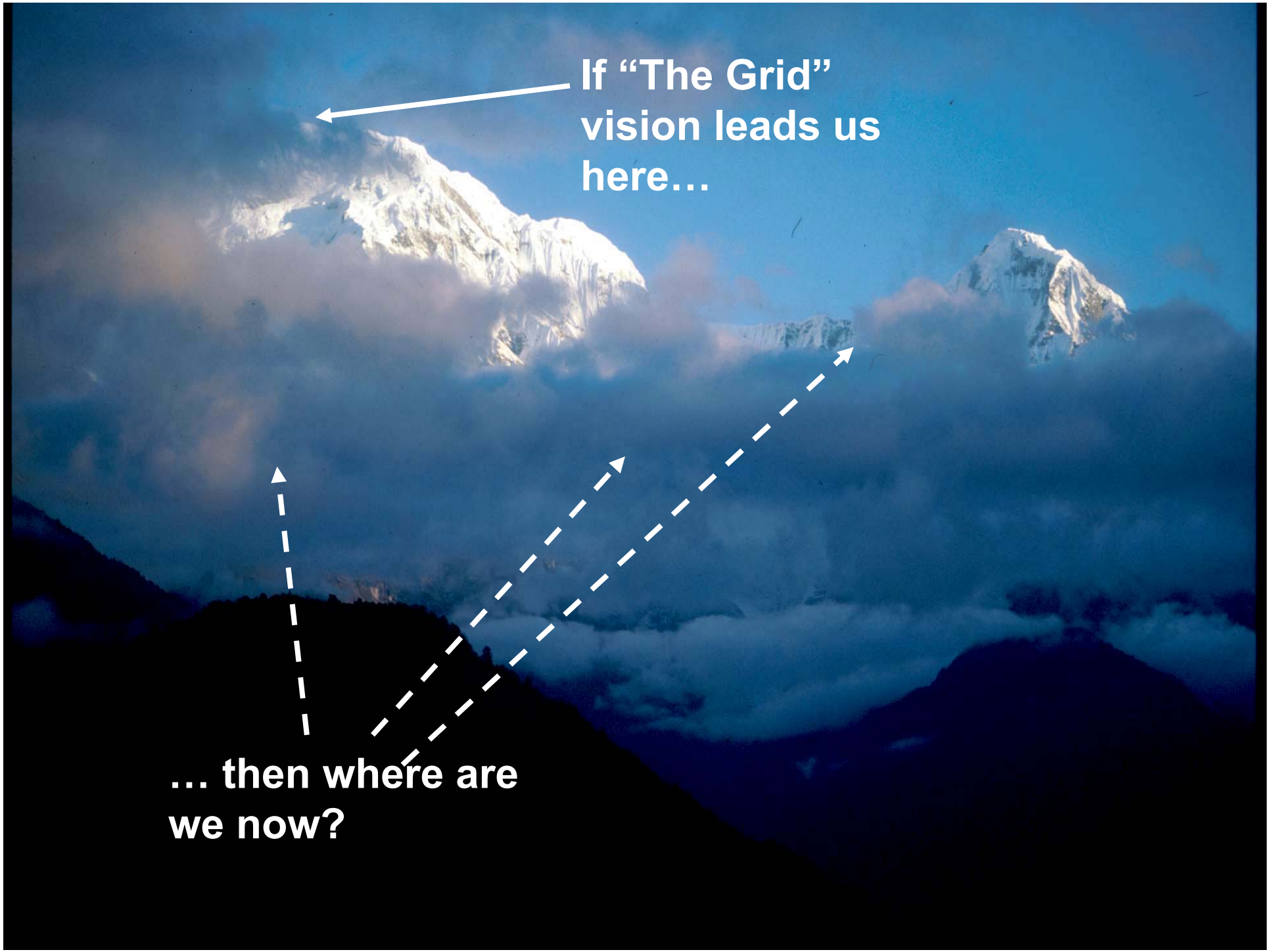


Biomedical community and the Grid, EGEE User Forum, March 1st 2006, I. Magnin

Example of Workflow



Biomedical community and the Grid, EGEE User Forum, March 1st 2006, I. Magnin


If “The Grid” vision leads us here…
… then where are we now?

Where are we now? –user's view

Networks

Web

Grids

Arts
Humanities
e-Soc-Sci

Sciences,
engineering

Early production grids:
International - EGEE

Types of use:

Service-oriented, workflow, "legacy" data

High throughput, new data

Grids: where are we now?

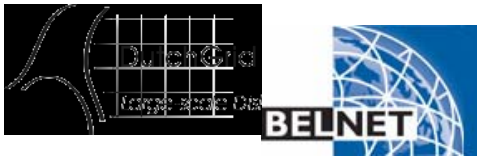
- **Many key concepts identified and known**
- **Many grid projects have tested, and benefit from, these**
  - Empowering collaborations
  - Resource-sharing
- **Major efforts now on establishing:**
  - **Production Grids *for multiple VO's***
    - "Production" = Reliable, sustainable, with commitments to quality of service
    - Each has
      - *One stack of middleware that serves many research communities*
      - *Establishing operational procedures and organisation*
    - Challenge for EGEE-II: federate these!
  - **Standards** (a slow process)
    - e.g. Open (formerly Global) Grid Forum, http://www.gridforum.org/
    - Extending web services
  - **Broadening range of research communities**
    - arts and humanities, social science …

- To obtain a Google map of the Grids in the Globus Interoperability Now initiative go to:

http://www.pparc.ac.uk/Nw/GIN.asp

- (You will need to install GoogleEarth)

National grid initiatives now include…



CroGrid

Grid security and trust

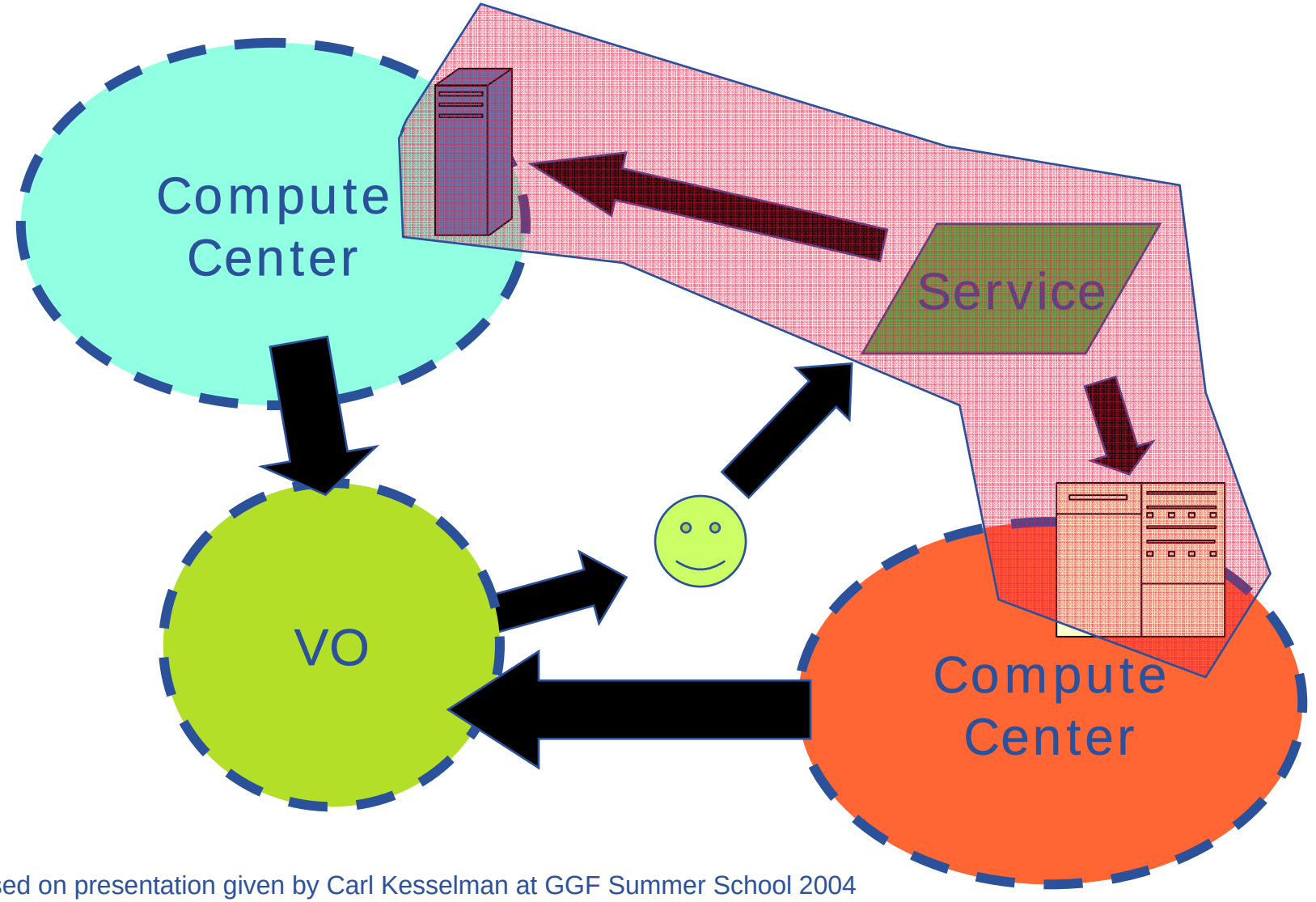
- **Providers of resources (computers, databases,..) need risks to be controlled: they are asked to trust users they do not know**

- **User's need**
  - single sign-on: to be able to logon to a machine that can pass the user's identity to other resources
  - To trust owners of the resources they are using

- **Build middleware on layer providing:**
  - *Authentication:* know who wants to use resource
  - *Authorisation:* know what the user is allowed to do
  - *Security:* reduce vulnerability, e.g. from outside the firewall
  - *Non-repudiation:* knowing who did what

- **The "Grid Security Infrastructure" middleware is the basis of (most) production grids**

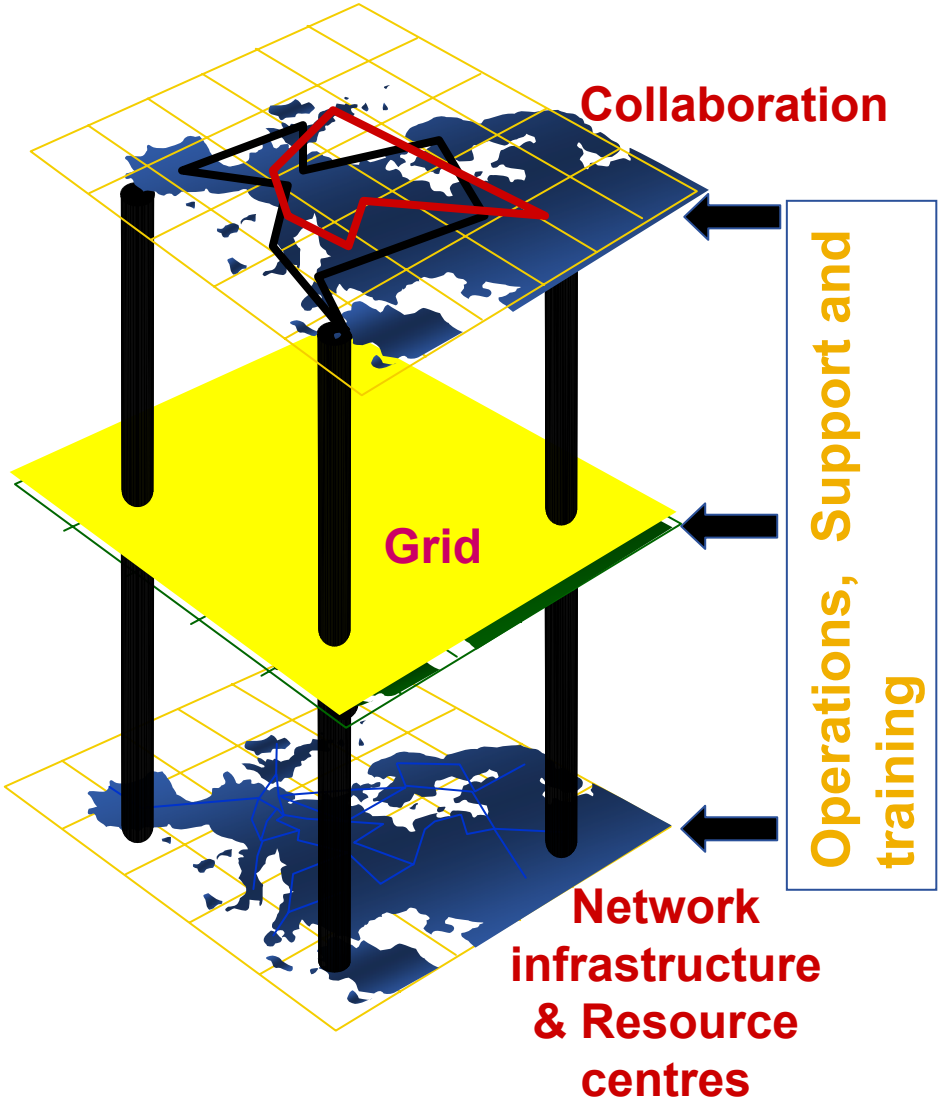
The Role of the Virtual Organisation (VO)



slide based on presentation given by Carl Kesselman at GGF Summer School 2004

What are Grids? - Summary

- **Grids enable virtual computing across administrative domains**
  - Resources share authorisation and authentication
  - Resources accessed thru abstractions
- **Motivations:**
  - Collaborative research, diagnostics, engineering, public service,..
  - Resource utilisation and sharing



More about the EGEE project: Enabling Grids for E-SciencE

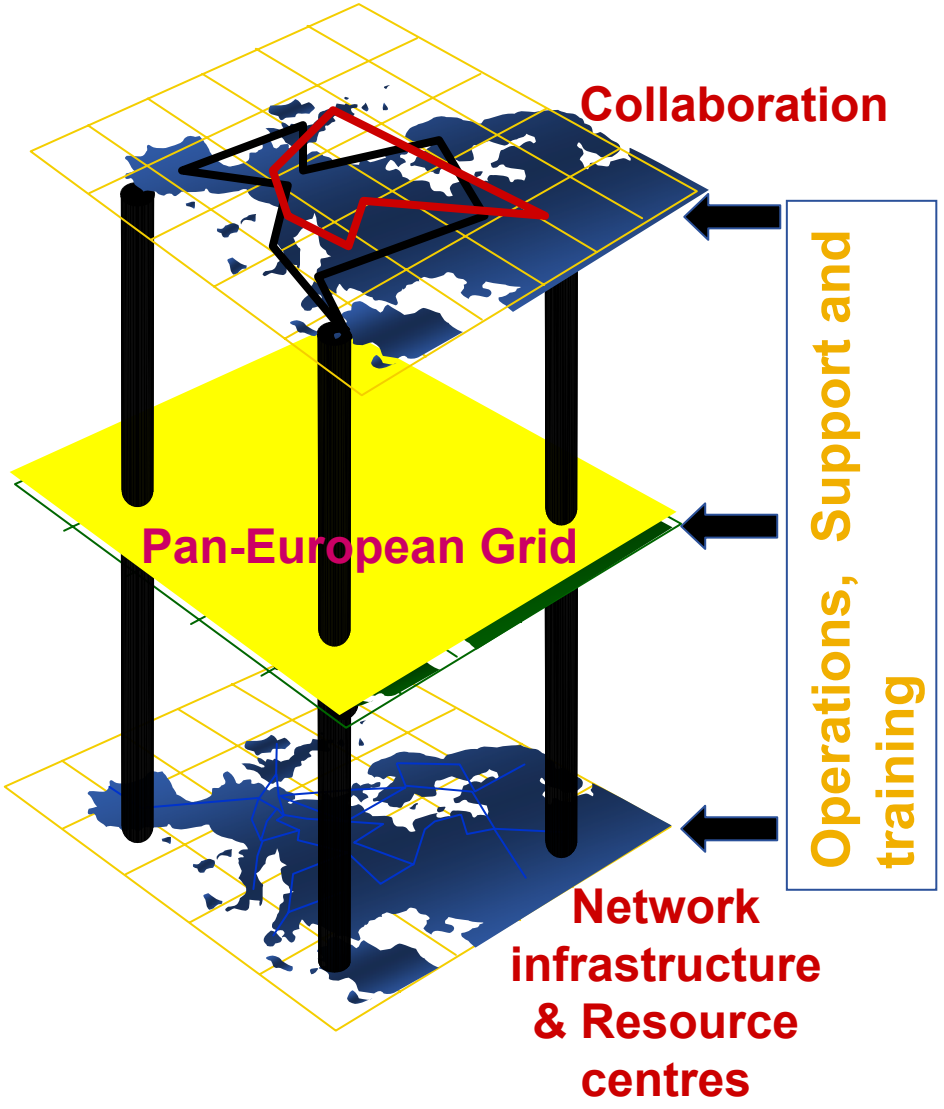
Contents

- **What is EGEE?**
  - Goals
  - Status
  - Activities
- **Grid services: gLite 3.0**
- **Sources of further information**

EGEE – international e-infrastructure

A four year programme:

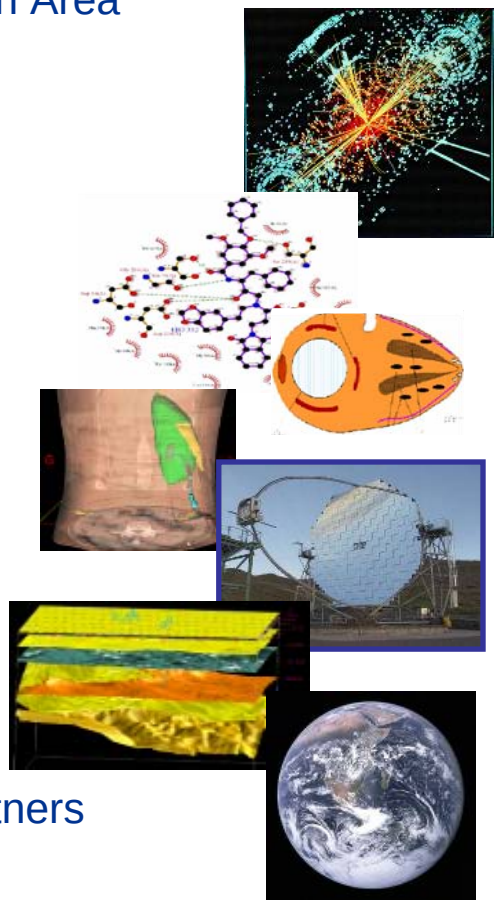
- **Build, deploy and operate a consistent, robust a large scale production grid service that**
  - Links with and build on national, regional and international initiatives
- **Improve and maintain the middleware in order to deliver a reliable service to users**
- **Attract new users from research and industry and ensure training and support for them**



EGEE Mission

- **Infrastructure**
  - Manage and operate production Grid for European Research Area
  - Interoperate with e-Infrastructure projects around the globe
  - Contribute to Grid standardisation efforts

- **Support applications from diverse communities**
  - High Energy Physics
  - Biomedicine
  - Earth Sciences
  - Astrophysics
  - Computational Chemistry
  - Fusion
  - Geophysics
  - Finance, Multimedia
  - …

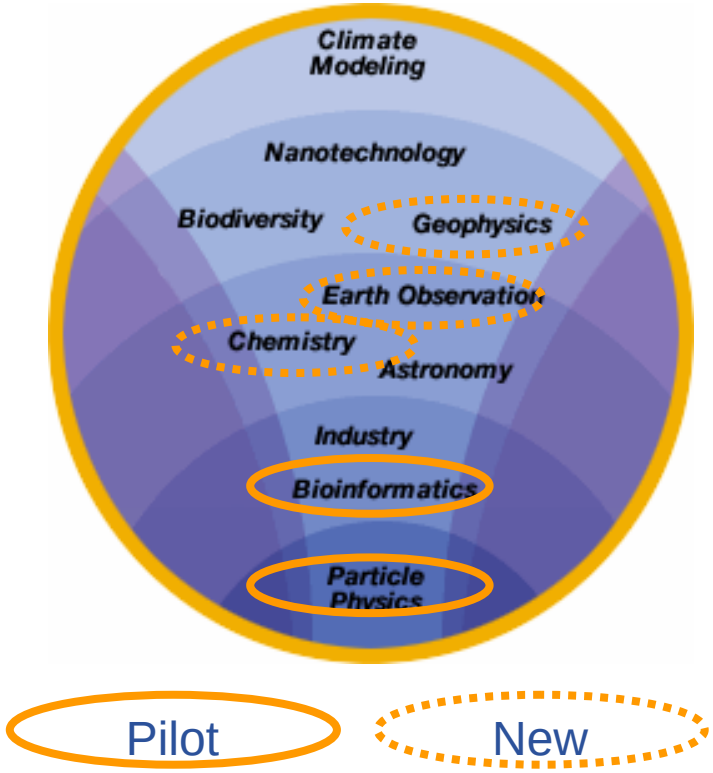


- **Business**
  - Forge links with the full spectrum of interested business partners

+ Disseminate knowledge about the Grid through training
+ Prepare for sustainable European Grid Infrastructure

In the first 2 years (to March 2006)

- **Established production quality sustained Grid services**
  - 3000 users from at least 5 disciplines
  - Goal was to integrate 50 sites into a common infrastructure → currently 180
  - offer 5 Petabytes (10^{15}) storage
- **Demonstrated a viable general process to bring other application communities on board**
- **Secured a second phase from April 2006**



- **Natural continuation of EGEE**
  - Expanded consortium
  - Emphasis on providing an infrastructure
    - → increased support for applications
    - → interoperate with other infrastructures
    - → more involvement from Industry

SA: service activities

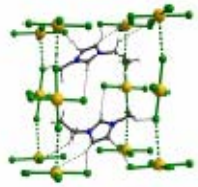
- establishing operations

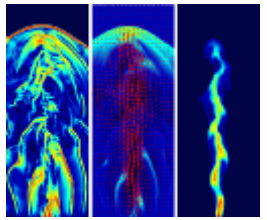
NA: network activities

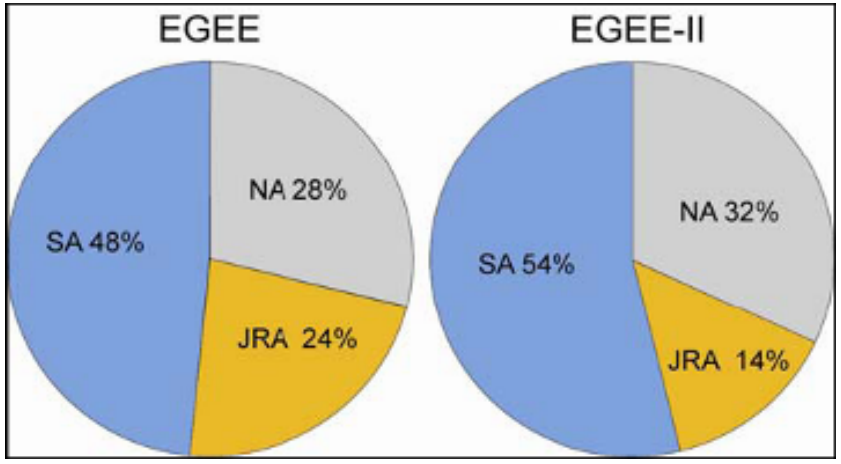
- supporting VOs

JRA: “joint research activities”

- e.g. hardening middleware







EGEE-II: Expertise & Resources

- More than 90 partners
- 32 countries
- 12 federations

→ Major and national Grid projects in Europe, USA, Asia

+ 27 countries through related projects:
  - BalticGrid
  - SEE-GRID
  - EUMedGrid
  - EUChinaGrid
  - EELA

The EGEE Infrastructure

Test-beds & Services

- Certification testbeds (SA3)
- Pre-production service
- Production service

Support Structures

- Operations Coordination Centre
- Regional Operations Centres
- Global Grid User Support
- EGEE Network Operations Centre (SA2)
- Operational Security Coordination Team

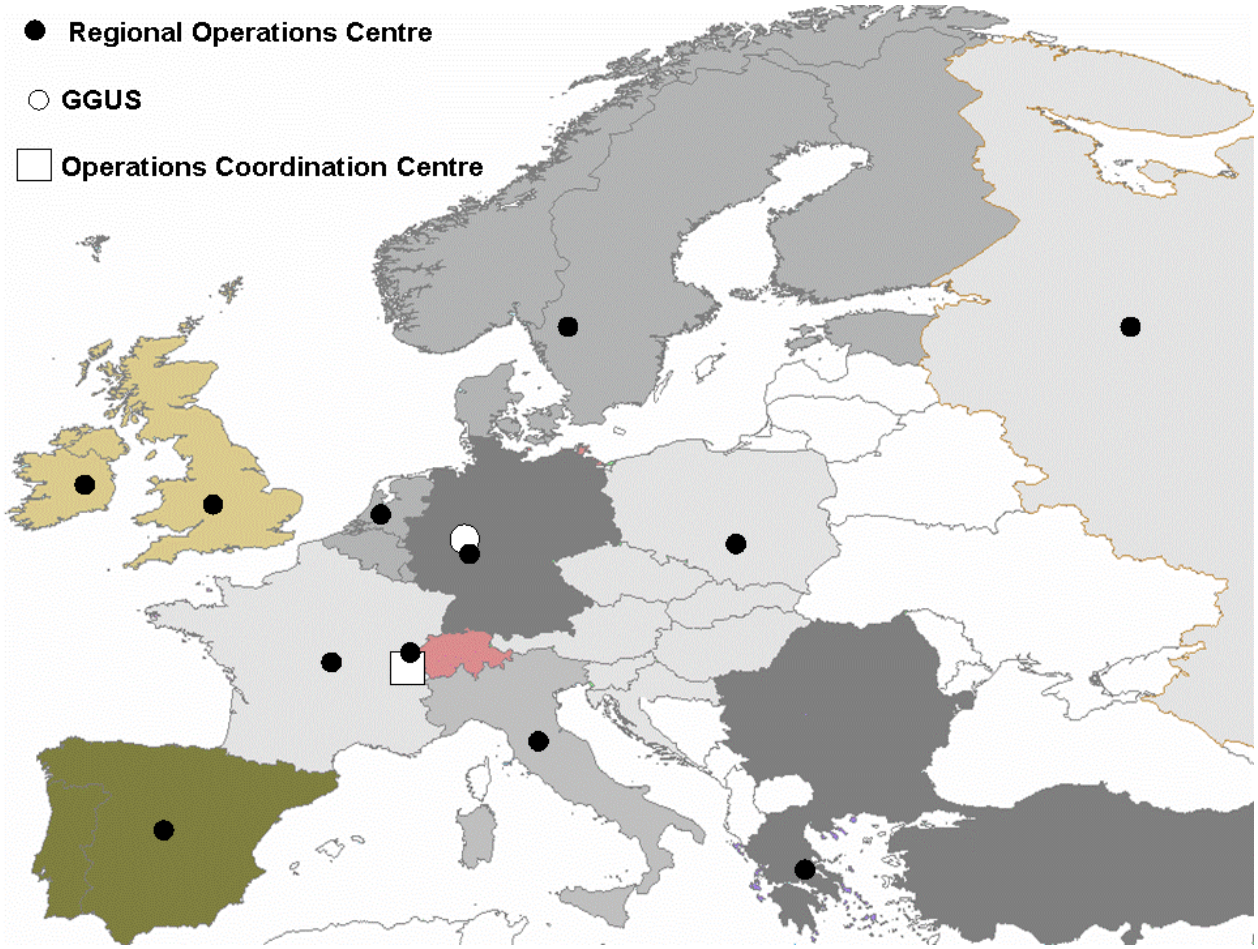
Infrastructure:

- Physical test-beds & services
- Support organisations & procedures
- Policy groups

Security & Policy Groups

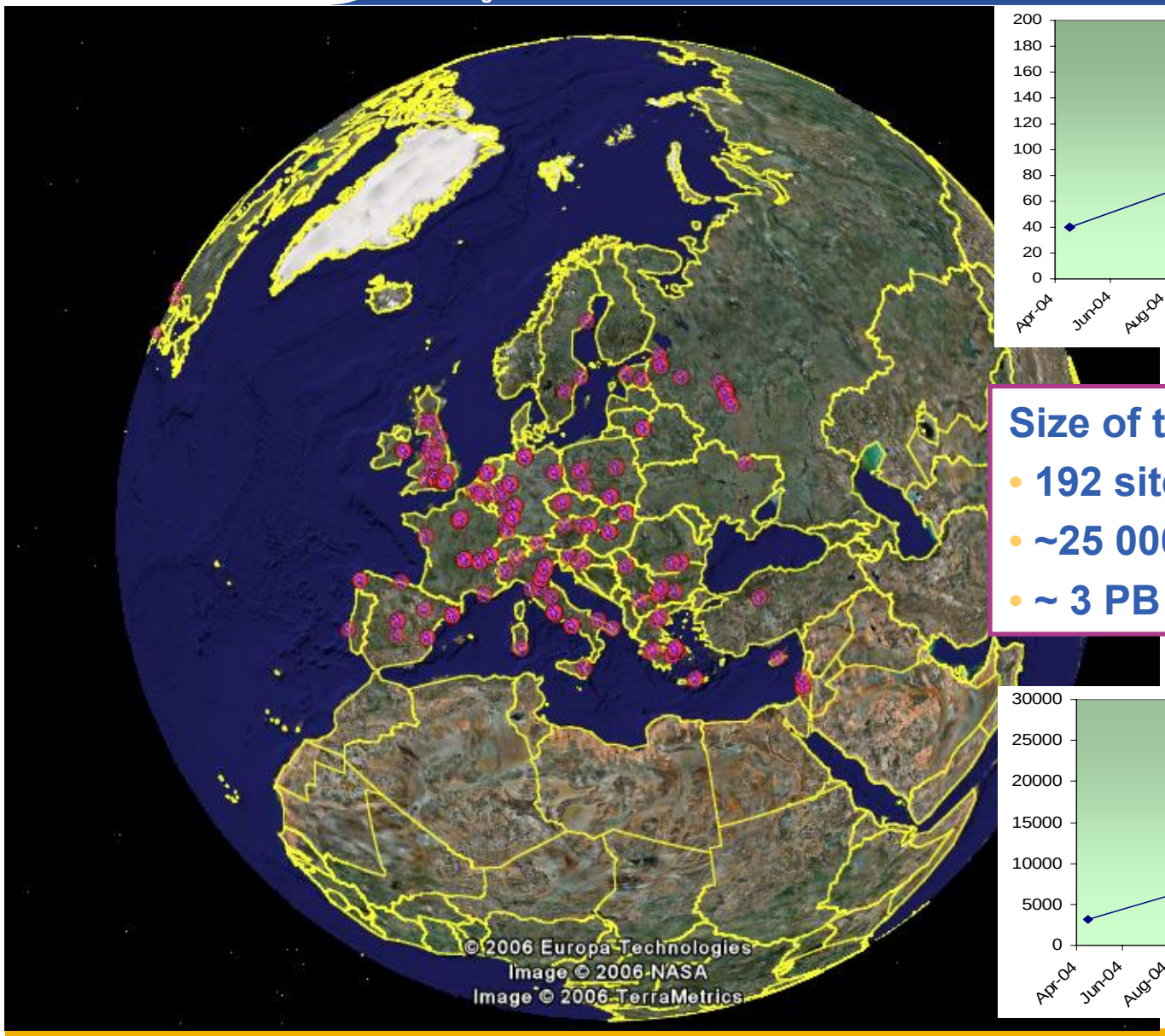
- Joint Security Policy Group
- EuGridPMA (& IGTF)
- Grid Security Vulnerability Group
- Operations Advisory Group (+NA4)

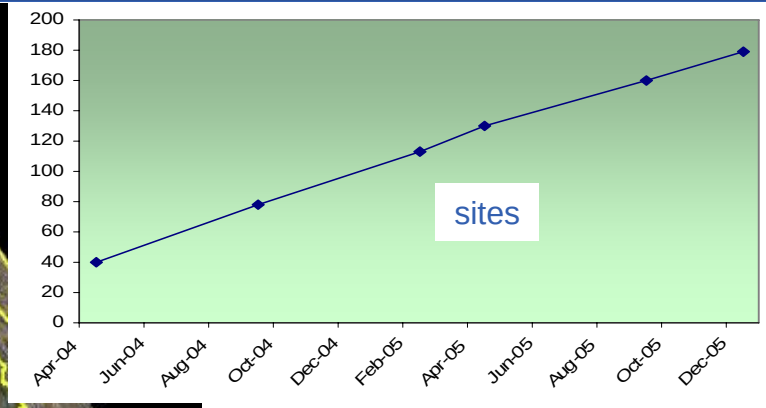
Grid management: structure



- **Operations Coordination Centre (OCC)**
  - management, oversight of all operational and support activities
- **Regional Operations Centres (ROC)**
  - providing the core of the support infrastructure, each supporting a number of resource centres within its region
  - **Grid Operator on Duty**
- **Resource centres**
  - providing resources (computing, storage, network, etc.);
- **Grid User Support (GGUS)**
  - At FZK, coordination and management of

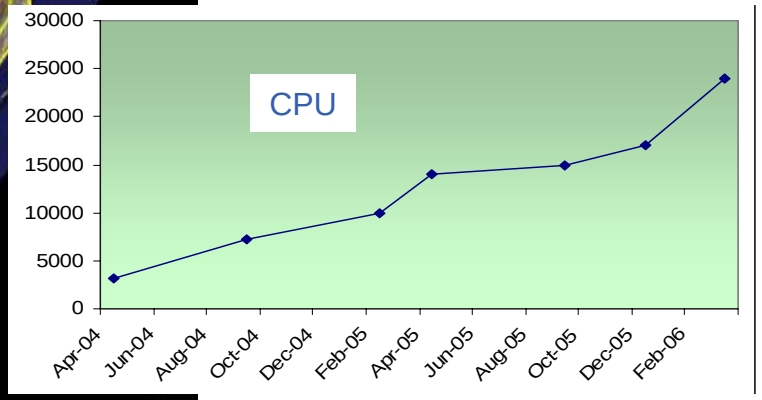
Production service



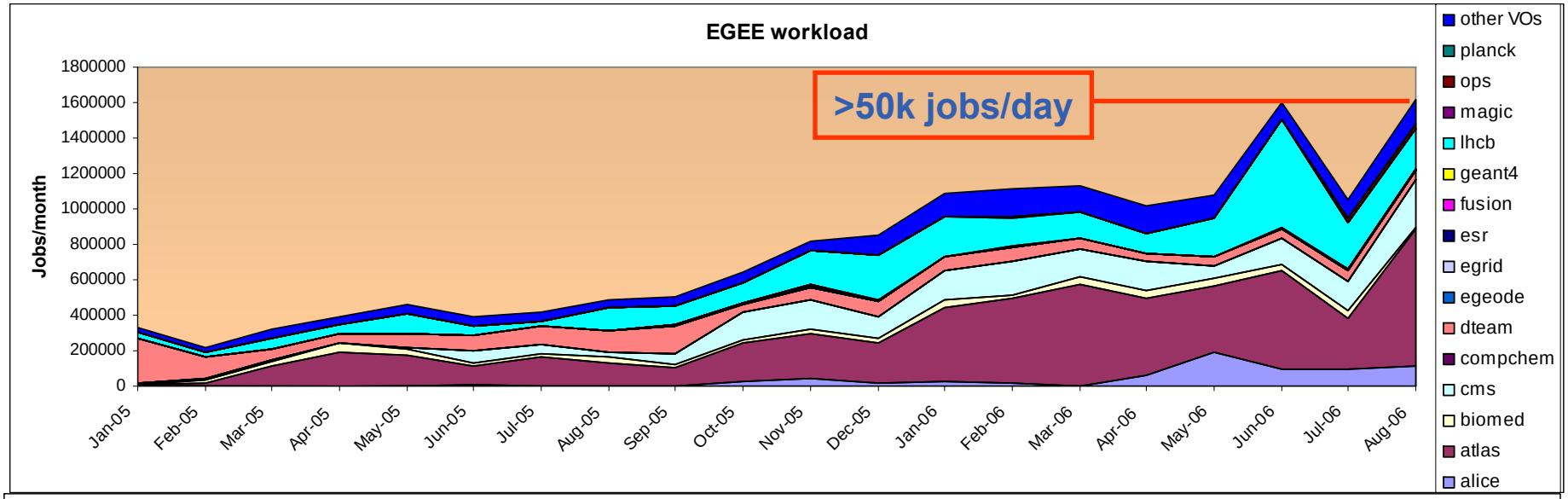


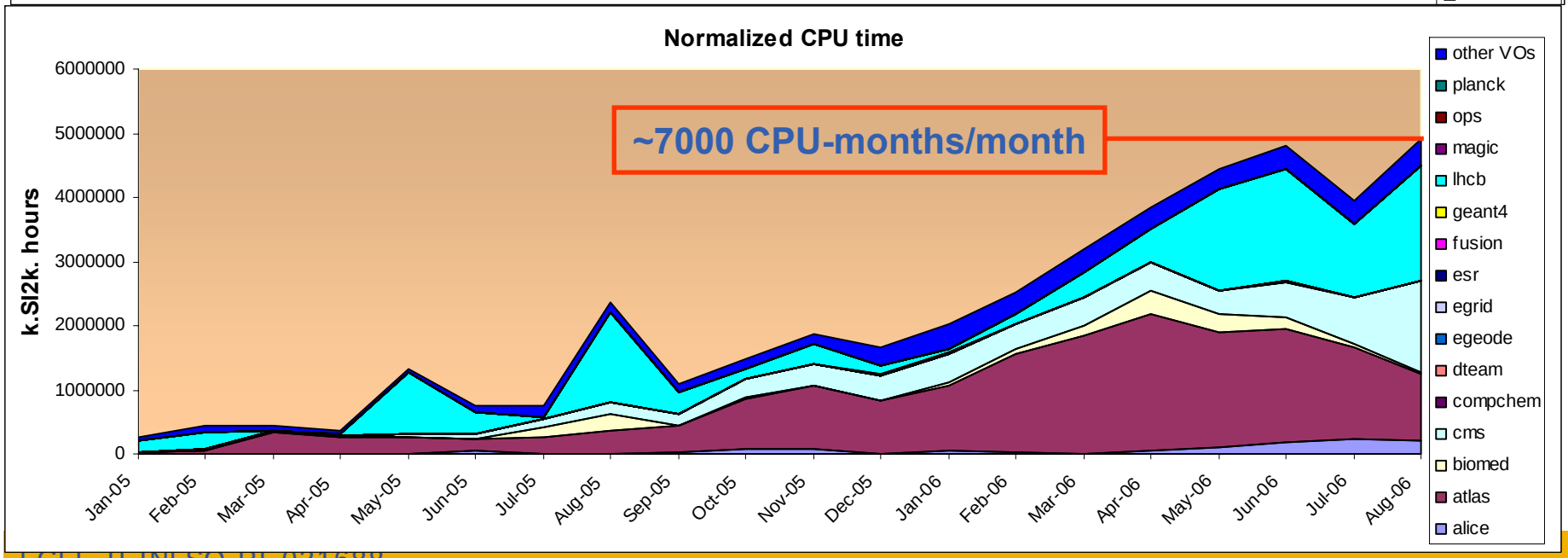
Size of the infrastructure today:

- **192 sites in 40 countries**
- **~25 000 CPU**
- **~ 3 PB disk, + tape MSS**

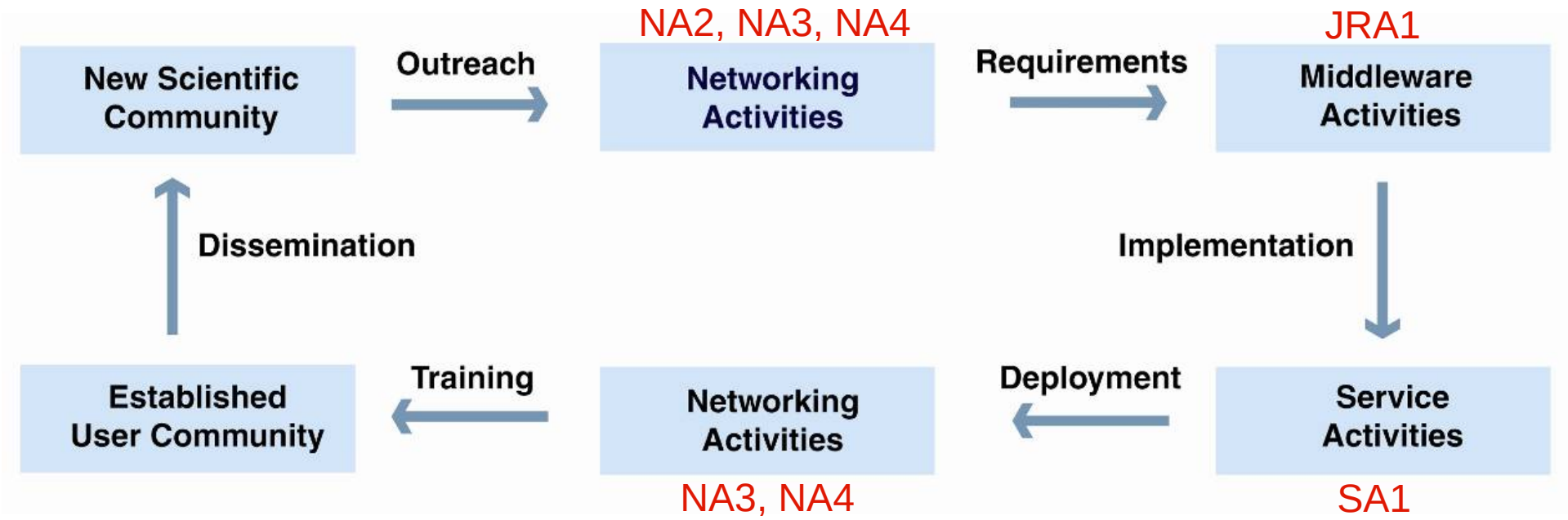


Usage of the infrastructure





The EGEE Virtuous Cycle



Building effective user communities

Related projects: infrastructure, engineering, education

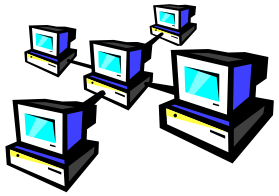
Name	Description
BalticGrid	EGEE extension to Estonia, Latvia, Lithuania
EELA	EGEE extension to Brazil, Chile, Cuba, Mexico, Argentina
EUChinaGRID	EGEE extension to China
EUMedGRID	EGEE extension to Malta, Algeria, Morocco, Egypt, Syria, Tunisia, Turkey
ISSeG	Site security
eIRGSP	Policies
ETICS	Repository, Testing
OMII-Europe	to provide key software components for building e-infrastructures;
BELIEF	Digital Library of Grid documentation, organisation of workshops, conferences
BIOINFOGRID	Biomedical
Health-e-Child	Biomedical – Integration of heterogeneous biomedical information for improved healthcare
ICEAGE	International Collaboration to Extend and Advance Grid Education

Main components

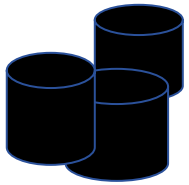
User Interface (UI): The place where users logon to the Grid

Resource Broker (RB): Matches the user requirements with the available resources on the Grid

Information System: Characteristics and status of CE and SE (Uses “GLUE schema”)

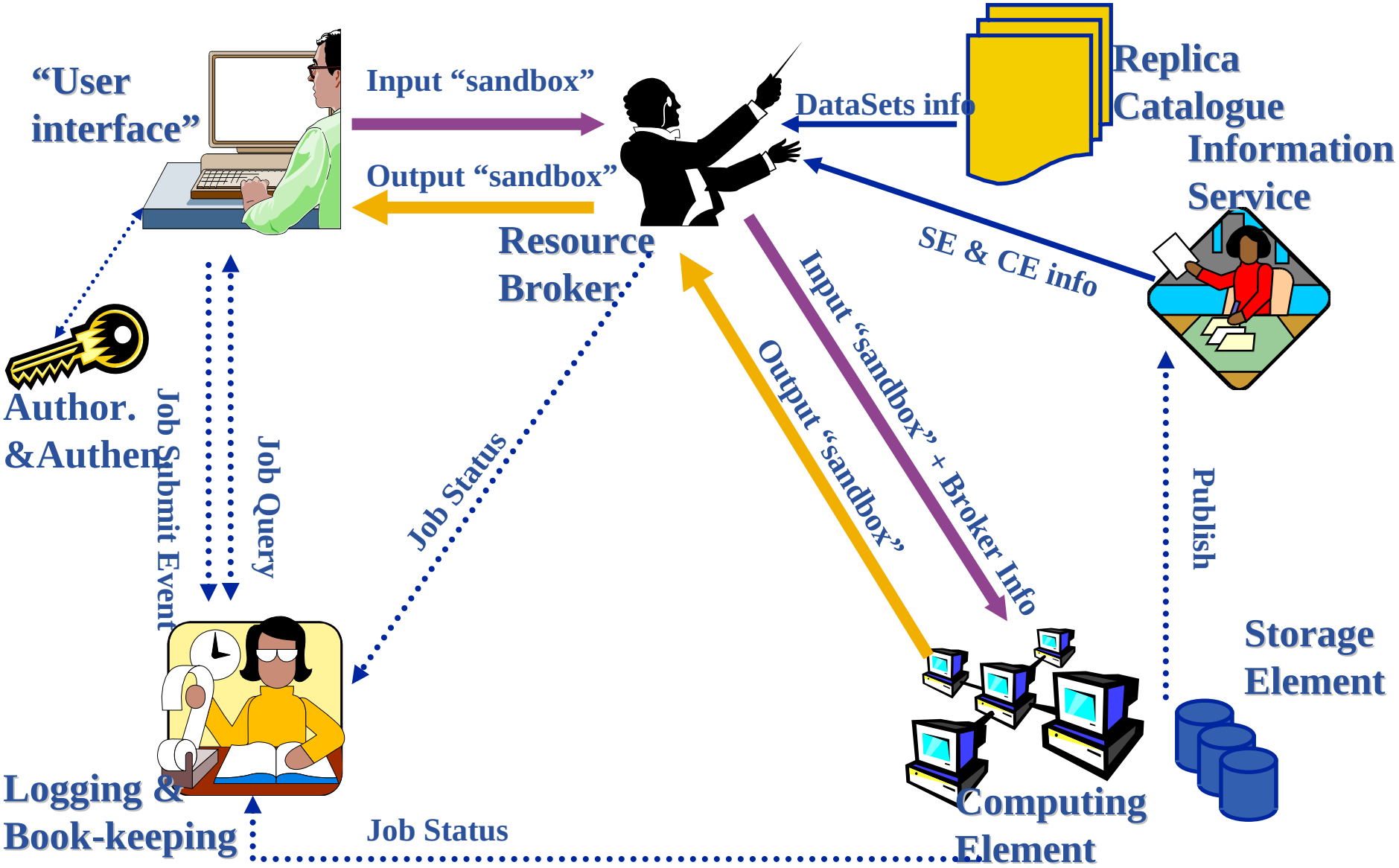


Computing Element (CE): A batch queue on a site’s computers where the user’s job is executed



Storage Element (SE): provides (large-scale) storage for files

Current production middleware



Building on basic tools and Information Service

- Submit job to grid via the “resource broker (RB)”,
- glite_job_submit *my.jdl*

Returns a “job-id” used to monitor job, retrieve output

Example JDL file

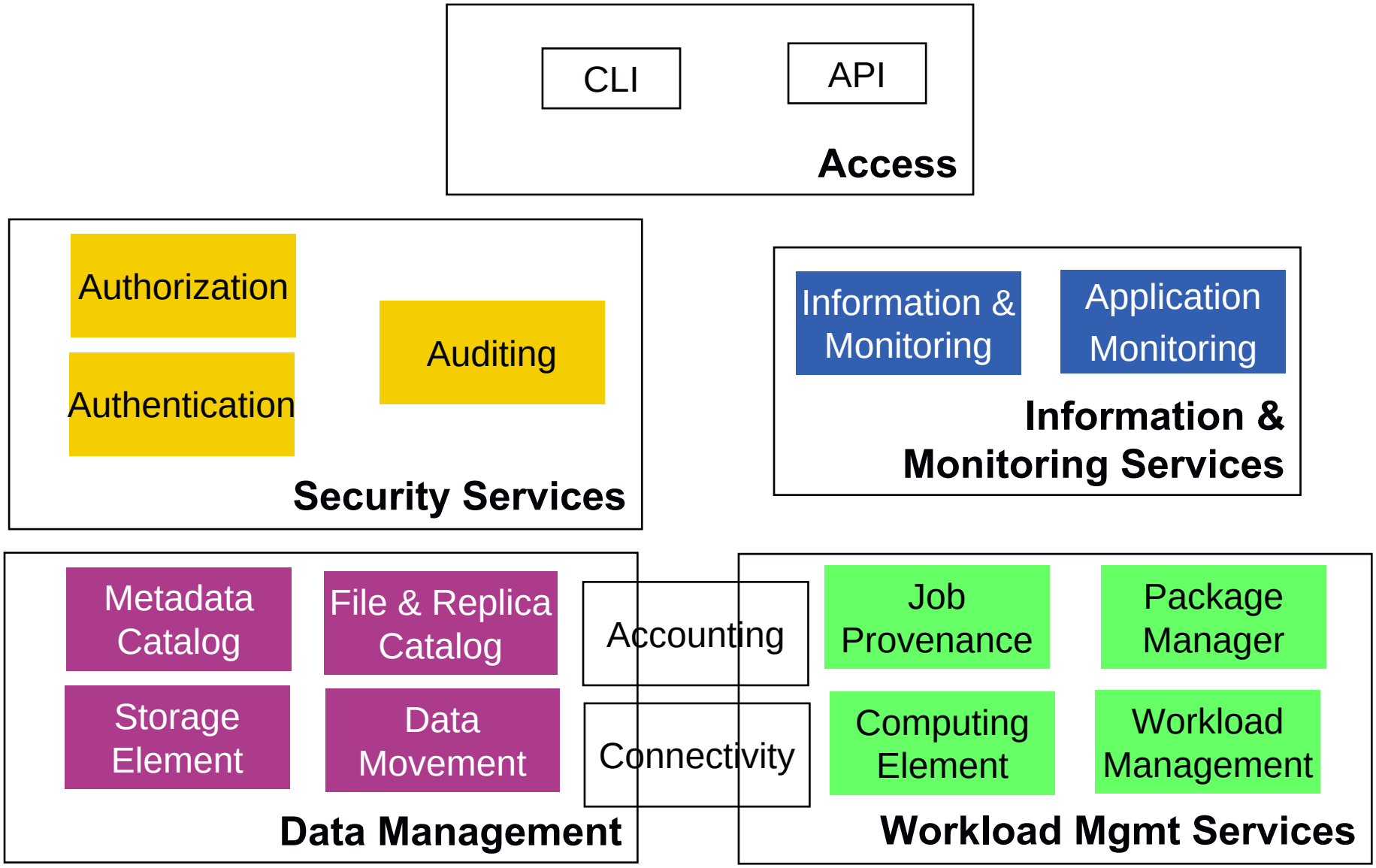
```
Executable = “gridTest”;
StdError = “stderr.log”;
StdOutput = “stdout.log”;
InputSandbox = {“/home/joda/test/gridTest”};
OutputSandbox = {“stderr.log”, “stdout.log”};
InputData = “lfn:/grid/gilda/training/testbed0-00019”;
DataAccessProtocol = “gridftp”;
Requirements = other.Architecture==“INTEL” && \
               other.OpSys==“LINUX”;
Rank = “other.GlueHostBenchmarkSF00”;
```

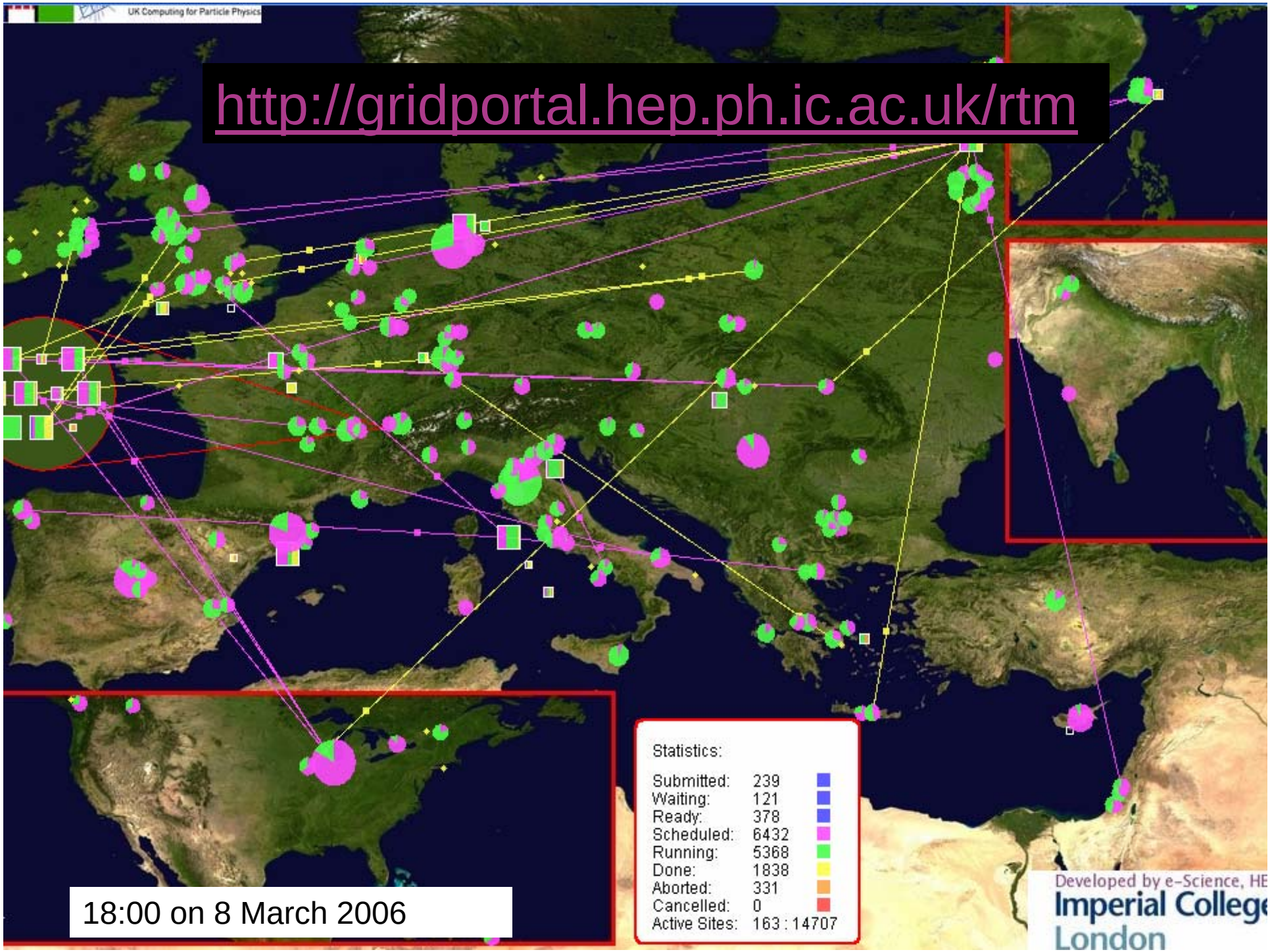
Who provides the resources?!

Service	Provider	Note
Access service	User / institute / VO	Computer with client software
Resource Broker (RB)	VO	
Information System:	Grid operations	
Computing Element (CE)	VOs - EGEE does not fund CEs	Scalability requires that VOs provide resources to match average need
Storage Element (SE)	VOs	

"VO": virtual organisation "Grid operations": funded effort

gLite Grid Middleware Services




UK Computing for Particle Physics
http://gridportal.hep.ph.ic.ac.uk/rtm
Statistics:
Submitted: 239
Waiting: 121
Ready: 378
Scheduled: 6432
Running: 5368
Done: 1838
Aborted: 331
Cancelled: 0
Active Sites: 163 : 14707
18:00 on 8 March 2006
Developed by e-Science, HE
Imperial College
London

What is happening now?

http://gridportal.hep.ph.ic.ac.uk/rtm

Further information

- **EGEE Conference: 25-29 September 2006**
  http://www.eu-egee.org/news/registration-open-for-egee201906-conference-September-2006-geneva/
- **EGEE digital library:** http://egee.lib.ed.ac.uk/
- EGEE www.eu-egee.org
- EGEE: 1st user Forum
  http://egee-intranet.web.cern.ch/egee-intranet/User-Forum

- gLite http://www.glite.org/

- Open Grid Forum http://www.gridforum.org/
- Globus Alliance http://www.globus.org/
- VDT http://www.cs.wisc.edu/vdt/

Further reading

- **Open Grid Forum** http://www.ggf.org/
- **The Grid Cafe** www.gridcafe.org
- **Grid Today** http://www.gridtoday.com/
- **Globus Alliance** http://www.globus.org/

Summary

- **EGEE is running the largest multi-VO grid in the world!**
- **Creating the "grid layer" in e-Infrastructure for research, public service and industry**
- **Key concepts for EGEE**
  - Sustainability – planning for the long-term
  - Production quality
  - And…
- **Grids are fundamentally about people**
- **… how people in different organisations commit to cooperate**
- **… and how that cooperation can be enabled by operations, training, support, and (most transient of all?) middleware**