Introduction to Grids and the EGEE project

EGEE-II's Programme for Industry

Ladislav Hluchý

Institute of Informatics, Slovak Academy of Sciences

Industry Day

Grid for Pharmaceutical & Biomedical Challenges

Bratislava, 19 September 2007

www.eu-egee.org

Information Society and Media

Lost in Definitions?

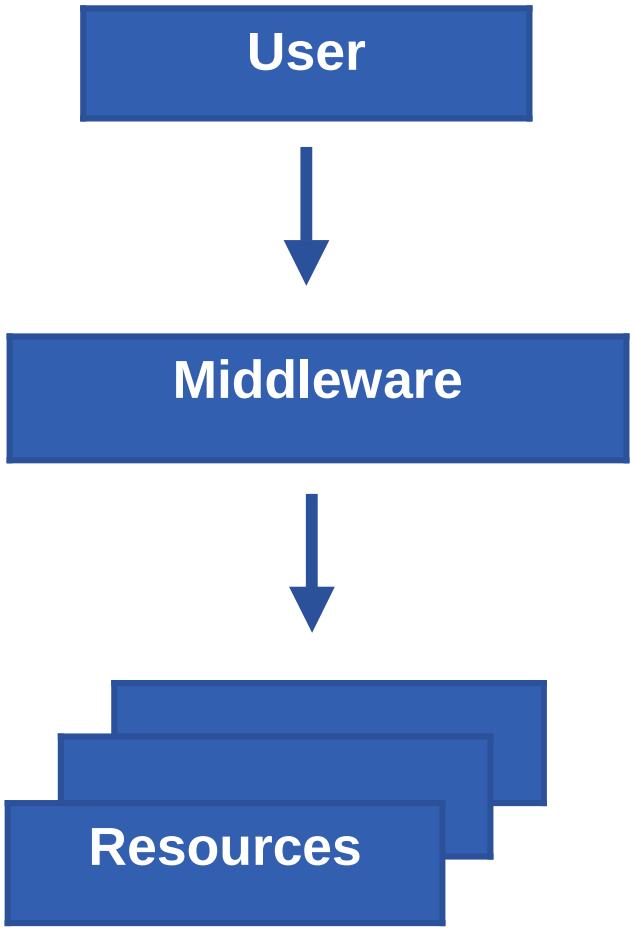
Defining the “Grid”:

- Access to (high performance) computing power
- Distributed parallel computing
- Improved resource utilization through resource sharing
- Increased memory provision
- Controlled access to distributed memory
- Interconnection of arbitrary resources (sensors, instruments, …)
- Collaboration between users/resources
- Higher abstraction layer above network services
- Corresponding security
- …

Defining the Grid

- A Grid is the combination of networked resources and the corresponding Grid middleware, which provides Grid services for the user.
- This interconnection of users, resources, and services for jointly addressing dedicated tasks is called a virtual organization.
- Comparison between Grids and Networks:
  - Networks realize message exchange between endpoints
  - Grids realize services for the users

  ➔ higher level of abstraction



Grids vs. Distributed Computing

- **Distributed applications already exist, but they tend to be *specialized systems* intended for a single purpose or user group**

- **Grids go further and take into account:**
  - Different kinds of *resources*
    - Not always the same hardware, data and applications
  - Different kinds of *interactions*
    - User groups or applications want to interact with Grids in different ways
  - *Dynamic* nature
    - Resources and users added/removed/changed frequently

Grid and Virtualisation

- **Virtual Organisations (VO's)= Group of users, federating resources**
  - Heterogeneous: people from different organisations
  - Cooperation: common goals
  - For sharing: to solve problems by using common resources
- **Virtualised shared computing and data resources**
  - Access to resources outside their institute for members of VO's
  - Resource providers negotiate with VO not with individual members
- **Virtualisation and sharing also possible for :**
  - Instruments, sensors, people, etc.

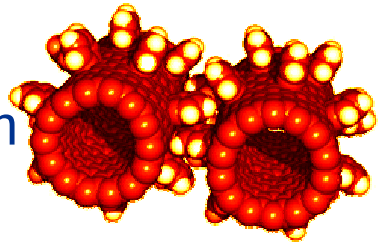
Virtualisation of resources is needed to hide their heterogeneity and present a simple interface to users

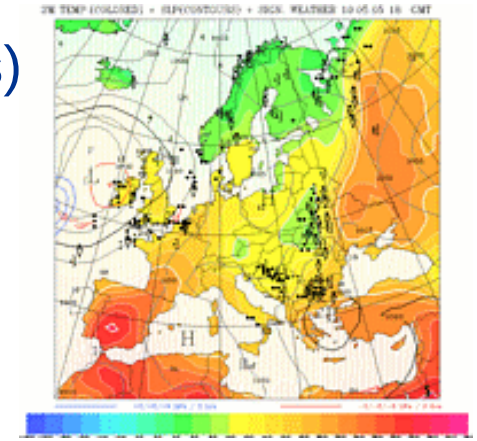
Defining the Grid

A Grid is the combination
of networked resources and the
corresponding Grid middleware, which
provides Grid services for the user.

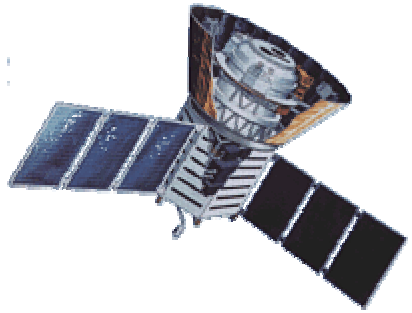
Why do we need a Grid?

- **Science is becoming increasingly digital and needs to deal with increasing amounts of data**
- **Simulations get ever more detailed**
  - e.g.Nanotechnology – design of new materials from the molecular scale
  - Modelling and predicting complex systems (weather forecasting, river floods, earthquakes)
  - Decoding the human genome





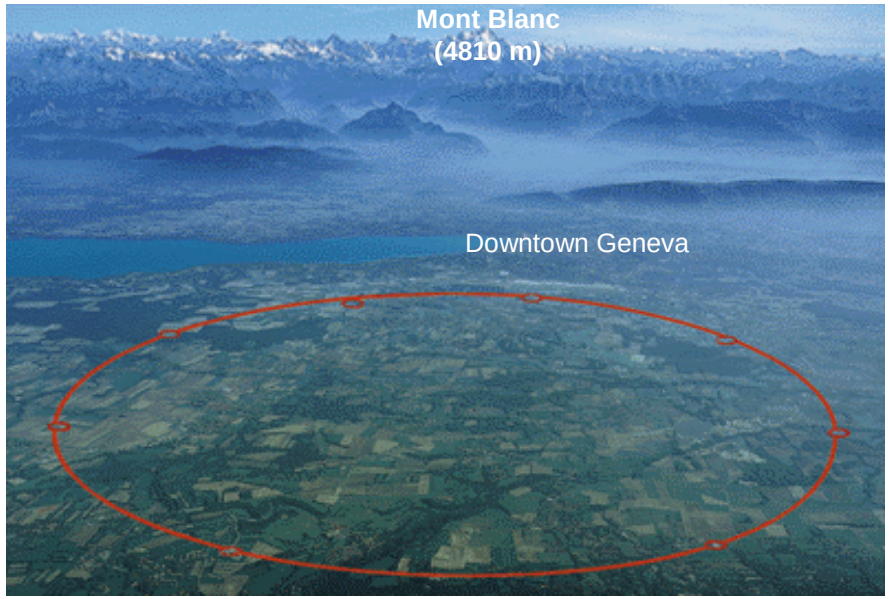
- **Experimental Science uses ever more sophisticated sensors to make precise measurements**
  - → Need high statistics
  - → Huge amounts of data
  - → Serves user communities around the world



The need for Grid in Particle Physics

- **CERN: the world's largest particle physics laboratory**
- **Particle physics requires special tools to create and study new particles: accelerators and detectors**

- **Large Hadron Collider (LHC):**
  - One of the most powerful instruments ever built to investigate matter
  - 4 experiments: ALICE, ATLAS, CMS, LHCb
  - 27 km circumference tunnel
  - Due to start up mid 2007



LHC Data

- 40 million collisions per second
- After filtering, 100 collisions of interest per second
- A Megabyte of data for each collision = recording rate of 0.1 Gigabytes/sec

- 10^{10} collisions recorded each year

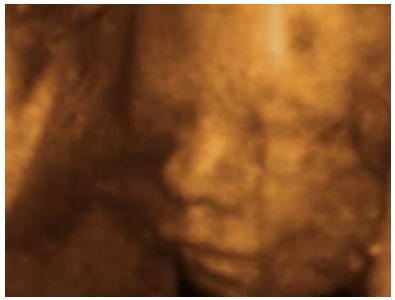
  ⇨ When LHC starts operation:

  will generate ~ 15 Petabytes/year of data*

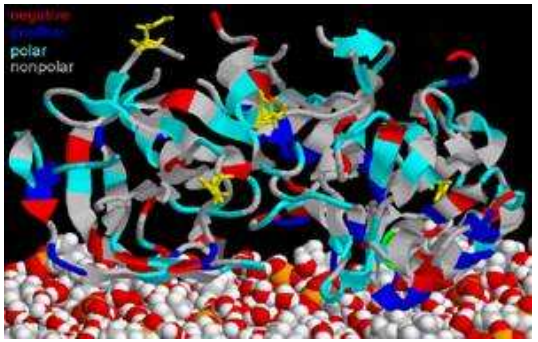
  *corresponding to more than 20 million CDs!

Biomedical applications

- **Biomedicine is also a pilot application area**
- **More than 20 applications deployed and being ported**



- **Three sub domains**
  - Medical image processing
  - Biomedicine
  - Drug discovery

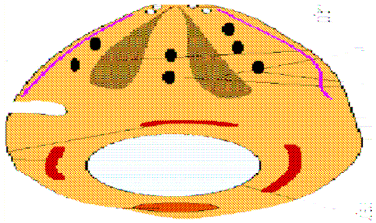


- **Use Grid as platform for collaboration**

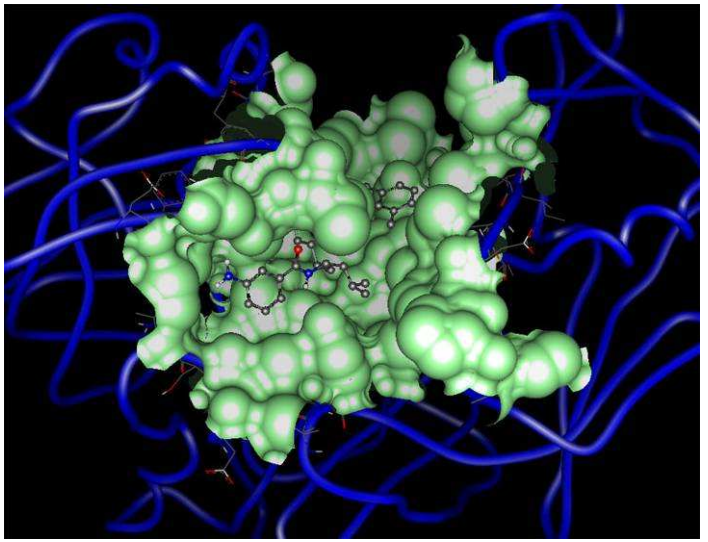
  (don’t need same massive processing power or storage as HEP)

Applications Example: WISDOM

- **Grid-enabled drug discovery process for neglected diseases**
  - *In silico* docking
    - compute probability that potential drugs dock with target protein
  - To speed up and reduce cost to develop new drugs



- **WISDOM (World-wide In Silico Docking On Malaria)**
  - First biomedical data challenge
  - 46 million ligands docked in 6 weeks
  - 1TB of data produced
  - 1000 computers in 15 countries
    - Equivalent to 80 CPU years
- **Second data challenge on Avian flu in April 2006**
  - 300,000 possible drug components tested
  - 8 different targets
  - 2000 computers used for 4 weeks



The EGEE project

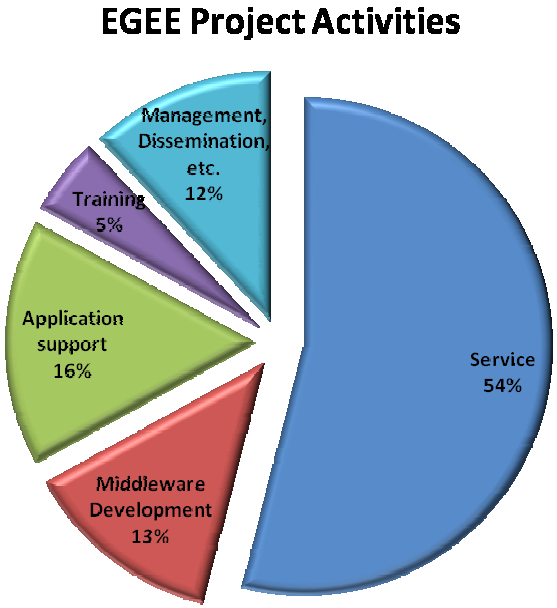
- **Flagship European grid infrastructure project, now in 2nd phase with 91 partners in 32 countries**

- **Objectives**
  - Large-scale, production-quality grid infrastructure for e-Science
  - Attracting new resources and users from industry as well as science
  - Maintain and further improve gLite Grid middleware

- **Structure**

  EGEE: 1 April 2004 – 31 March 2006

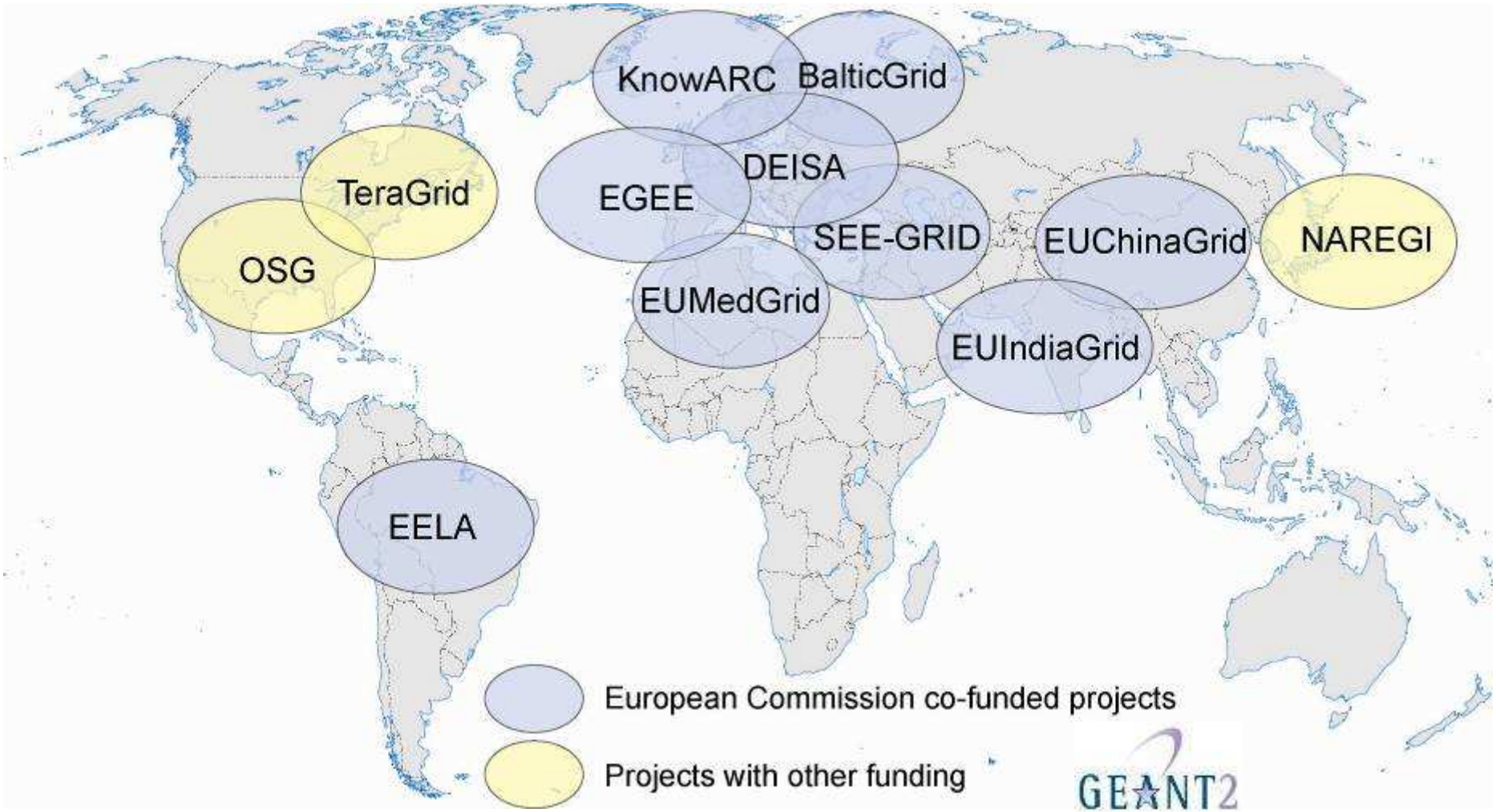
  EGEE-II: 1 April 2006 – 31 March 2008
  - Leveraging national and regional grid activities worldwide
  - Funded by the EC at a level of ~37 M Euros for 2 years
  - Support of related projects for infrastructure extension, application, specific services



Timeframe

From EGEE06: Projects present at the conference

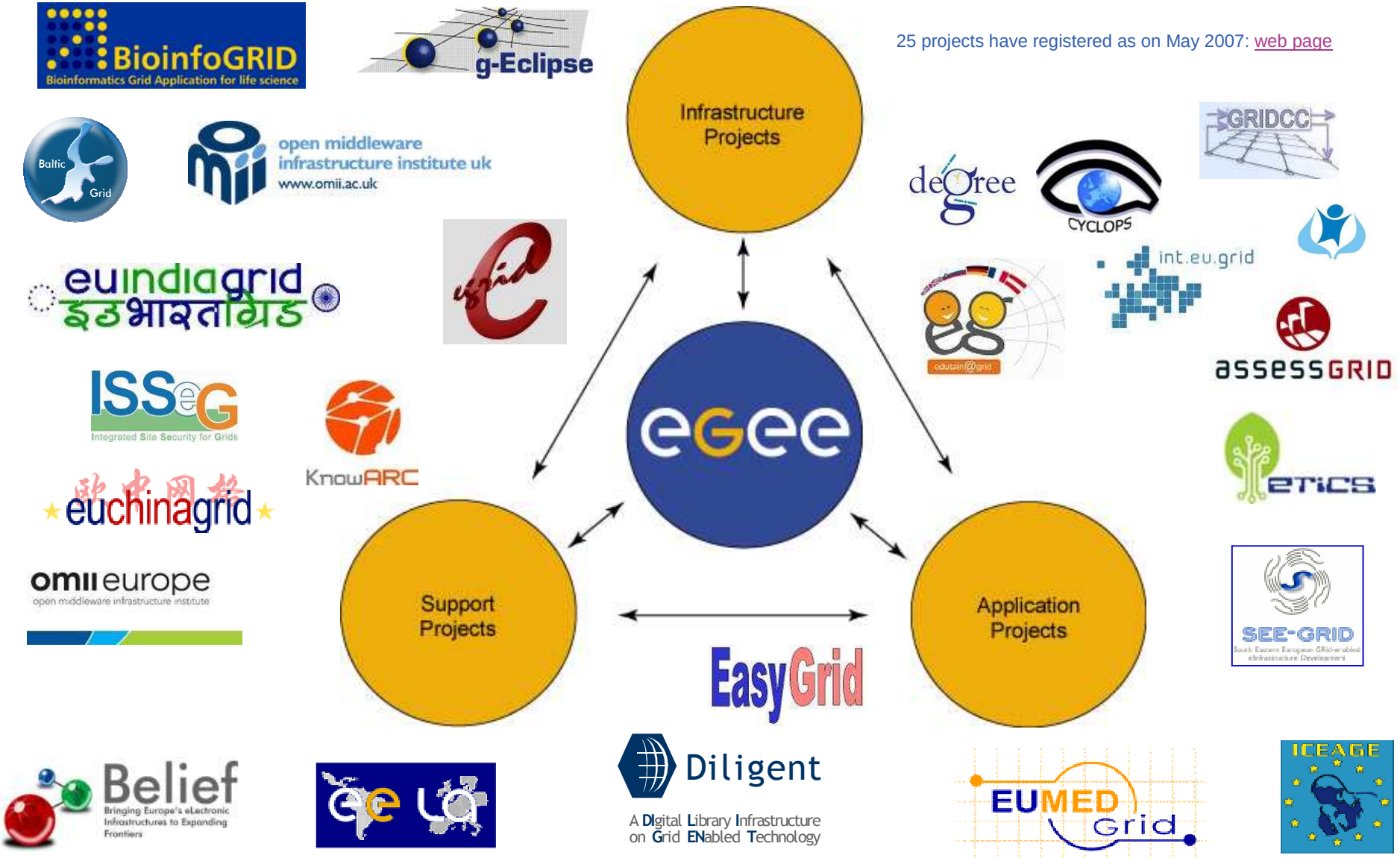
Collaborating e-Infrastructures



Potential for linking ~80 countries by 2008

Related projects

25 projects have registered as on May 2007: web page



Achievements

- **Infrastructure**
  - ~ 240 sites
  - > 36 000 CPUs
  - > 5 PB storage
  - 98k jobs/day
  - > 200 Virtual Organisations
- **Middleware**
  - Now at gLite release 3.0
    - Focus on basic services, easy installation and management
    - Industry friendly open source license
- **Many applications from a growing number of doma**
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics
  - Earth Sciences
  - Fusion
  - High-Energy Physics
  - Life Sciences

Grids in Europe

- **Great investment in developing Grid technology**
- **Sample of National Grid projects:**
  - Austrian Grid Initiative
  - Netherlands: DutchGrid
  - France: Grid'5000
  - Germany: D-Grid; Unicore
  - Greece: HellasGrid
  - Grid Ireland
  - Italy: INFNGrid; GRID.IT
  - NorduGrid
  - Swiss Grid
  - UK e-Science: National Grid Service; OMII; GridPP

- **EGEE provides a framework for national, regional and thematic Grids**

Grid projects at II SAS

- **MEDIGRID** - EU 6FP: Mediterranean Grid of Multi-Risk Data and Models (2004-2006)
- **NATO project**: Flood Forecasting on Grid Infrastructures (2004-2006)
- **K-Wf Grid** - EU 6FP: Knowledge-based Workflow System for Grid Applications (2004-2007)
- **DEGREE** - EU 6FP: Dissemination and Exploitation GRids in Earth sciencE (2006-2008)
- **int.eu.grid** - EU 6FP: Interactive European Grid (2006-2008)
- **EGEE** - EU 6FP: Enabling Grids for E-sciencE (2004-2006)
- **EGEE-II** - EU 6FP: Enabling Grids for E-sciencE II (2006-2008)

Grid computing in Slovakia

Národné gridové infraštruktúry v strednej Európe spájajú národné a regionálne gridové infraštruktúry a vytvárajú silný a bezpečný grid.

240 zúčastnených miest (site) zo 41 krajín s viac ako 36.000 CPU je organizovaných do 12 federácií.

Stredoeurópsku federáciu tvoria: Slovensko, Poľsko, Česko, Rakúsko, Maďarsko, Slovinsko a Chorvátsko.

Väčšina krajín našej federácie má národnú gridovú iniciatívu (ako projekt financovaný na národnej úrovni alebo ako centrá, ktoré koordinujú národné gridové aktivity).

National grid infrastructures in central Europe aim to join national and regional grid infrastructures and create robust and secure grid available to scientists.

240 participating sites from 41 countries with more than 36.000 CPUs are organized to 12 federations.

Slovakia with Poland, Czech Republic, Austria, Hungary, Slovenia and Croatia form the **Central European federation**.

Most of the countries in our federation have **national grid initiatives** either as projects funded on national level or as centres coordinating national grid activities.

Grid computing in Slovakia

Slovensko nemá národný (financovaný) program na podporu rozvoja gridových technológií a budovanie národnej gridovej infraštruktúry.

V súčasnosti je do EÚ gridovej infraštruktúry zapojených 5 miest (site) zo Slovenska.

Ústav informatiky SAV poskytuje 64 CPU do niekoľkých gridových infraštruktúr.

ÚI SAV prevádzkuje sajt (site) a poskytuje používateľskú podporu ako aj činnosť Slovenskej Gridovej Certifikačnej Autority a vydávanie digitálnych gridových certifikátov pre používateľov a hostiteľské počítače (host) zo Slovenska.

Unlike other countries in central Europe, Slovakia has no national funding program for supporting the development of grid technologies and building national grid infrastructure.

Currently there are five grid sites included in EU grid infrastructures.

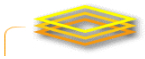
Institute of Informatics SAS provides 64 CPUs in multiple grid infrastructures.

II SAS also provides site and user support including operation of

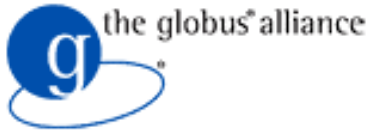
Slovak Grid Certification Authority issuing digital grid certificates for users and hosts from Slovakia.

Projects Worldwide

- **Infrastructure projects**
  - OSG, Teragrid (US)
  - Naregi (Japan)
  - APAC (Australia)
  - and many more
  - …



- **Middleware projects**
  - Condor
  - Globus
  - Legion
  - and many more
  - …



→ **Collaboration with EGEE**

Standards are key

- **Need standards for the Grid to:**
  - Build confidence
  - Facilitate interoperability
  - Required for Business use

- **EGEE contributes to standards**
  - In OGF: contributes to 15 WGs and RGs, provides 2 Area Directors
  - Also work with IETF (Internet Engineering Task Force), OASIS (Organisation for the Advancement of Structured Information Standards) , e-IRG (e-Infrastructure Reflection Group) on standards
  - Common work with OSG, NAREGI, NORDUGRID/ARC, GIN (Grid Interoperation Now)

Summary

- **Grids represent a powerful new tool for science**

→Today we have a **window of opportunity** to move Grids from research prototypes to production systems (as networks did a few years ago)

- **EGEE offers:**
  - A mechanism for **linking** together **people**, **resources** and **data** for many scientific communities
  - A basic set of middleware for gridifying applications, together with documentation, training and support
  - Regular forums to discuss with Grid experts, other communities and industry