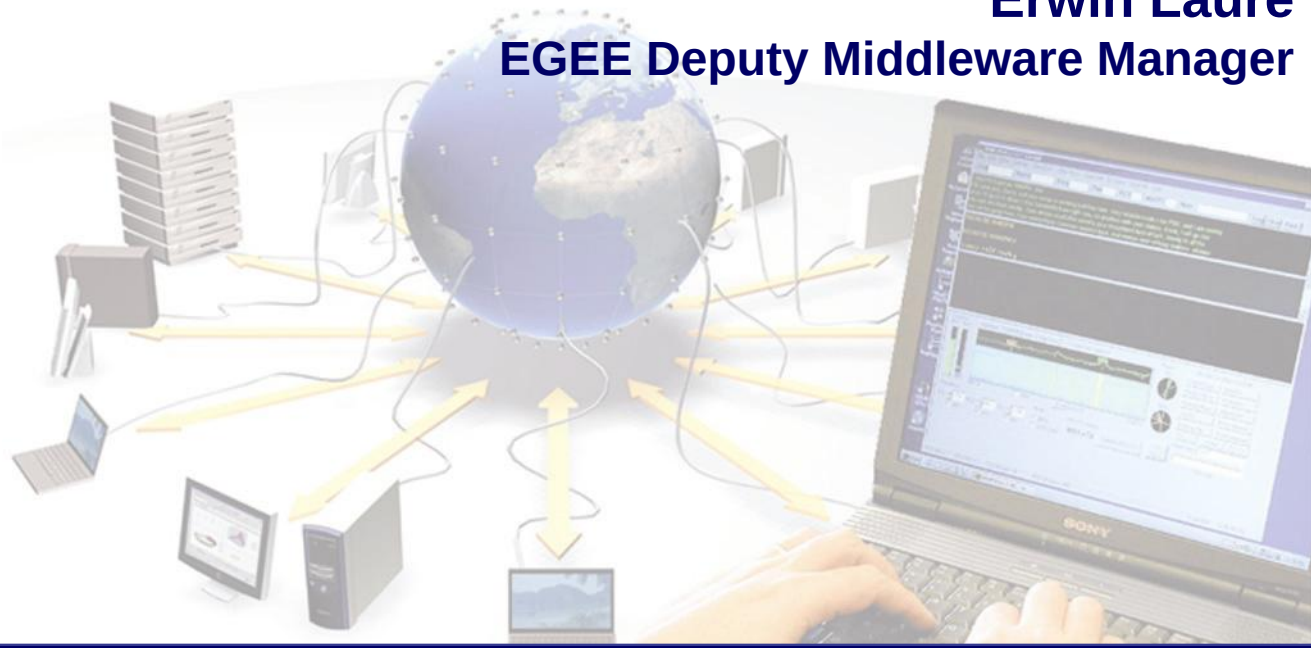
Enabling Grids for E-science in Europe

www.eu-egee.org

Cern School of Computing, Aug 30 – Sep 10 2004

Introduction to Grid Computing and the EGEE Project

Erwin Laure
EGEE Deputy Middleware Manager

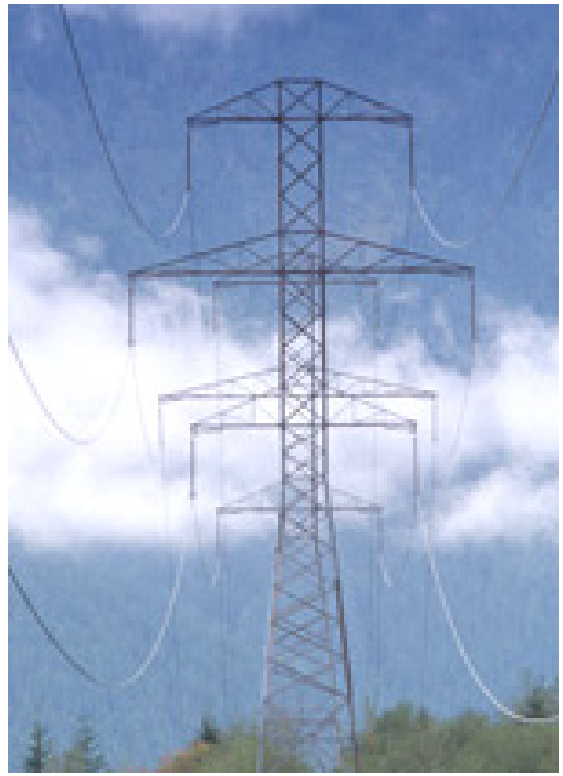


EGEE is a project funded by the European Union under contract IST-2003-508833

What is the Grid?

Enabling Grids for E-science in Europe

- The World Wide Web provides seamless access to information that is stored in many millions of different geographical locations
- In contrast, the Grid is a new computing infrastructure which provides seamless access to computing power and data distributed over the globe
- The name Grid is chosen by analogy with the electric power grid: plug-in to computing power without worrying where it comes from, like a toaster

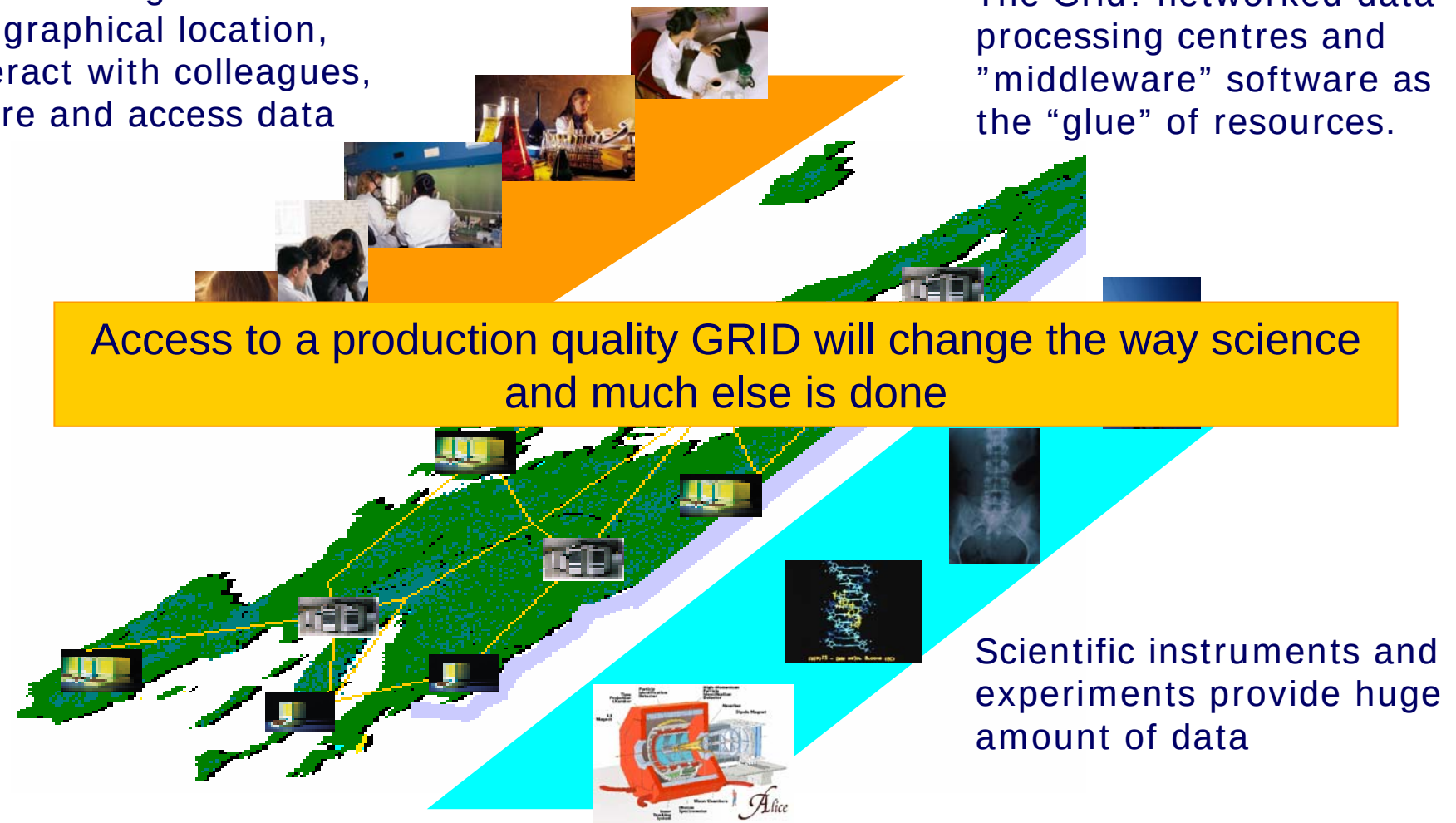


The Grid Vision

Enabling Grids for E-science in Europe

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.



Access to a production quality GRID will change the way science and much else is done

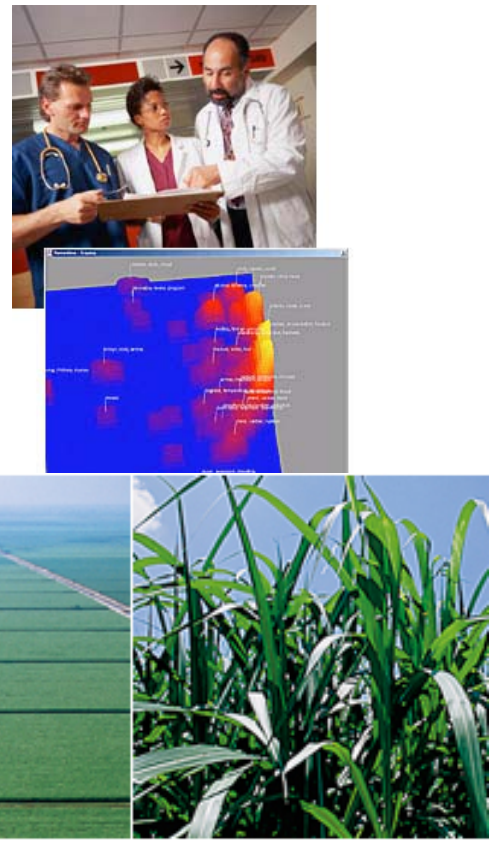
Scientific instruments and experiments provide huge amount of data

What is driving grid development?

Enabling Grids for E-science in Europe

Data and compute intensive sciences are next generation applications that have extreme needs but are likely to become mainstream in the next 5 years

- Physics/Astronomy: data from different kinds of research instruments
- Medical/Healthcare: imaging, diagnosis and treatment
- Bioinformatics: study of the human genome and proteome to understand genetic diseases
- Nanotechnology: design of new materials from the molecular scale
- Engineering: design optimization, simulation, failure analysis and remote Instrument access and control
- Natural Resources and the Environment: weather forecasting, earth observation, modeling and prediction of complex systems: river floods and earthquake simulation



How does the grid work?

Enabling Grids for E-science in Europe

- The Grid relies on advanced software, called middleware, which ensures seamless communication between different computers and different parts of the world
- The Grid search engine not only finds the data the scientist needs, but also the data processing techniques and the computing power to carry them out
- It distributes the computing task to wherever in the world there is available capacity, and sends the result back to the scientist

Grids vs. Distributed Computing

Enabling Grids for E-science in Europe

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

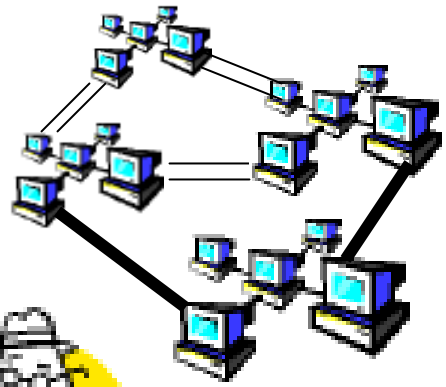
- Grids go further and take into account:
  - Different kinds of *resources*
    - Not always the same hardware, data and applications
  - Different kinds of *interactions*
    - User groups or applications want to interact with Grids in different ways
  - *Dynamic* nature
    - Resources and users added/removed/changed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

What are the characteristics of a Grid system?

Enabling Grids for E-science in Europe

Numerous Resources

Ownership by Mutually Distrustful Organizations & Individuals



Connected by Heterogeneous, Multi-Level Networks

Different Security Requirements & Policies Required

Different Resource Management Policies

Potentially Faulty Resources

Geographically Separated

Resources are Heterogeneous

Grid projects

Enabling Grids for E-science in Europe

Many Grid development efforts — all over the world

- NASA Information Power Grid
- DOE Science Grid
- NSF National Virtual Observatory
- NSF GriPhyN
- DOE Particle Physics Data Grid
- NSF TeraGrid
- DOE ASCI Grid
- DOE Earth Science Grid
- DARPA CoABS Grid
- NEESGrid
- DOH BIRN
- NSF iVDGL

- EGEE (CERN, …)
- DataGrid (CERN, ...)
- EuroGrid (Unicore)
- DataTag (CERN,…)
- Astrophysical Virtual Observatory
- GRIP (Globus/Unicore)
- GRIA (Industrial applications)
- GridLab (Cactus Toolkit)
- CrossGrid (Infrastructure Components)
- EGSO (Solar Physics)

- UK – OGSA-DAI, RealityGrid, GeoDise, Comb-e-Chem, DiscoveryNet, DAME, AstroGrid, GridPP, MyGrid, GOLD, eDiamond, Integrative Biology, …
- Netherlands – VLAM, PolderGrid
- Germany – UNICORE, Grid proposal
- France – Grid funding approved
- Italy – INFN Grid
- Eire – Grid proposals
- Switzerland - Network/Grid proposal
- Hungary – DemoGrid, Grid proposal
- Norway, Sweden - NorduGrid

Standards are the key

Enabling Grids for E-science in Europe

- The systems developed and deployed by the plethora of Grid projects need eventually converge
- Standardization efforts at multiple levels:

- Basic Infrastructure:
  - OASIS (http://www.oasis-open.org/)
    Web Services, XML, WSRF

- Higher level services
  - GGF (http://www.ggf.org/)
    Open Grid Service Architecture (OGSA)
    51 Research and Working groups organized in 7 areas

What do we expect?

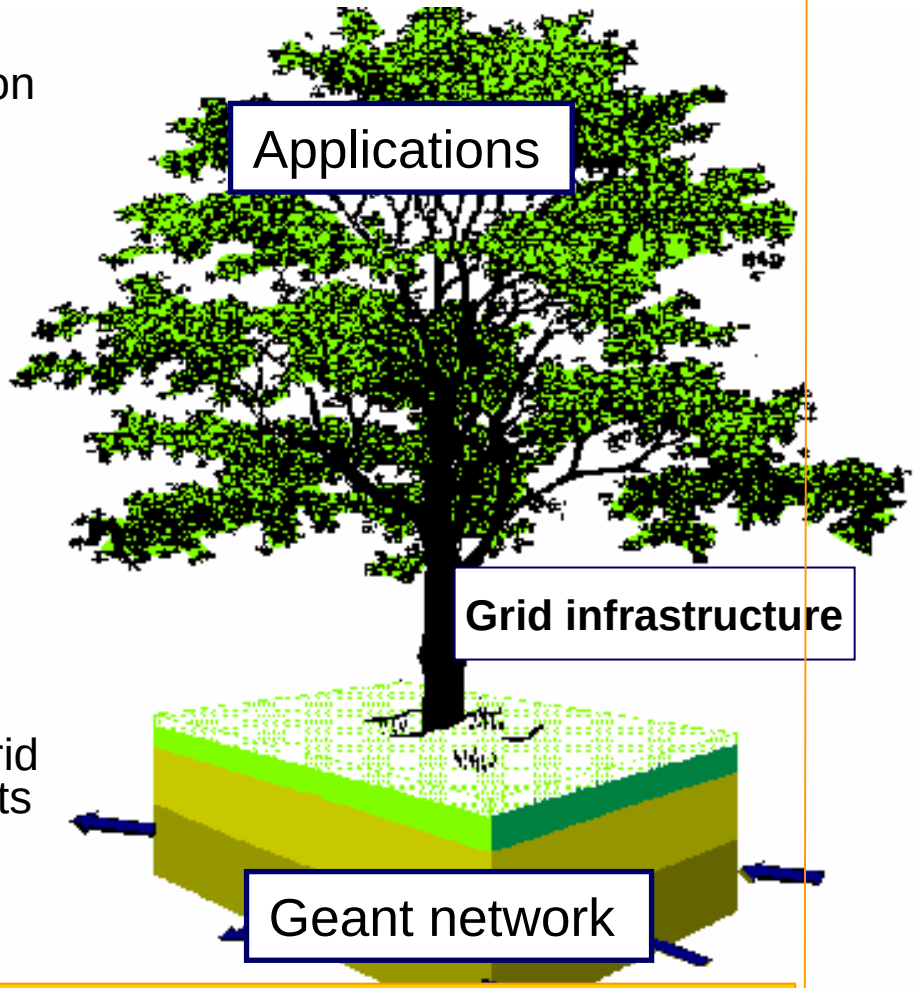
Enabling Grids for E-science in Europe

- The Grid will provide:
  - Access to a world-wide virtual computing laboratory with almost infinite resources
  - Possibility to organize distributed scientific communities in Virtual Organizations (VOs)
  - Transparent access to distributed data and easy workload management
  - Easy to use application interfaces

Introduction to EGEE

Enabling Grids for E-science in Europe

- **Goal**
  - Create a wide European Grid production quality infrastructure on top of present and future EU RN infrastructure
- **Build On**
  - EU and EU member states major investments in Grid Technology
  - International connections (US and AP)
  - Several pioneering prototype results
  - Large Grid development teams in EU require major EU funding effort
- **Approach**
  - Leverage current and planned national and regional Grid programmes
  - Work closely with relevant industrial Grid developers, NRENs and US-AP projects



Despite its name EGEE is an International project involving in particular Israel, Russia and US

What is EGEE?

Enabling Grids for E-science in Europe

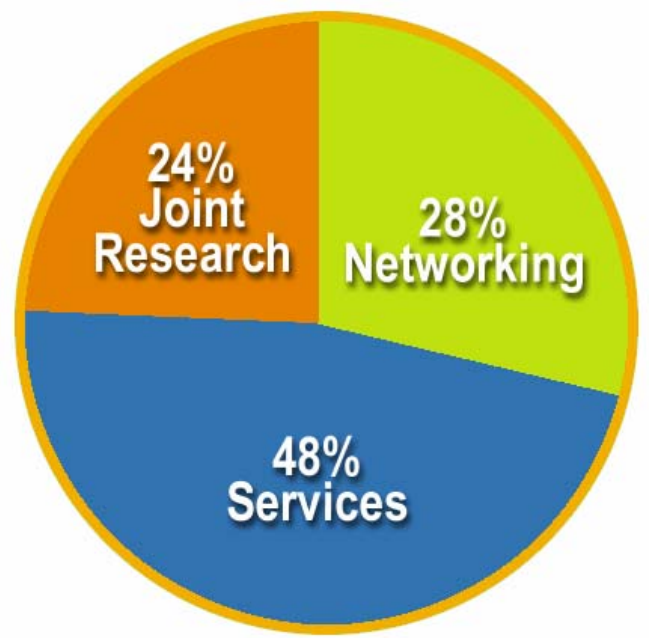
- 70 leading institutions in 27 countries, federated in regional Grids
- 32 M Euros EU funding (2004-5), O(100 M) total budget
- Aiming for a combined capacity of over 20’000 CPUs (the largest international Grid infrastructure ever assembled)
- ~ 300 dedicated staff

EGEE Activities

Enabling Grids for E-science in Europe

- Emphasis on operating a production grid and supporting the end-users

- **48 % service activities** (Grid Operations, Support and Management, Network Resource Provision)
- **24 % middleware re-engineering** (Quality Assurance, Security, Network Services Development)
- **28 % networking** (Management, Dissemination and Outreach, User Training and Education, Application Identification and Support, Policy and International Cooperation)



EGEE and LCG

- **EGEE builds on the work of LCG to establish a grid operations service**

- **LCG (LHC Computing Grid) - Building and operating the LHC Grid**
- A collaboration between:
  - The physicists and computing specialists from the LHC experiment
  - The projects in Europe and the US that have been developing Grid middleware
  - The regional and national computing centres that provide resources for LHC
  - The research networks

- **Mission:**
  - Prepare and deploy the computing environment that will be used by the experiments to analyse the LHC data
  - Started September 2001

- **Strategy:**
  - Integrate thousands of computers at dozens of participating institutes worldwide into a global computing resource
  - Rely on software being developed in advanced grid technology projects, both in Europe and in the USA (EDG, VDT, others)

EGEE infrastructure

- Access to networking services provided by **GEANT** and the **NRENs**

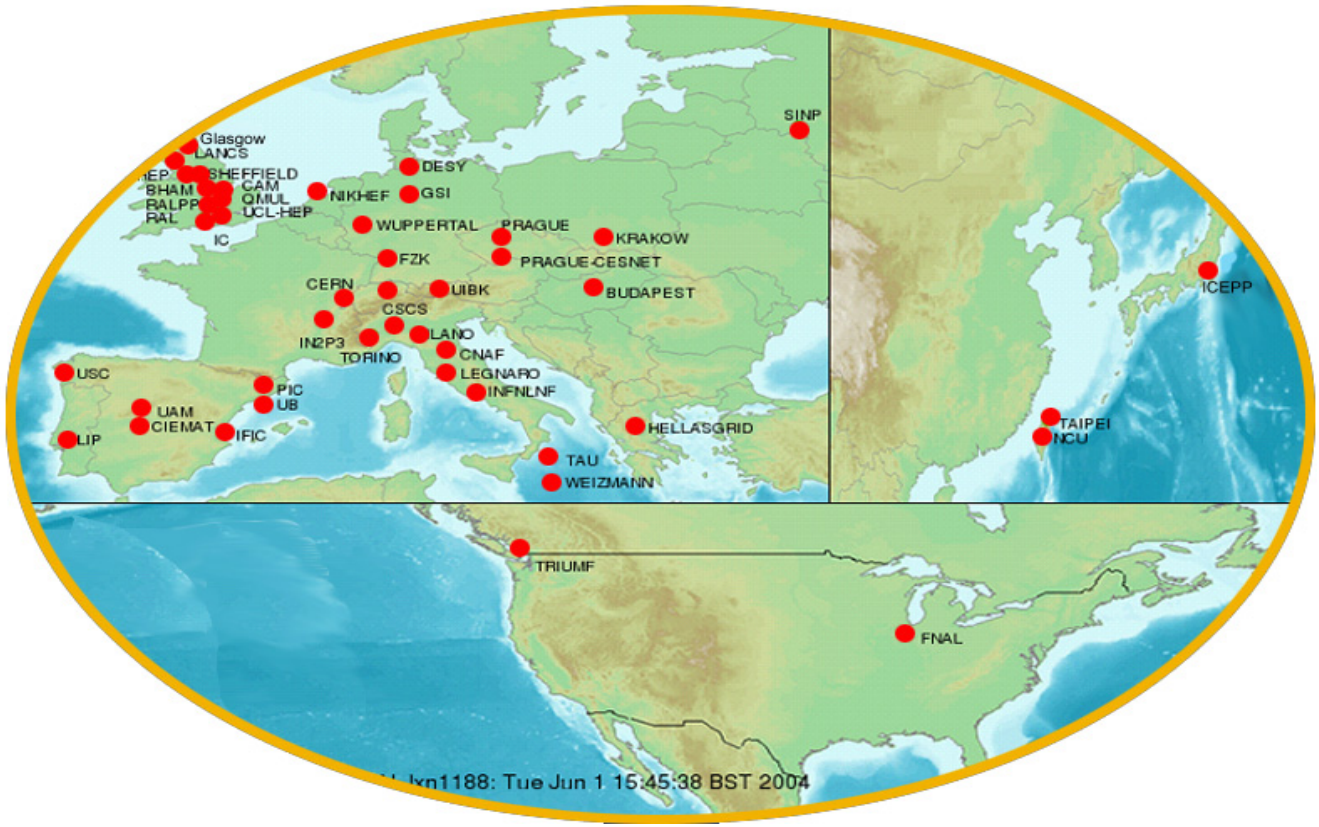
- Production Service:
  - in place (based on HEP LCG-2)
  - for production applications
  - MUST run reliably, runs only proven stable, debugged middleware and services
  - Will continue adding new sites in EGEE federations

- Pre-production Service:
  - For middleware re-engineering

- Certification and Training/Demo testbeds

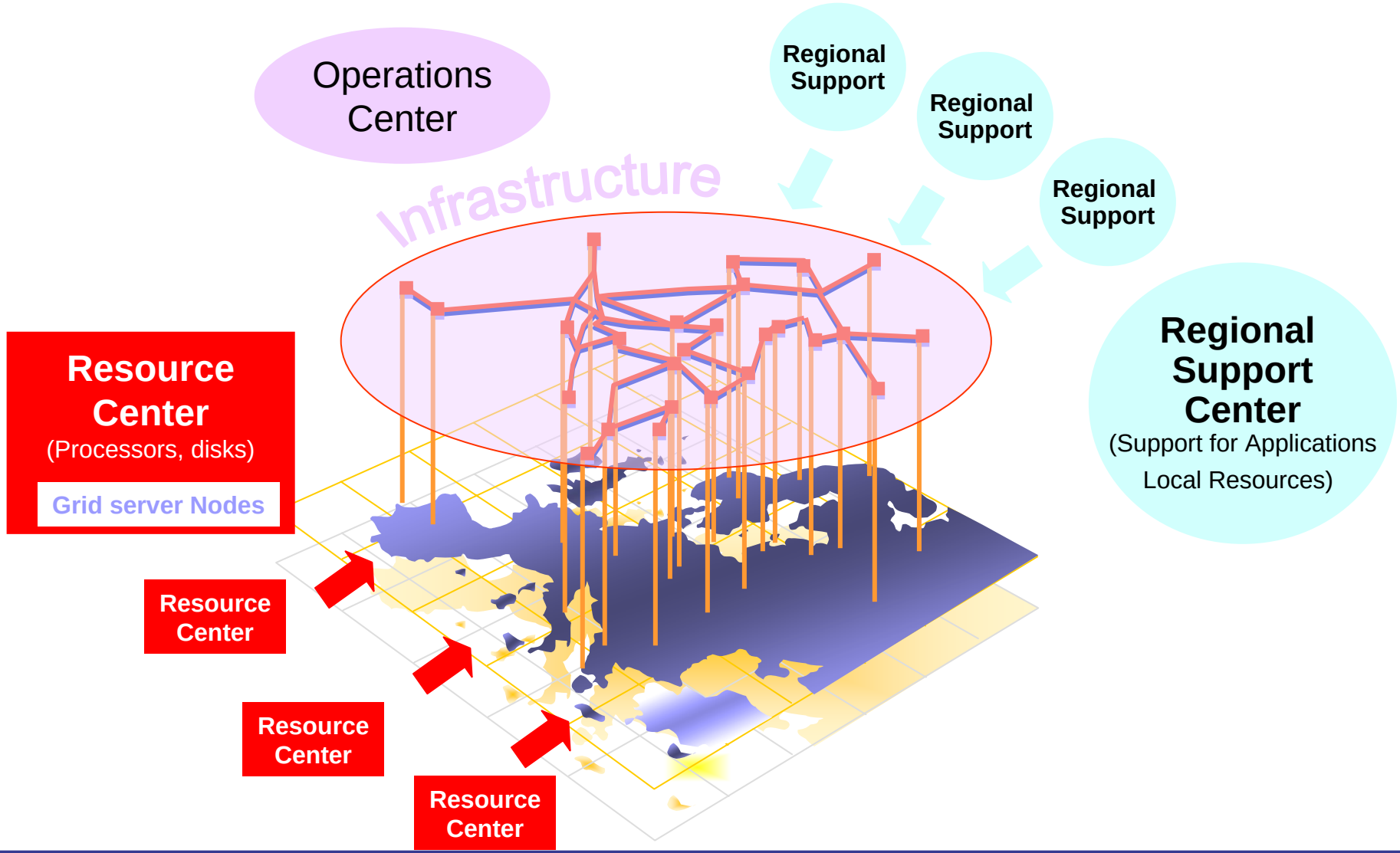
LCG-2/EGEE-0 (I)

- Based on HEP-LCG testbed: more than 60 sites worldwide (+ few non-HEP); over 6.000 CPUs



EGEE Operations

Enabling Grids for E-science in Europe



EGEE Operations (I): OMC and CIC

Enabling Grids for E-science in Europe

- Operation Management Centre
  - located at CERN, coordinates operations and management
  - coordinates with other grid projects
- Core Infrastructure Centres
  - behave as single organisations
  - operate core services (VO specific and general Grid services)
  - develop new management tools
  - provide support to the Regional Operations Centres

EGEE Operations (II): ROC

Enabling Grids for E-science in Europe

- Regional Operations Centre responsibilities and roles:
  - Testing (certification) of new middleware on a variety of platforms before deployment
  - Deployment of middleware releases + coordination + distribution inside the region
  - integration of ‘Local’ VO
  - Development of procedures and capabilities to operate the resources
  - First-line user support
  - Bring new resources into the infrastructure and support their operation
  - Coordination of integration of national grid infrastructures Provide resources for pre-production service

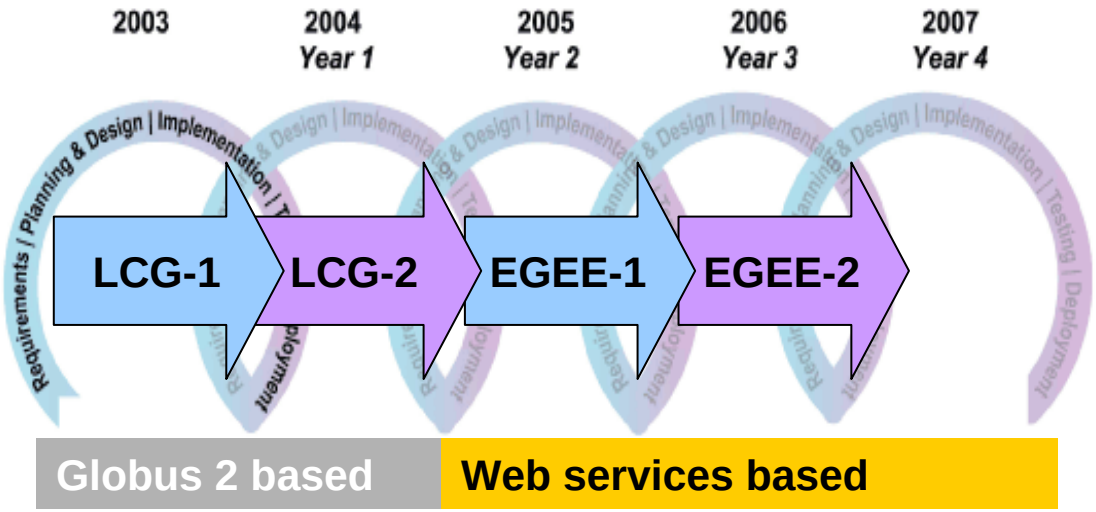
Deployment Issues

Enabling Grids for E-science in Europe

- **Need to expand on existing LCG service while maintaining stability**
  - Add more sites/resources (some have no previous experience with grids)
    - Experience has shown that this can be effort consuming
    - Problematic sites have been causing problems for the whole system
  - Introduce applications and VOs from non-HEP (Bio-medical)
    - Need to clarify processes and information flow
- **Portability**
  - Support for further platforms (currently just RedHat 7.3)
  - Middleware dependencies and packaging
- Middleware Support
  - Deterministic Support Model has been formalized
  - Essential to have (so far excellent) VDT support for Condor/Globus
- **“24x7” operational support**
  - Currently have GOC at RAL http://goc.grid-support.ac.uk/
  - Being replicated at Taipei (and maybe Canada?)
  - Prototype accounting system (based on R-GMA) ready for the release in April 2004 (testing, documentation and packaging done)

EGEE Implementation

- **From day 1 (1st April 2004)**
  - Production grid service based on the LCG infrastructure running LCG-2 grid middleware (SA)
  - LCG-2 will be maintained until the new generation has proven itself (fallback solution)
- **In parallel develop a "next generation" grid facility**
  - Produce a new set of grid services according to evolving standards (Web Services)
  - Run a development service providing early access for evaluation purposes
  - Will incrementally replace LCG-2 on production facility in 2005

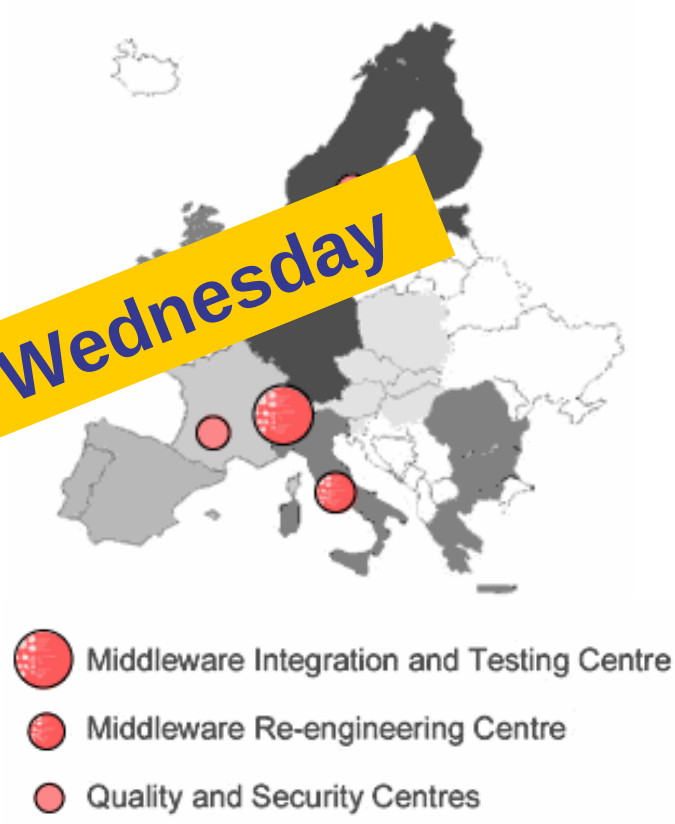


EGEE Middleware Activity

Enabling Grids for E-science in Europe

- Middleware selected based on requirements of Applications and Operations
- Harden and re-engineer existing middleware functionality, leveraging experience of partners
- Provide robust, [illegible] components
- Support components evolution towards a service-oriented approach (Web Services)

More details will be given on Wednesday



Exploit established standards where possible
Contribute to standardization efforts (e.g. GGF)

EGEE Middleware: gLite

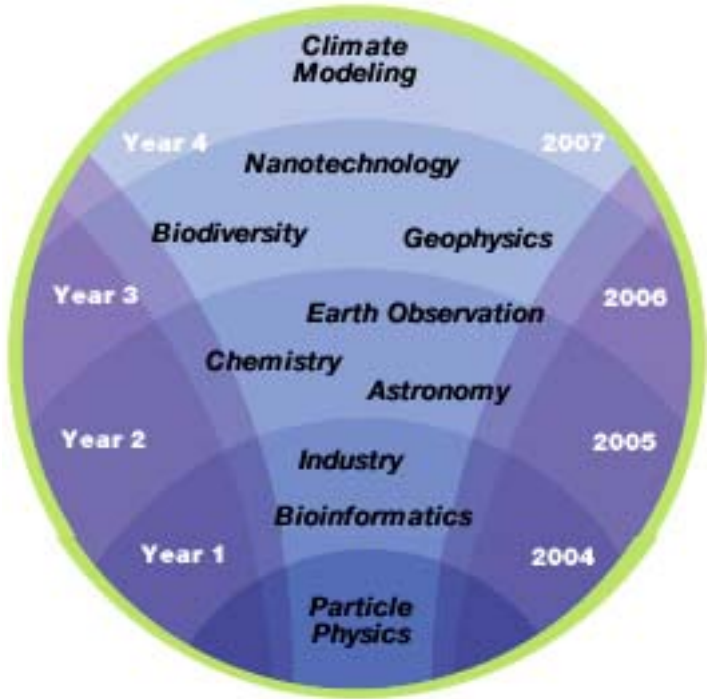
Enabling Grids for E-science in Europe

- **gLite**
  - Exploit experience and existing components from VDT (CondorG, Globus), EDG/LCG, AliEn, and others
  - Develop a lightweight stack of generic middleware useful to EGEE applications (HEP and Biomedics are pilot applications).
    - Should eventually deploy dynamically (e.g. as a globus job)
    - Pluggable components – cater for different implementations
  - Focus is on re-engineering and hardening
  - Early prototype and fast feedback turnaround envisaged

EGEE Applications

Enabling Grids for E-science in Europe

- EGEE Scope : ALL-Inclusive for academic applications (open to industrial and socio-economic world as well)
- The major success criterion of EGEE: how many satisfied users from how many different domains ?
- 5000 users (3000 after year 2) from at least 5 disciplines
- Two pilot applications selected to guide the implementation and certify the performance and functionality of the evolving infrastructure: Physics & Bioinformatics



Application domains and timelines are for illustration only

EGEE pilot application: HEP

Enabling Grids for E-science in Europe

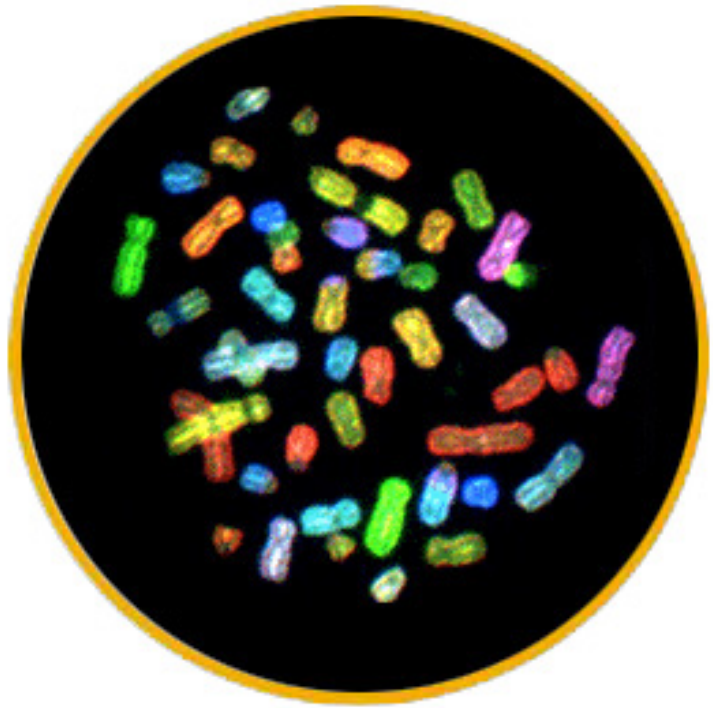
- **HEP**
  - Running large distributed computing systems for many years
  - Focus for the future is on computing for LHC (LCG)
  - The 4 LHC experiments and other current HEP experiments use grid technology e.g. Babar,CDF,D0, ..
  - LHC experiments are currently executing large scale data challenges (DCs) involving thousands of processors world-wide and generating many Terabytes of data

  - Moving to so-called ‘chaotic’ use of grid with individual user analysis (thousands of users interactively operating within experiment VOs)

EGEE pilot application: Biomedics

Enabling Grids for E-science in Europe

- Biomedics
  - Bioinformatics (gene/proteome databases distributions)
  - Medical applications (screening, epidemiology, image databases distribution, etc.)
  - Interactive application (human supervision or simulation)
  - Security/privacy constraints
    - Heterogeneous data formats,
      Frequent data updates
      Complex data sets
      Long term archiving



- BioMed applications deployed and expect to run first job on LCG-2 by September

Generic Application Support

Enabling Grids for E-science in Europe

- Getting new scientific and industrial communities interested and committed to use the grid infrastructure built by EGEE is key to the success of the project

- Questionnaire to get information and first requirem… new communities interested in using the EGEE In… e (http://alipc1.ct.infn.it/grid/egee/na4/qu…/na4-genapp-questionnaire.doc)
- Feed-backs received so fa… (http://alipc1.ct.infn.it/…na4/questionnaire):
  - Astrophysic… Planck satellite)
  - Earth… on (ozone maps, seismology, climate)
  - …raries (DILIGENT Project)
  - …d Search Engines (GRACE Project)
  - Industrial applications (SIMDAT Project)

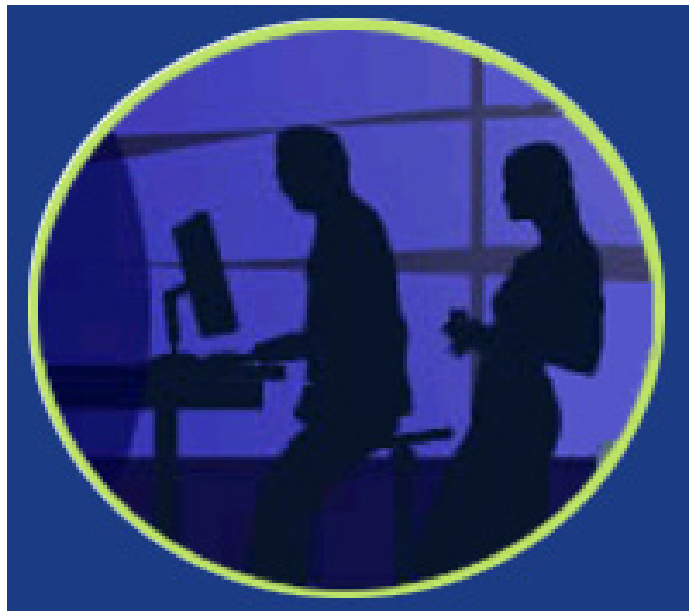
More details will be given on Wednesday

- Interest also from Computational Chemistry (Italy and Czech Republic), Civil Engineering (Spain), and Geophysics (Switzerland and France) communities

User training and induction

Enabling Grids for E-science in Europe

- Training material and courses from introductory to advanced level
- Train a wide variety of users both internal to the EGEE consortium and from external groups from across Europe
- 7 courses/presentations already held and 5 more planned through July
- Experience with GENIUS portal and GILDA testbed (provided by INFN)
- Courses inline with the needs of the projects and applications



Dissemination

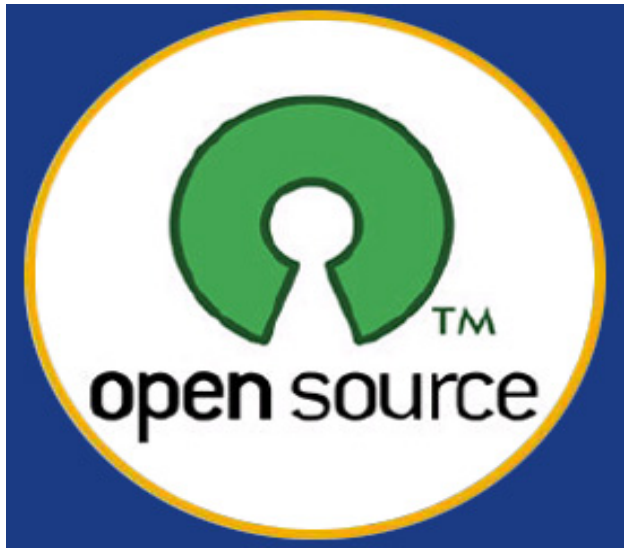
- 1st project conference
  - Over 300 delegates came to the 4 day event during April in Cork Ireland
  - Kick-off meeting bringing together representatives from the 70 partner organisations

- Websites, Brochures and press releases
  - For project and general public **www.eu-egee.org**
  - Information packs for the general public, press and industry

Security & Intellectual Property

Enabling Grids for E-science in Europe

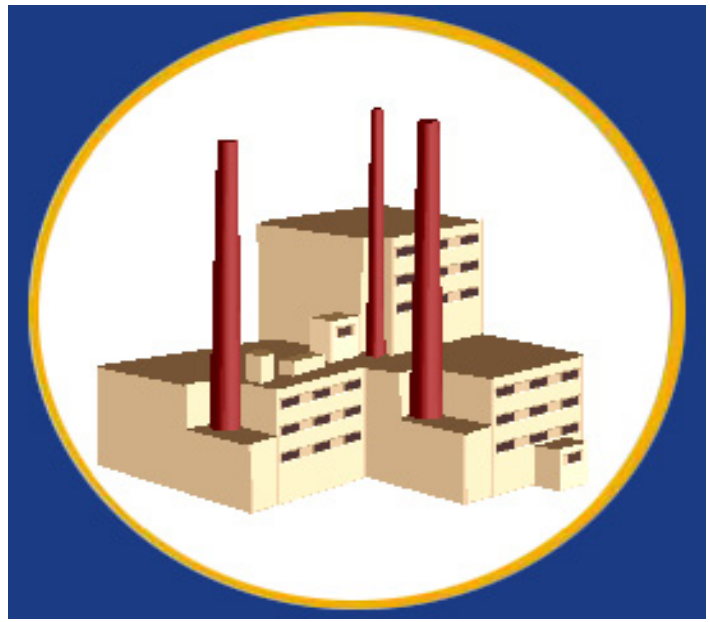
- The existing EGEE grid middleware is distributed under an Open Source License developed by EU DataGrid
  - No restriction on usage (scientific or commercial) beyond acknowledgement
  - Same approach for new middleware
- Application software maintains its own licensing scheme
  - Sites must obtain appropriate licenses before installation



EGEE Industry Forum

Enabling Grids for E-science in Europe

- EGEE Industry Forum
  - raise awareness of the project in industry to encourage industrial participation in the project
  - foster direct contact of the project partners with industry
  - ensure that the project can benefit from practical experience of industrial applications



- For more info:

  www.eu-egee.org

Expected Developments in 2004

Enabling Grids for E-science in Europe

- **General:**
  - LCG-2 will be the service run in 2004 – aim to evolve incrementally
  - Goal is to run a stable service
- **Some functional improvements:**
  - Extend access to MSS – tape systems, and managed disk pools
  - Distributed vs replicated replica catalogs
    - To avoid reliance on single service instances
- **Operational improvements:**
  - Monitoring systems – move towards proactive problem finding, ability to take sites on/offline; experiment monitoring
  - Continual effort to improve reliability and robustness
  - Develop accounting and reporting
- **Address integration issues:**
  - With large clusters, with storage systems
  - Ensure that large clusters can be accessed via grid
  - Issue of integrating with other applications and non-LHC experiments

Overview of EGEE - Summary

Enabling Grids for E-science in Europe

- EGEE is expected to deliver a production Grid infrastructure for scientific applications
- The project started 5 months ago
  - We have a running grid service based on LCG-2
  - All EGEE activities are well advanced
  - Next generation middleware being designed – first prototype made available to applications
- Biomedical and physics are the pilot applications domains that will lead the exploitation of the EGEE Grid infrastructure
- The first project conference held in Cork (Ireland) 18-22nd April
  - http://public.eu-egee.org/kickoff/index.html
- Project homepage
  - http://www.eu-egee.org/