Introduction to Grids and the EGEE project

General presentation

Last update End May 2007
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
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
• 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
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
• 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!
• Aim: to develop, build and maintain a distributed computing environment for the storage and analysis of data from the four LHC experiments
  ▪ Ensure the computing service
  ▪ … and common application libraries and tools

• “Tier” infrastructure with Tier-0 at CERN, 11 Tier-1 centres and more than 100 Tier-2, and Tier-3 centres

• Phase I – 2002-05 – Development & planning

• Phase II – 2006-2008 – Deployment & commissioning of the initial services

➔ LCG is not a development project – it relies on EGEE (and other Grid projects) for Grid middleware development, application support, Grid operation and deployment
• **Purpose**
  – Understand what it takes to operate a real Grid service
  – Trigger and verify Tier-1 & large Tier-2 planning and deployment –
    - tested with realistic usage patterns
  – Get the essential grid services ramped up to target levels of reliability,
    availability, scalability, end-to-end performance

• **Four progressive steps from October 2004 to September 2006**
  – End 2004 - SC1 – data transfer to subset of Tier-1s
  – Spring 2005 – SC2 – include mass storage, all Tier-1s, some Tier-2s
  – 2\textsuperscript{nd} half 2005 – SC3 – Tier-1s, >20 Tier-2s –first set of baseline
    services
  – Jun-Sep 2006 – SC4 – pilot service

→ Autumn 2006  – LHC service in continuous operation
  – ready for data taking in 2007
Collaboration with CERN openlab

- **CERN openlab**
  - Industry consortium for Grid-related technologies with common interests
  - Testbed for cutting-edge Grid software and hardware
  - Training ground for a new generation of engineers to learn about Grid

- **Partners in openlab (2003-2005)**
  - Enterasys, HP, IBM, Intel, Oracle

- **openlab-II (2006-2008)**
  - Platform Competence Centre
    - Platform virtualisation
    - Software and hardware optimisation
  - Grid Interoperability Centre – in collaboration with EGEE
    - Integration and certification of Grid middleware
    - Standardisation
  - Security activities
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
Collaborating e-Infrastructures

Enabling Grids for E-sciencE

Potential for linking ~80 countries by 2008

European Commission co-funded projects
Projects with other funding

EGEE-II INFSO-RI-031688
Introduction to Grids and the EGEE project
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 domains**
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics
  - Earth Sciences
  - Fusion
  - High-Energy Physics
  - Life Sciences

---

Encourage inter-disciplinary research and increase data inter-operability!
• 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
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)
• 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