EGI-InSPIRE Current Requirements & Outlook

Maria Girone, CERN-IT &
EGI-InSPIRE

Outline

- The EGI-InSPIRE project: Activities and Communities
- Common solutions across diverse communities
- Current Requirements
  - Data Management
- The European Digital Agenda

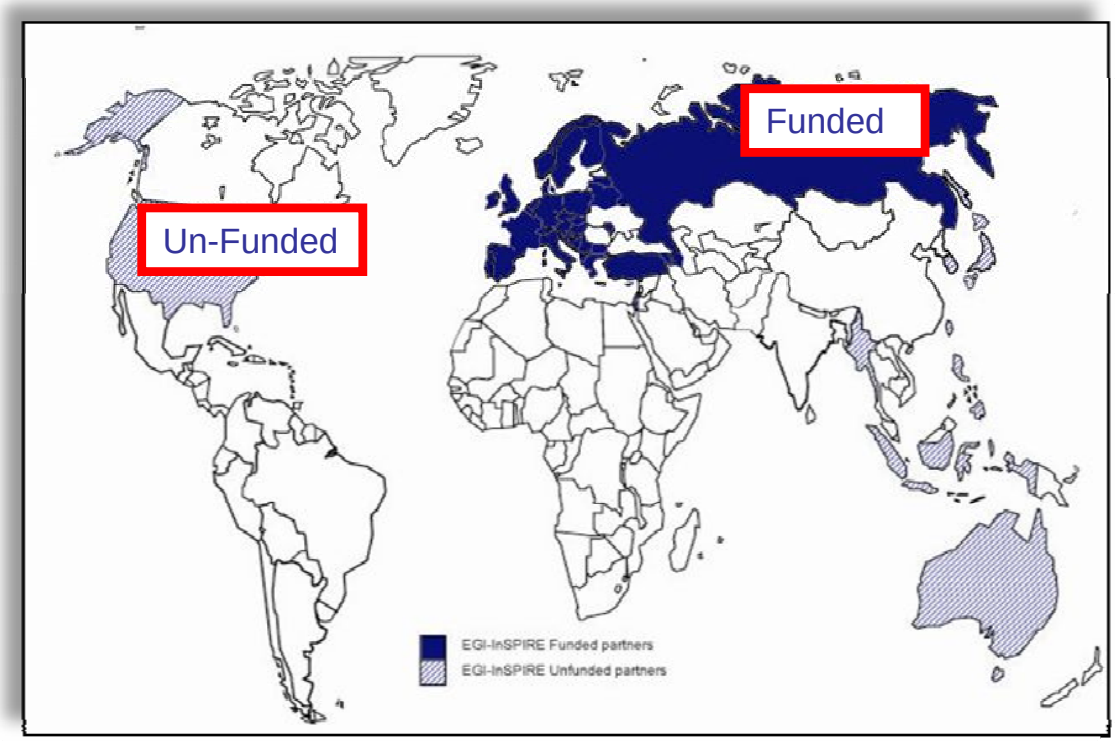
EGI-InSPIRE Project

European Grid Initiative - Integrated Sustainable Pan-European Infrastructure for Researchers in Europe

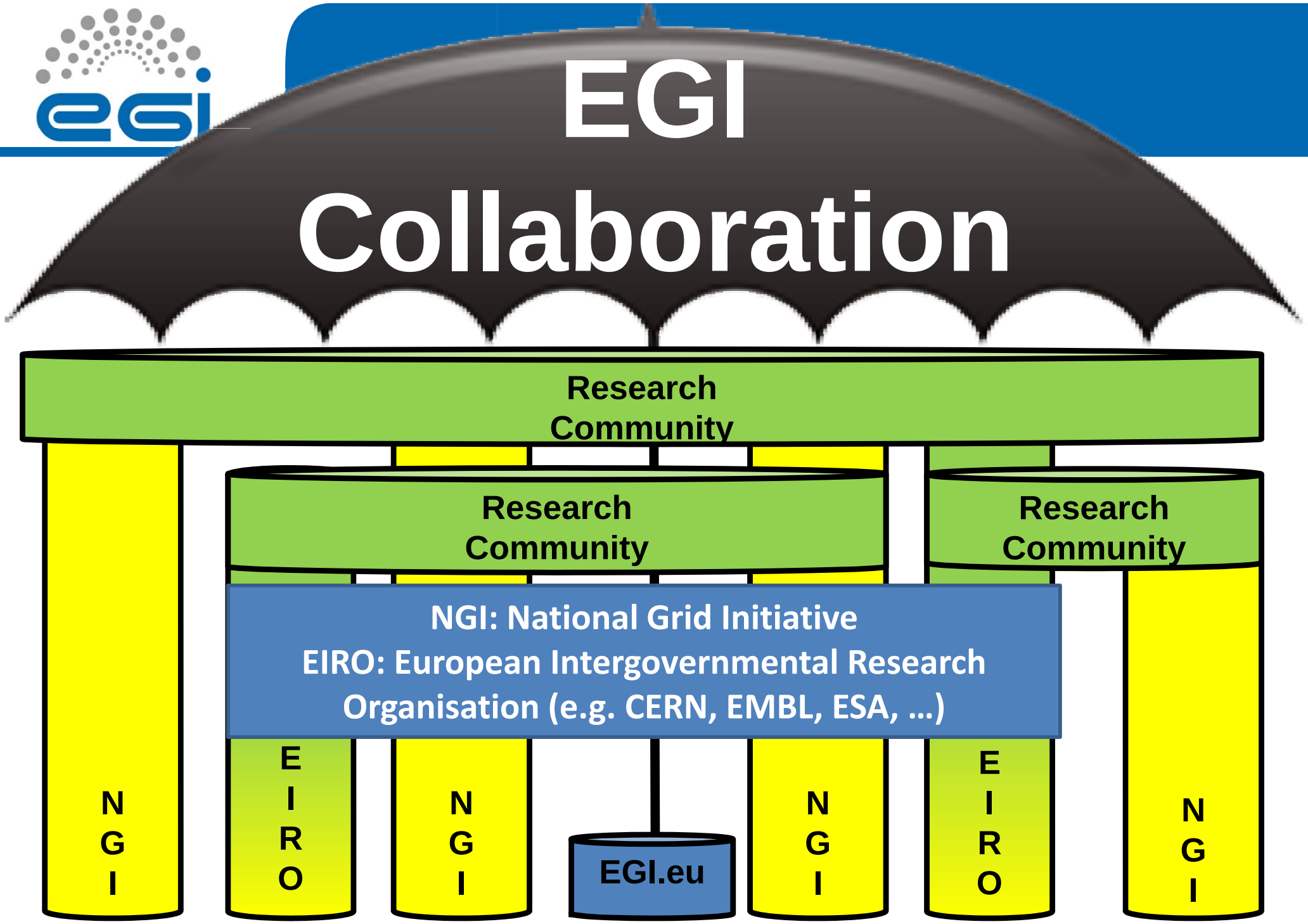
A 4 year project with €25M EC contribution

- Total Effort ~€330M
- Effort: 9261PMs

Project Partners (51)
EGI.eu, 38 NGIs, 2 EIROs
Asia Pacific (9 partners)



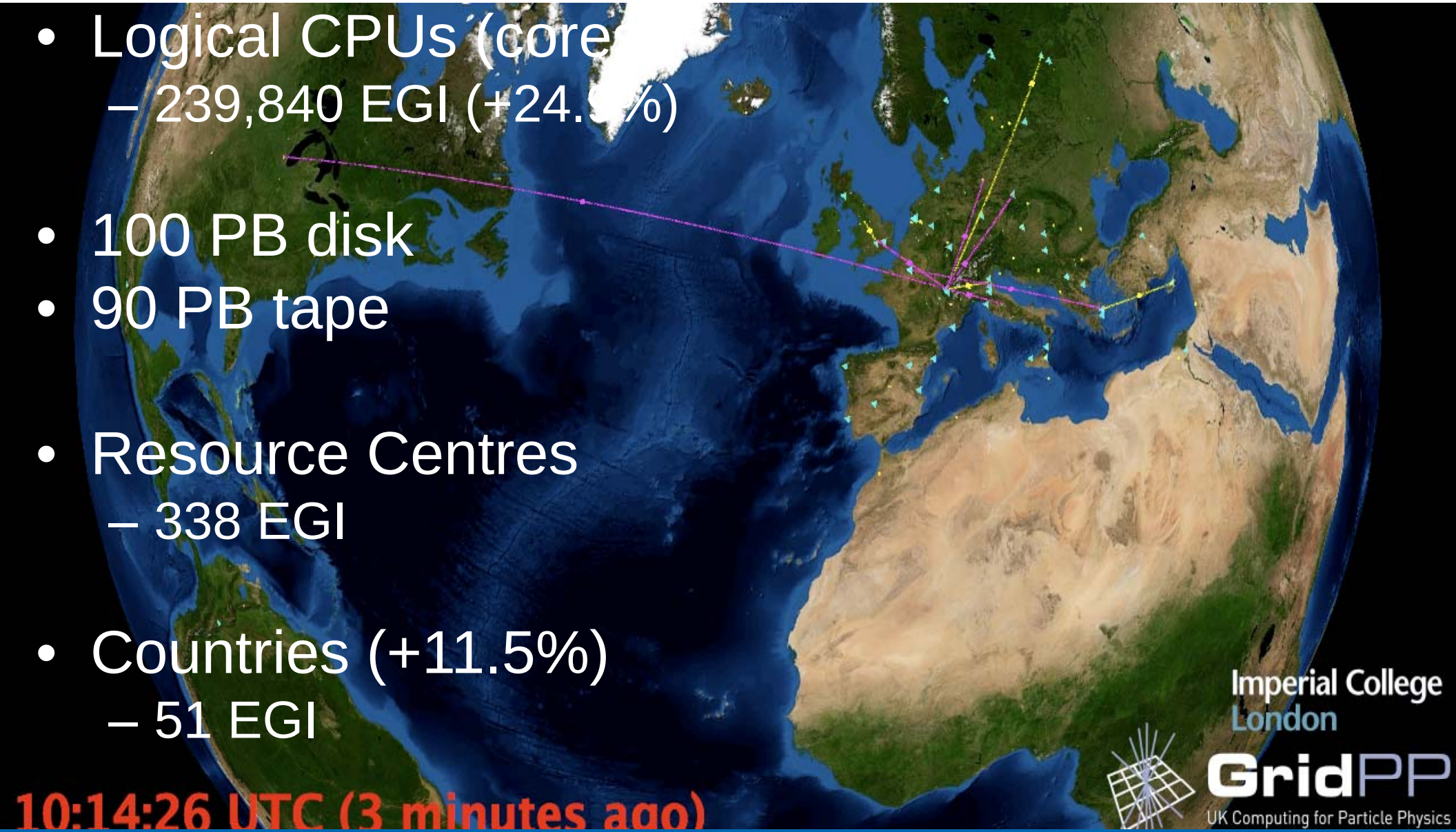
EGI Collaboration



EGI-InSPIRE Goals

- **Multi-disciplinary** project aimed at identifying and providing common, sustainable solutions
- Goals:
  - The continued **operation** and expansion of today's production **infrastructure** that can be increasingly sustained outside of specific project funding
  - The support for current heavy **users** (HUCs) as they move to sustainable support models for their own communities
  - Allow the integration of **new** technologies (e.g. **clouds** and **virtualization**) and heterogeneous resources (e.g. HTC & HPC)

European Grid Infrastructure
(April 2011 and yearly increase)



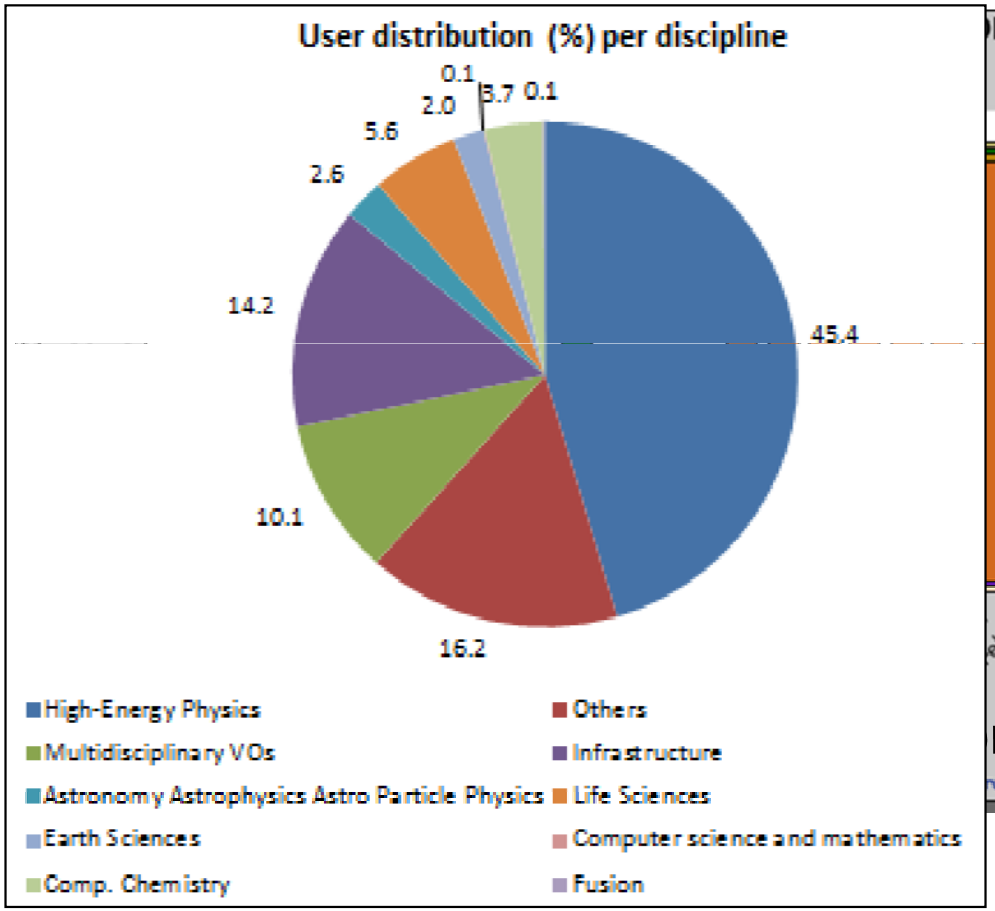
- Logical CPUs (core
  - 239,840 EGI (+24.[illegible]%)
- 100 PB disk
- 90 PB tape
- Resource Centres
  - 338 EGI
- Countries (+11.5%)
  - 51 EGI

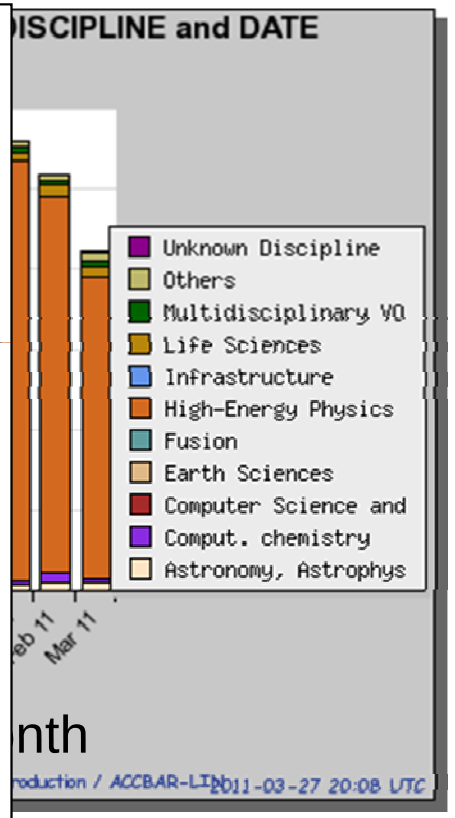
EGI Usage (April 2011)

~11000 End-users
220 VOs
~30 active VOs

User Communities

- Archeology
- Astronomy
- Astrophysics
- Civil Protection
- Comp. Chemistry
- Earth Sciences
- Finance
- Fusion
- Geophysics
- High Energy Physics
- Life Sciences
- Multimedia
- Material Sciences
- …





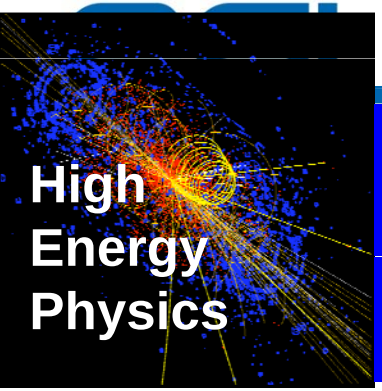
Average usage 2010-2011 vs 2009-2010

- 1Mjobs/day (+420%)
- 74.6M CPU wall clock hours/month (+86.5%)
- 551M HEP-SPEC06 CPU wall clock hours/month (+99.4%)

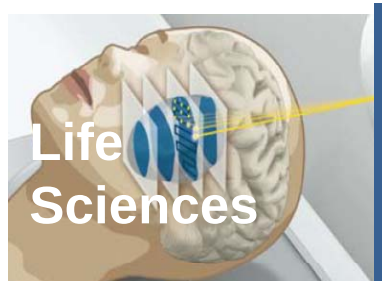
Work Package Breakdown

Work Package	Goal
SA1	Operations
SA2	Software Provisioning
SA3	**Support for Heavy User Communities (HUCs)**
NA1	Project Management
NA2	Dissemination
NA3	User Community Support (other than HUCs)
JRA1	Operational Tools

Communities & Activities

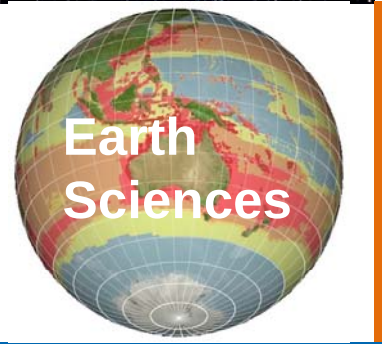


The four LHC experiments use **grid** computing for data distribution, processing and analysis. Strong focus on common tools and solutions. Areas supported include: Data Management, Data Analysis and Monitoring. Main VOs: ALICE, ATLAS, CMS, LHCb



Focuses on medical, biomedical and bioinformatics sectors to connect worldwide laboratories, share resources and ease access to data in a secure and confidential way. Supports 4 VOs (biomed, lsgri, vlemed and pneumogrid) across 6 NGIs via the Life Science Grid Community.

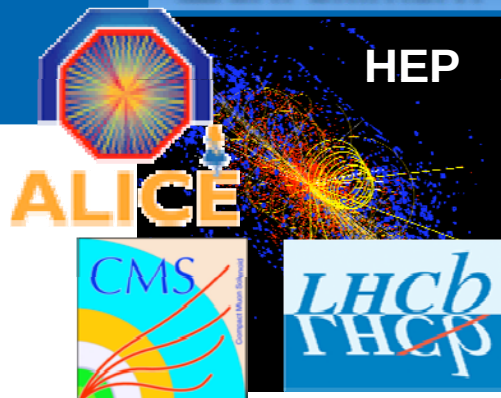
Covers a variety of projects including the European Extremely Large Telescope (E-ELT), the Square Kilometre Array (SKA) and Cerenkov Telescope Array (CTA). Activities focus on visualisation tools and database/catalog access from the grid.

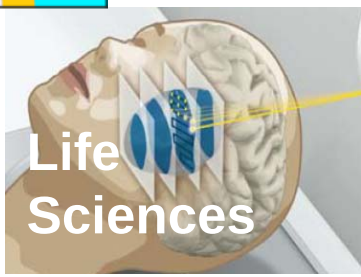


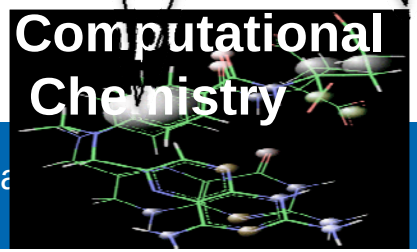
Covers seismology, atmospheric modelling, meteorological forecasting, flood forecasting and climate change. Provides access from the **grid** to resources within the Ground European Network for Earth Science Interoperations - Digital Repositories (GENESI-DR). Also assists scientists working on climate change via the Climate-G testbed.

Common Solutions

Connect various tools with multiple communities: the main disciplines and also Fusion, Computational Chemistry







Common Solutions

Ganga: job definition and management in HEP (ATLAS and LHCb, Compass, Harp), Fusion, L&E Sciences.

HEP

Life Sciences

Astronomy & Astrophysics

Earth Sciences

Computational Chemistry

Fusion

Common Solutions

Hammercloud: a site stress testing system to validate site usability. Used by ATLAS, CMS and LHCb.

HEP

Life Sciences

Astronomy & Astrophysics

Earth Sciences

Computational Chemistry

Fusion

Common Solutions

HEP

Life Sciences

Astronomy & Astrophysics

Earth Sciences

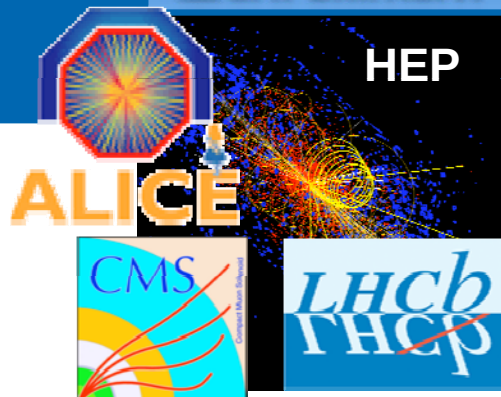
Fusion

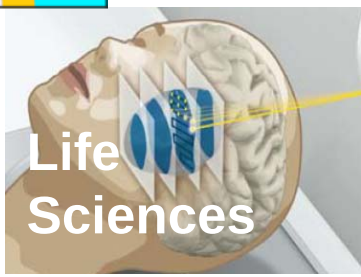
Computational Chemistry

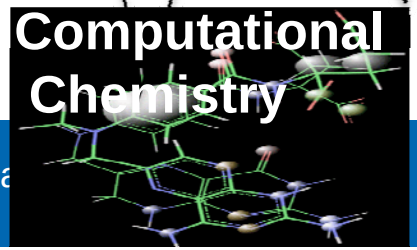
Dashboards: single entry point to monitoring of all 4 LHC experiment activities on the grid. (Mini) Dashboard also used by E&L Sciences.

Common Solutions

DIRAC: workload and data management used by LHCb plus other HEP experiments with interest from Earth Sciences.

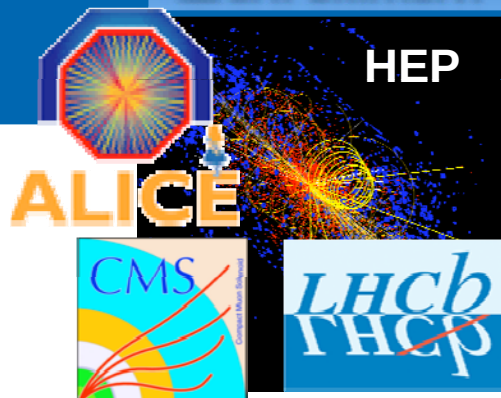


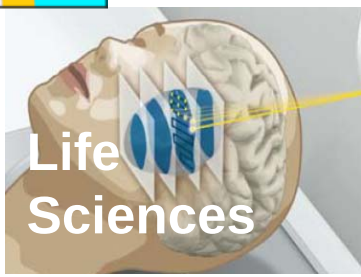


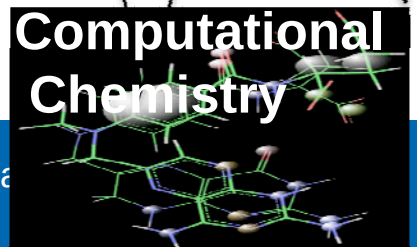


Common Solutions

MPI: used by A&A, Fusion, Computational Chemistry to handle parallel execution in grid environments.

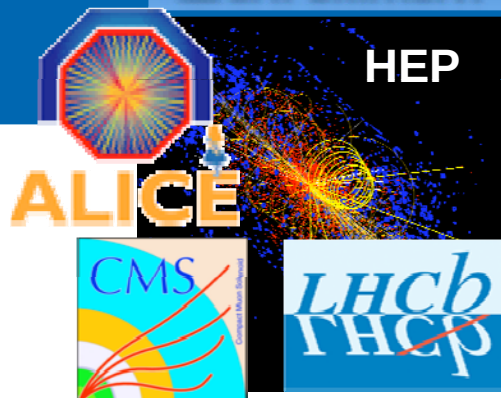


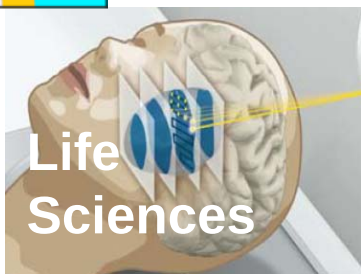


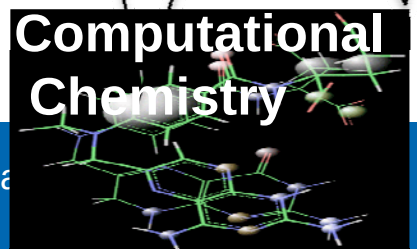


Common Solutions

GRelC: set of advanced data grid services to manage Databases on the Grid. Used by L&E Sciences and A&A.







Current Requirements - HEP

- Have described so far activities that began prior to EGI-InSPIRE but have been continued and/or extended
- EGI-InSPIRE has explicitly foreseen the need to address **issues arising** from first LHC data taking and production
- Key examples are:
  - **Data Management** in general, including:
    - Data Placement / Dynamic Caching & Evolving Network Architectures
    - Catalogue / Storage Element Consistency
  - **Data analysis** support
- Work done in collaboration with WLCG and the LHC experiments as well as technology providers such as EMI (see P. Fuhrmann talk)

Data Management – 1

- Crucial area for LHC and other HEP experiments
  - Data volumes: tens of PB/year, rates: up to 200TB/day between sites, several hundred active analysis users / expt, 1M analysis jobs / day (see M. Ernst talk)
- Experience from first data taking has shown that some assumptions on data placement are no longer optimal
  - Based on MONARC, the initial phase of LHC **data distribution** was based on static pre-placement
    - Significant fraction of such data never read!
- Computing models now driving towards dynamic data placement (see I. Fisk talk)
  - Replication is based on usage ("popularity") – this results in more efficient network and storage utilization
- Implemented first for ATLAS, now for CMS and LHCb

Data Placement / Dynamic Caching & Evolving Network Architectures

Data Management – 2

- With 50PB of data storage across a large number of sites worldwide, inconsistencies can easily arise
  - Data that resides on Storage Elements but not in various catalogs (grid, experiment) referred to as “Dark Data”
    - One site recently reported 70TB dark data!
- Using a messaging-based system, various catalogs and SEs can talk to each other and implement lazy synchronization

Catalogue / Storage Element Consistency

Data Analysis Support

- Covers the final stage of data processing leading on to publication
  - Large number of users/month (~1000) and analysis jobs/day (~1M) running across (~100) Tier2 and other sites – “chaotic” data access
  - All frameworks support heterogeneous back-ends
- Ganga (ATLAS, LHCb) used by 10 other communities and 500 – 600 users
- Common site stress testing system (Hammercloud) used by ATLAS, CMS and LHCb
- New areas of commonality and optimization
  - Move to “community support” model
  - Simplify data access and improve monitoring
  - Use common components and frameworks, such as for job submission and file transfer (built on gLite / EMI FTS)
    - An area of potential future common work and simplification

The Next Challenge

- EGI-InSPIRE has demonstrated that multiple disciplines **can** work together, which leads naturally into the next challenge
- The EU intends to invest significantly from 2014 in what it refers to as "the Digital Agenda"
- This should address as wide a range of disciplines as possible – from science to humanities and beyond
  - eHealth, climate change, science and education explicitly mentioned amongst many others
- As a multi-disciplinary project that is largely data-driven, EGI-InSPIRE offers an attractive foundation on which to build such an effort
- An initial step is a data management requirements matrix across the EGI-InSPIRE communities: review matrix in **Panel on Requirements**

Draft Requirements Matrix

Discipline	Preservation	Volume	Access	Security
Earth Sciences	**Millennia**	No intrinsic limit – broken down into many different areas	Cross-correlation	Some data clearly sensitive
Astronomy & Astrophysics	**Millennia**	No limit – current projects range from 100 TB – low PB range	Cross-correlation between different observations	Explicit policy on placing data in public domain after 1 year
Life Sciences	**1+ centuries**	# people (life-forms) x data	By Individual By condition Evolution	**Patient privacy;** **Copyrighted tools;** **Competing industries**
High Energy Physics	**A few decades** More for educational purposes?	**100PB – 1EB** today; Previous generations, e.g. LEP, are in commodity market	Range from sequential access to bulk data to many // accesses	Often traded for performance & scalability

Summary

EGI-InSPIRE is a **multi-disciplinary** project that supports large-scale, long-term projects

- Peta-scale in terms of computing resources, global in terms of communities and deployment
- High level of service both quantitatively and qualitatively
- Common, sustainable solutions with more to come

Long-term means adapting to changing requirements and technology

- Multi-core; virtualization; clouds; network capacity and topology
- Have to be integrated / adopted whilst constantly ramping up the service in terms of capacity