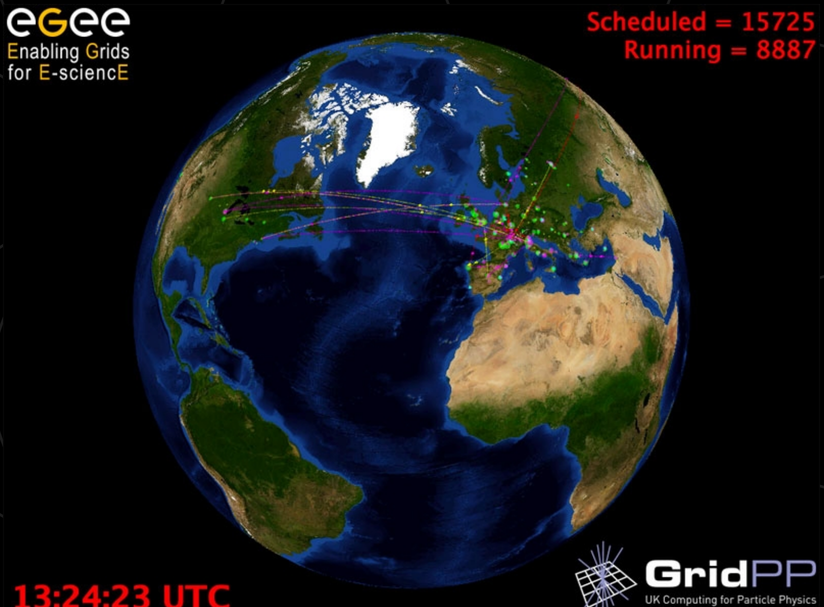
vl·e virtual laboratory for e·science

Grid: data delen op wereldschaal

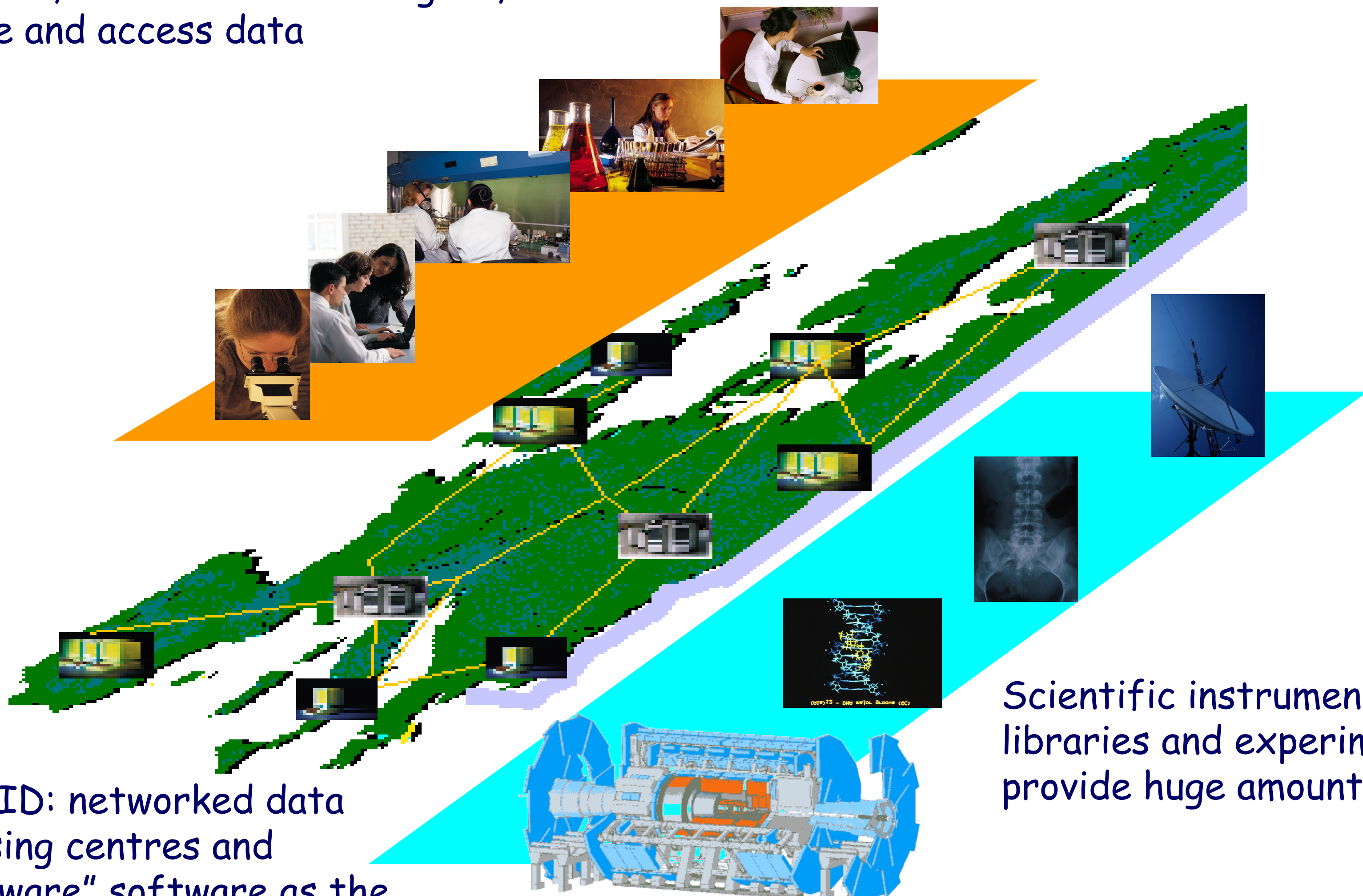
David Groep, NIKHEF



Graphics: Real Time Monitor,
Gidon Moont, Imperial College London, see http://gridportal.hep.ph.ic.ac.uk/rtm/

Grid from 10 000 feet

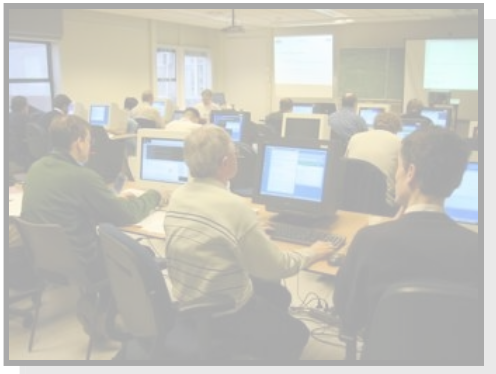
Work regardless of geographical location, interact with colleagues, share and access data



The GRID: networked data processing centres and "middleware" software as the "glue" of resources.

Scientific instruments, libraries and experiments provide huge amounts of data

What is Grid?



Cycle scavenging

- harvest idle compute power
- improve RoI on desktops

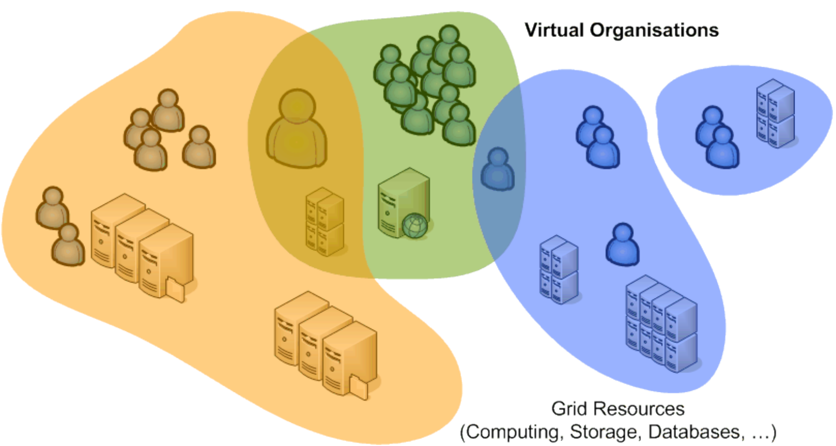
Cluster computing and storage

- What-if scenarios
- Physics event analysis
- Improve Data Centre Utilization

Cross-domain resource sharing

- more than one organisation
- more than one application
- more than one …

- open protocols
- collective service



Why would we need it?

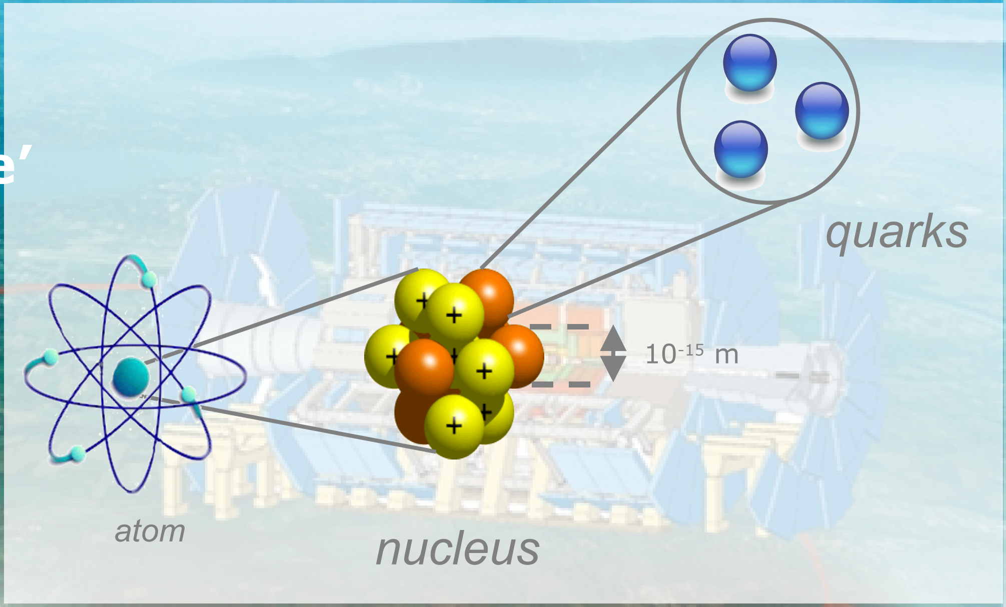
Collected data in science and industry grows exponentially:

The Bible	**5**	**MByte**
X-ray image	5	MByte/image
Functional MRI	1	GByte/day
Bio-informatics databases	500	GByte each
Refereed journal papers	1	TByte/yr
Satellite world imagery	5	TByte/yr
US LoC contents	20	TByte
Internet Archive 1996-2002	100	TByte
Particle Physics today	1	PByte/yr
LHC era physics	**20**	**PByte/yr**

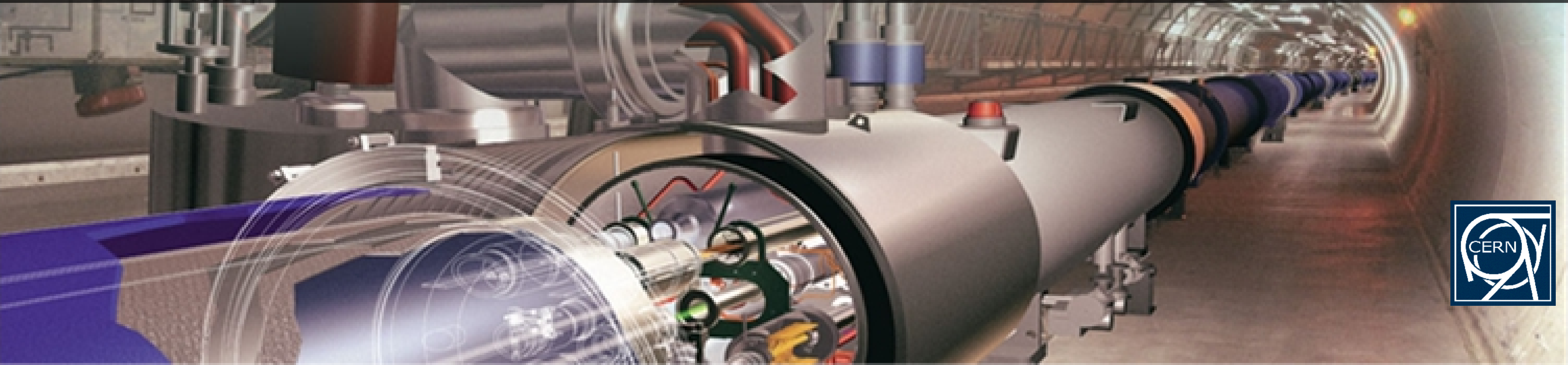
Some use cases: LHC Computing

Large Hadron Collider

- ‘the worlds largest microscope’
- ‘looking at the fundamental forces of nature’
- 27 km circumference
- Located at CERN, Geneva, CH



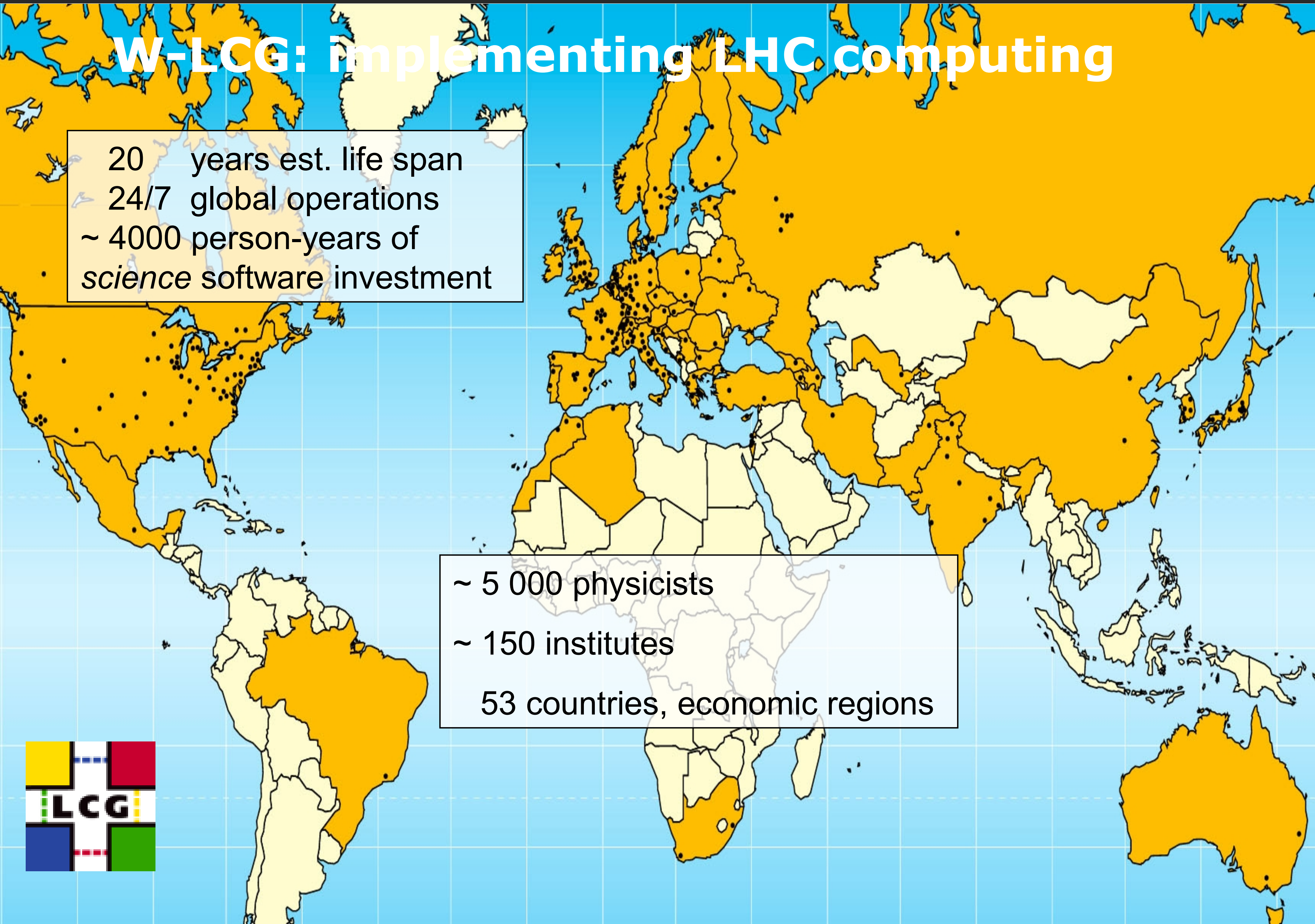
~ 20 PByte of data per year, ~ 50 000 modern PC style computers



W-LCG: implementing LHC computing

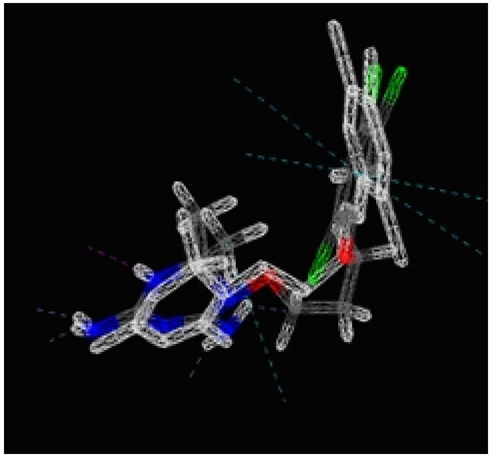
20 years est. life span
24/7 global operations
~ 4000 person-years of *science* software investment

~ 5 000 physicists
~ 150 institutes
53 countries, economic regions



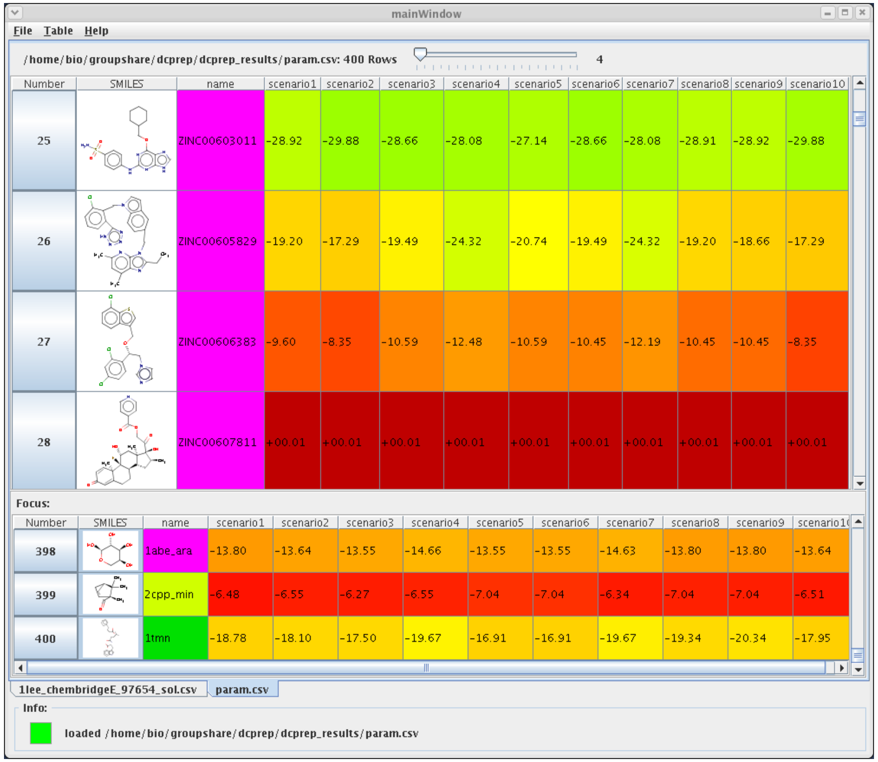
WISDOM: drug discovery

Wide-area In-Silico Docking On Malaria



over 46 million ligands virtually docked on malaria and H5N1 avian flu viruses in less than a month

used 100 *years* of CPU power
speedup ~ 100 times!



- **47 sites**
- **15 countries**

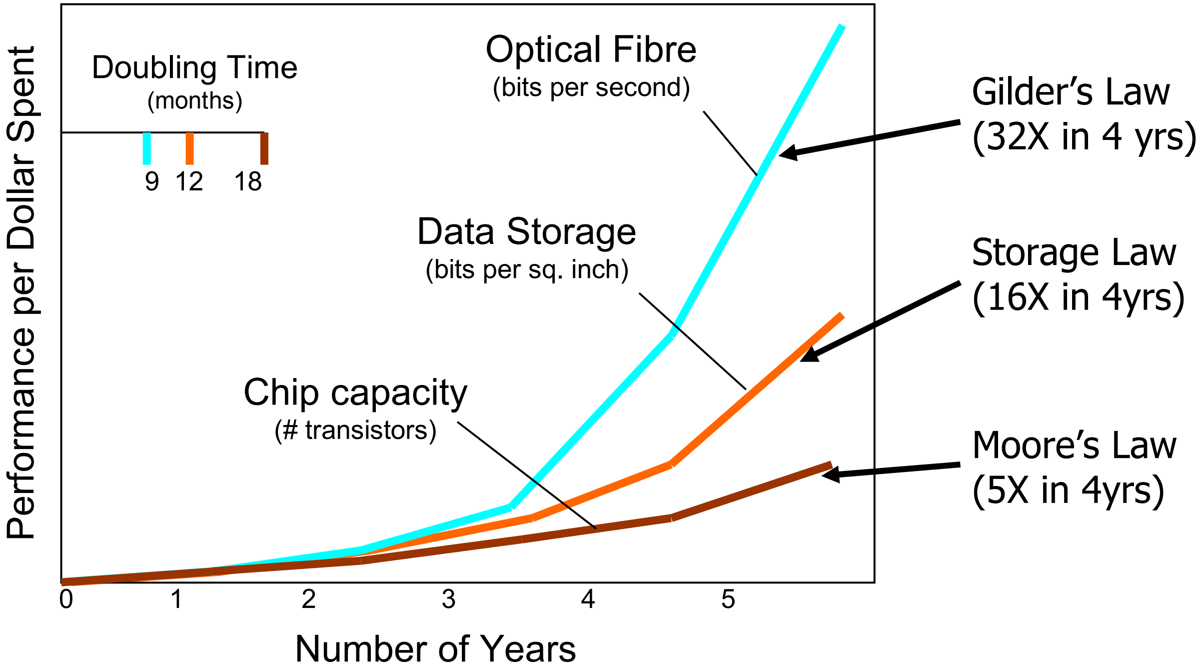
- 3000 CPUs
- 12 TByte disk

vl·e

egee
Enabling Grids for E-sciencE

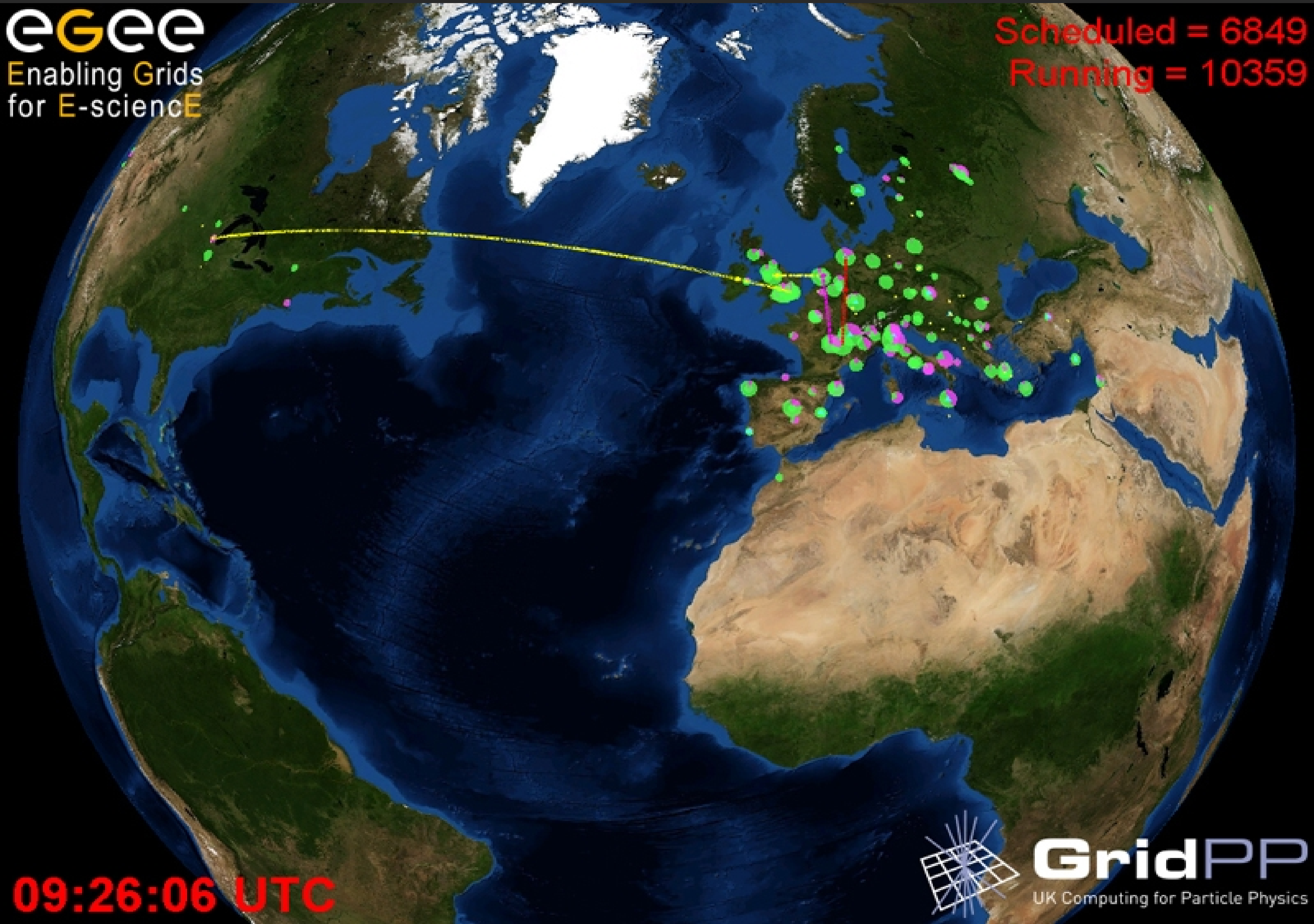
Why Grid computing – today?

- New applications need larger amounts of data or computation
- Larger, and growing, distributed user community
- Network grows faster than compute power/storage



Graphic: "The Triumph of Light", Scientific American, January 2001

Making the Grid …



Three essential ingredients for Grid

'Access computing like the electrical power grid'

A grid combines resources that

- Are not managed by a single organization
- Use a common, open protocol … that is general purpose
- Provide additional qualities of service, *i.e.*, are usable as a collective and transparent resource

GRID today

DAILY NEWS AND INFORMATION FOR THE GLOBAL GRID COMMUNITY / JULY 22, 2002: VOL. 1 NO. 6

WHAT IS THE GRID? A THREE POINT CHECKLIST
By Ian Foster Argonne National Lab & University of Chicago

The recent explosion of commercial and scientific interest in the Grid makes it timely to revisit the question: What is the Grid, anyway? I propose here a three-point checklist for determining whether a system is a Grid. I also discuss the critical role that standards must play in defining the Grid.

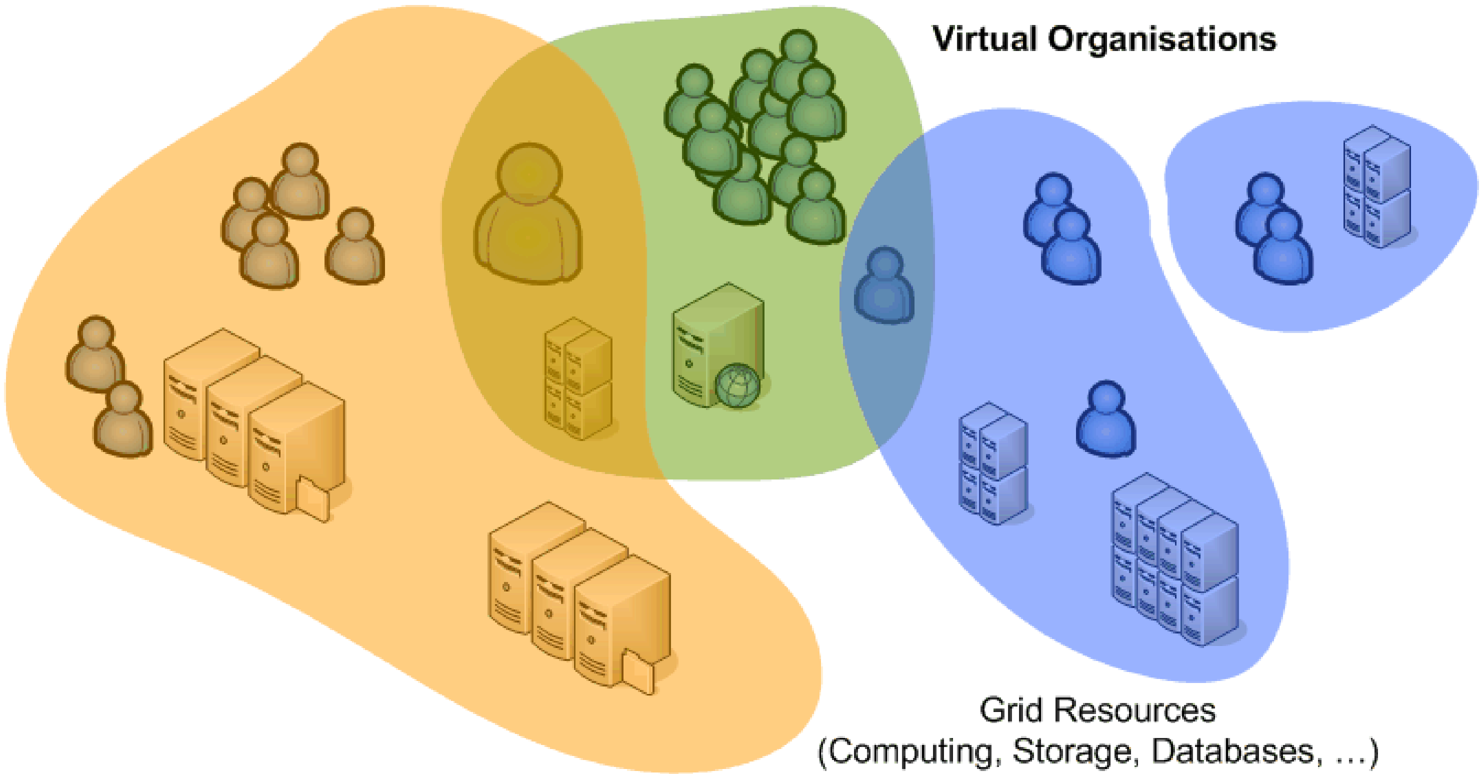
The Need for a Clear Definition Grids have moved from the obscurely academic to the highly popular. We read about Compute Grids, Data Grids, Science Grids, Access Grids, Knowledge Grids, Bio Grids, Sensor Grids, Cluster Grids, Campus Grids, Tera Grids, and Commodity Grids. The skeptic can be forgiven for wondering if

Source: Ian Foster in *Grid Today*, July 22, 2002; Vol. 1 No. 6, see http://www-fp.mcs.anl.gov/~foster/Articles/WhatIstheGrid.pdf

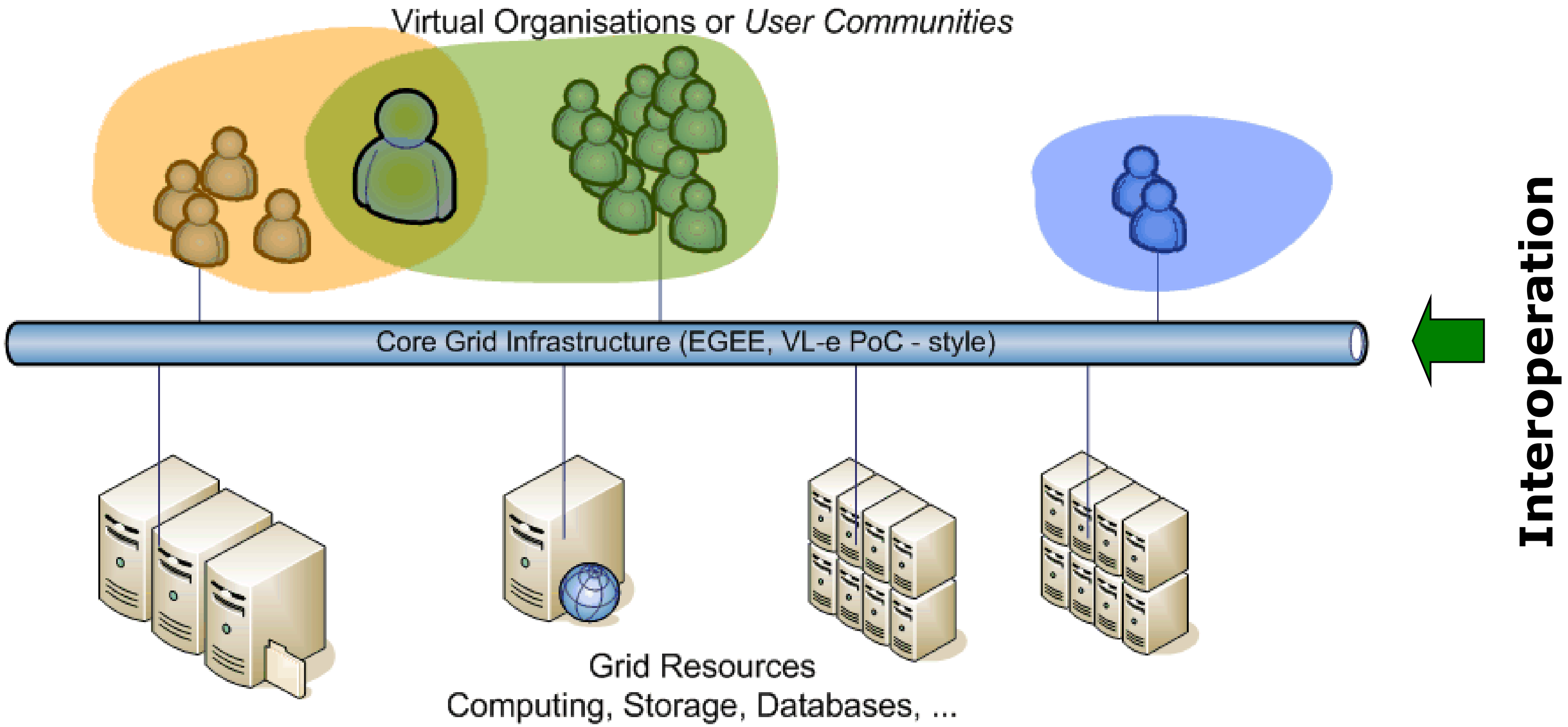
Virtual Organisations

The communities that make up the grid:

- **not under single hierarchical control**,
- (temporarily) **joining forces** to solve a particular problem at hand,
- bringing to the collaboration a subset of their resources,
- sharing those **at their discretion** and each **under their own conditions**.



Building Grid Infrastructures



- Protocols: common syntax and sematics for grid operations
- APIs: making grid concepts accessible from the applications
- Portals and workflows: bridging the end-user gap

Standards

OpenGridForum

- Standards, such as those by IETF, OASIS, OGF, &c aid interoperability and reduce vendor lock-in

- as you go higher up the stack, you get less synergy
  - Transport: IP/TCP, HTTP, TLS/SSL, &c well agreed
  - Web services: SOAP used to be the solution for all …
    … but 'Web 2.0' shows alternatives tailored to specific applications gaining popularity
  - Grid standards:
    low-level job submission (BES, JSDL), management (DRMAA), basic security (OGSA-BSP Core, SC) there
  - higher-level services still need significant work …

see also http://www.ogf.org/

Grid Infrastructure

Realizing ubiquitous computing requires a *persistent infrastructure*, based on standards

Hardware infrastructure

clusters, supercomputers, databases, mass storage, visualisation

Software infrastructure

execution services, workflow, resource information systems, database access, storage management, meta-data

Application infrastructure

user support, and ICT experts … with domain knowledge

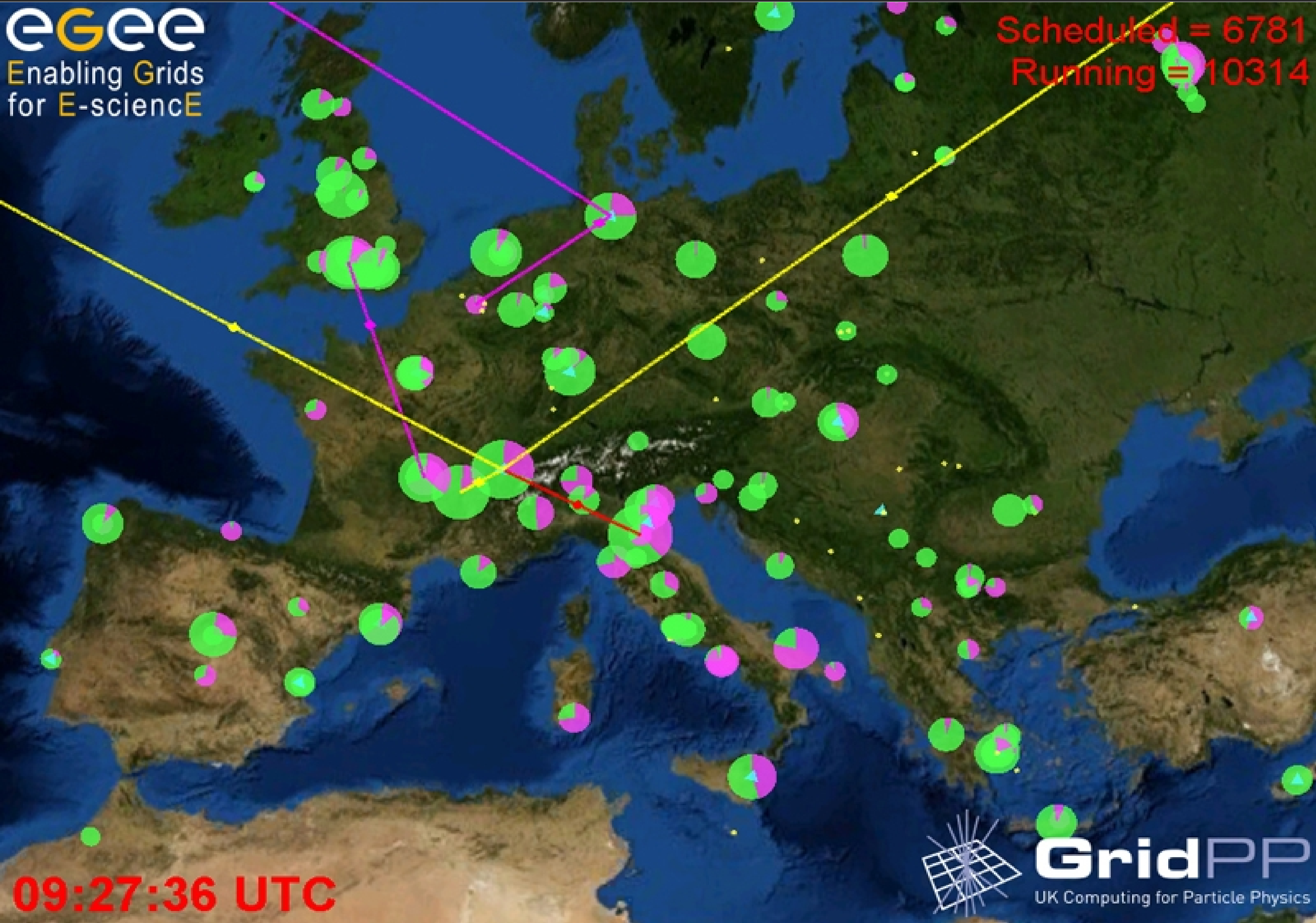
Interoperation and standards

- Standards are essential for adoption
  - resource providers are not inclined to provide *n* different interfaces

- But a pragmatic approach is needed today
  - GIN (Grid Interoperation Now) leverage existing de-facto agreements
  - be agnostic to changes at the protocol level e.g. by leveraging higher-level APIs (SAGA)

  - *do not get married to a particular protocol hype*

Where do we stand today?



Issues for today and tomorrow

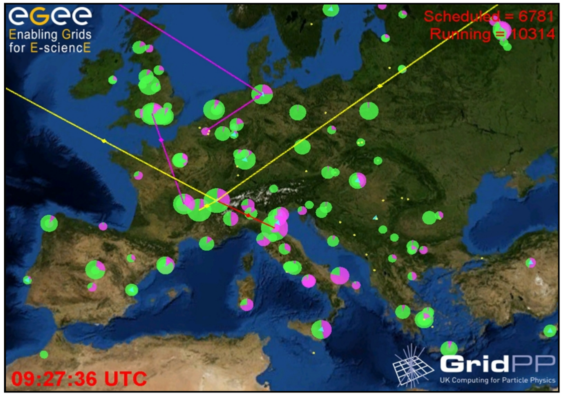
- Distributed security
  - any computer, desktop and laptop, must be assumed compromised

  - identity vetting and community membership assertions needed in cross-domain grids
  - trust between organisations needed
    - we demonstrated this in science – globally!
    - federated access to a wide range of resources coming
  - security, privacy policies must be coordinated
    - essential for a mainstream, sustained, infrastructure

strike balance between security and usability …

- help with identity federations, on-line credentials
- portals and canned (web) applications

Working at scale



Grid is an error amplifier …

‘passive’ controls are needed to push work away from failing resources

Resource information systems are the backbone of any real-life grid

Grid is much like the ‘Wild West’

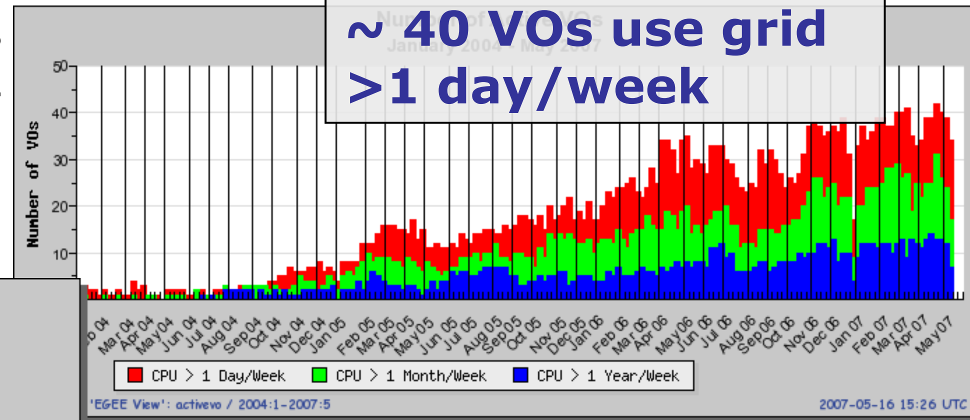
- almost unlimited possibilities – but as a community plan for scaling issues, and a novel environment
- users and providers *need to interact* and articulate needs

Grid Infrastructures Work

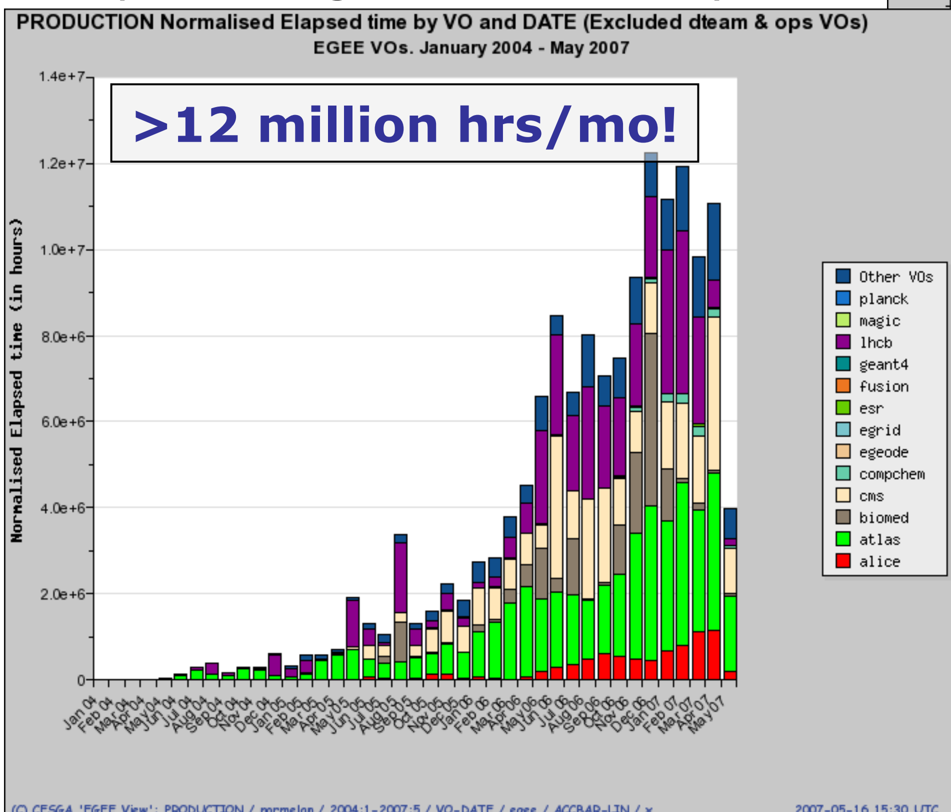
egee
Enabling Grids for E-sciencE

Number of **active** VOs in EU since 2004

260 VOs total in EU
~ 40 VOs use grid >1 day/week



Compute usage since 2004 by VO



>12 million hrs/mo!

over 20 VOs hosted in NL

www.biggrid.nl

A reliable Grid Infrastructure needs operational support:

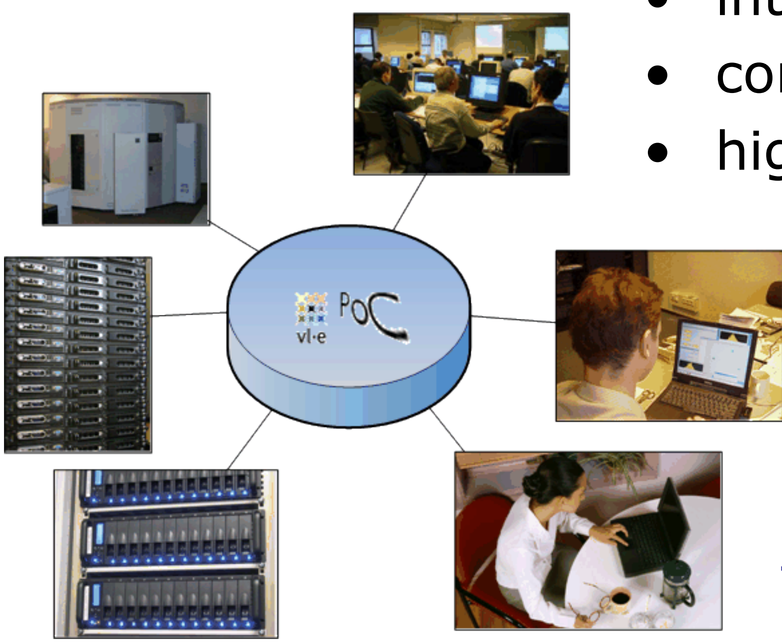
- availability monitoring
- reporting and follow-up
- user support

data: EGEE monitoring, RAL and CESGA, http://goc.grid-support.ac.uk/gridsite/accounting/

Common environment

Common infrastructure for e-Science in NL
provided in the *VL-e Proof-of-Concept*

- interoperable interfaces to resources
- common software environment
- higher-level ‘virtual lab’ services



Central Facilities:
SARA, NIKHEF, RC-RUG, Philips

Join yourself: user-interfaces,
distributed clusters, storage

http://poc.vl-e.nl/distribution/

vl·e

http://www.vl-e.nl/