GRID COMPUTING & E-SCIENCE

Presentation by:

Lê Minh Nhựt	10070491
Tô Vũ Song Phương	11070468

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

WHAT IS E-SCIENCE?

- First introduced in 1999 by John Taylor from the United Kingdom's Office of Science and Technology.
- “e-Science is about global collaboration in key areas of science and the next generation of [computing] infrastructure that will enable it.” - *John Taylor , the United Kingdom's Office of Science and Technology.*
- “collaborative science that is made possible by the sharing across the Internet of resources (data, instruments, computation, people’s expertise...).” -*Richard Hopkins , Rutherford Appleton Laboratory.*

WHAT IS E-SCIENCE?

- "E-Science (or eScience) is computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing; the term sometimes includes technologies that enable distributed collaboration." (wikipedia).
- Offers scientists a scope to store, interpret, analyse and network their data to other work groups.
- A virtual organization: often involves collaboration between scientists who share resources.

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

THE AIM OF E-SCIENCE

- Ensure the advancement of computer technology continues.
- Enable better research in all disciplines.
  - Development of new computational tools and infrastructures to support scientific discovery.
  - Develop new methods to analyze vast amounts of data accessible over the internet using vast amounts of computational resources.

THE AIM OF E-SCIENCE

- Invention and exploitation of advanced computational methods
  - To generate and analyse research *data*
  - To develop and explore *models and simulations*
  - To enable *dynamic* distributed virtual organisations

THE AIM OF E-SCIENCE

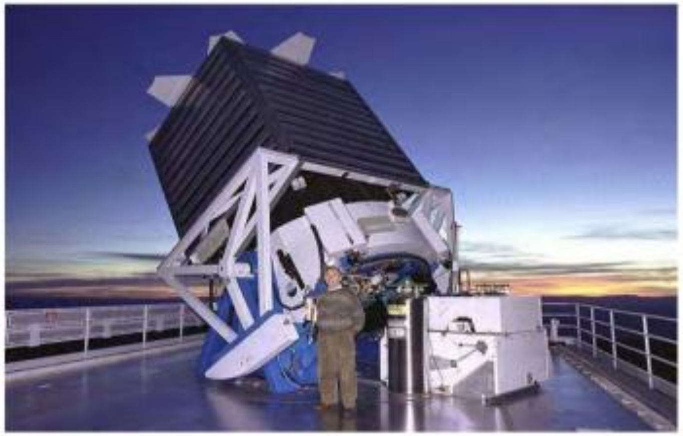
- Data intensive science
- Computation intensive science
- Simulation-based science
- Remote access to experimental apparatus
- Collaborative working

DATA-INTENSIVE SCIENCE

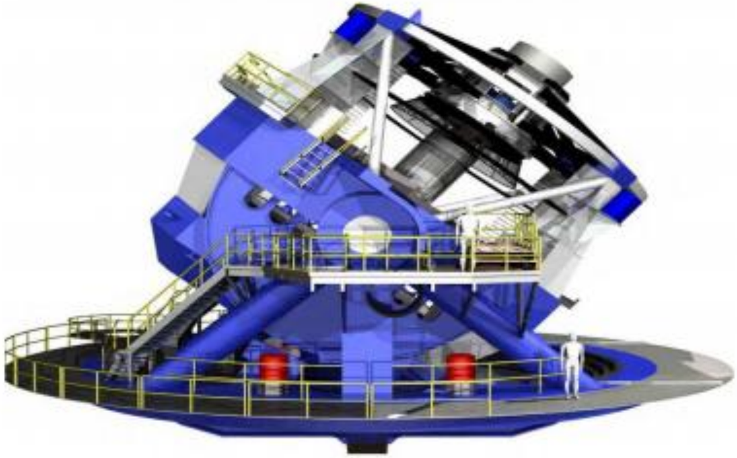
- The way of science researching changes from few data, lots of thinking, to lots of data & analysis.
- Need to develop new methods to store, analyze and share vast amounts of data.
- Including both creating new data and accessing very large data collections, data deluges from new technologies.

DATA-INTENSIVE SCIENCE

LHC(Large Hadron Collider) 60TB/day



Apache Point Telescope SDSS 15TB/day



Large Synoptic Survey Telescope 30TB/day

Illumina Genome Analyzer 1TB/day

COMPUTATION INTENSIVE SCIENCE

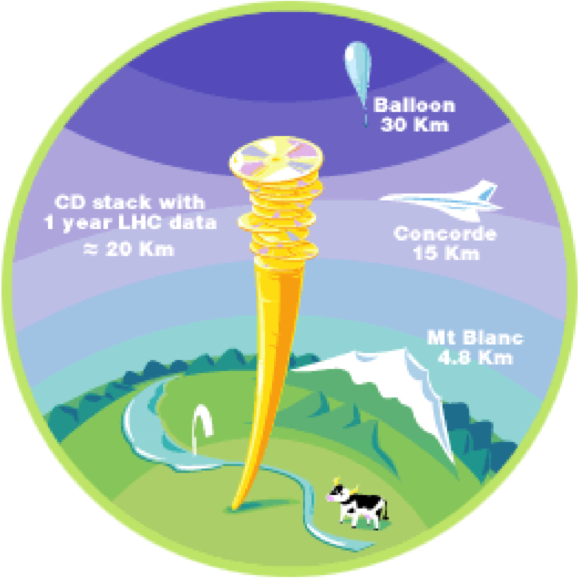
- Scientific instruments and experiments provide huge amount of data.
- Analysis and process these huge amounts of data requires very high computing power of today's fastest PC processors.

SIMULATION-BASED SCIENCE

- Simulation represents another new problem-solving methodology which physical experiments cannot easily be performed but computational simulations are feasible.
- Simulation, reconstruction, analysis huge amounts of data requires very high computing power.
- The Japanese Earth Simulator:
  - Allowing simulations to be performed at an unprecedented 10-km horizontal resolution and generating tens of terabytes of data in a single run.

SIMULATION-BASED SCIENCE

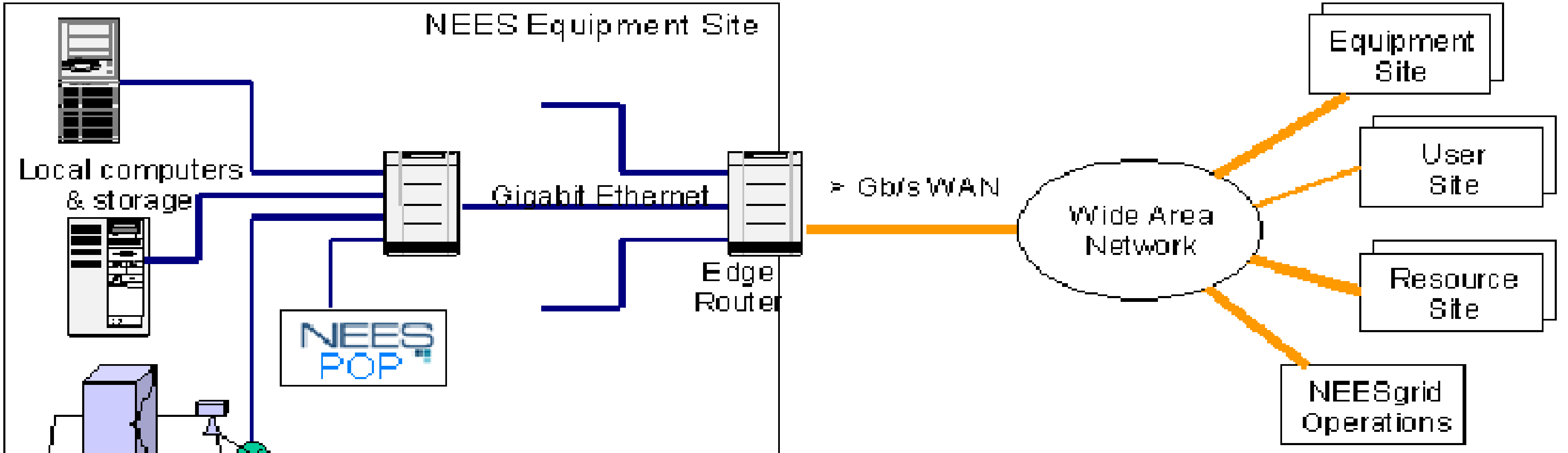
- LHC project
- Data Challenge:
  - 10 Petabytes/year of data !!!
  - 20 million CDs each year!
- Simulation, reconstruction, analysis:
  - LHC data handling requires computing power equivalent to ~100,000 of today's fastest PC processors!
- Operational challenges
  - Reliable and scalable through project lifetime of decades



REMOTE ACCESS TO EXPERIMENTAL APPARATUS

- The emergence of high-speed networks facilitate to integrate the experimental apparatus into the scientific problem-solving process.
- Earthquake Engineering Simulation (NEES)
  - An ambitious national program whose purpose is to advance the study of earthquake engineering and to find new ways to reduce the hazard earthquakes represent to life and property.
  - Collaborative tools aid (middleware) in experiment planning and allow engineers at remote sites to perform teleobservation and teleoperation of experiments, and enable access to computational resources and open source analytical tools for simulation and analysis of experimental data.

REMOTE ACCESS TO EXPERIMENTAL APPARATUS



Certificate Authority
MyProxy
Account Mgmt Tools
Index Service
Monitoring Tools
NEESgrid Website
Bugzilla
Mailing Lists

NEES Grid

COLLABORATIVE WORKING

- Collaborative working is fundamental to eScience.
- The global collaborative will lead to create a virtual community science.
- AccessGrid is multi-site videoconferencing over the Internet:
  - up to four video streams from each site
  - full echo cancellation on audio
  - distributed PowerPoint tool
  - other data integrated through one of the video streams

COLLABORATIVE WORKING

AccessGrid at Oxford

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

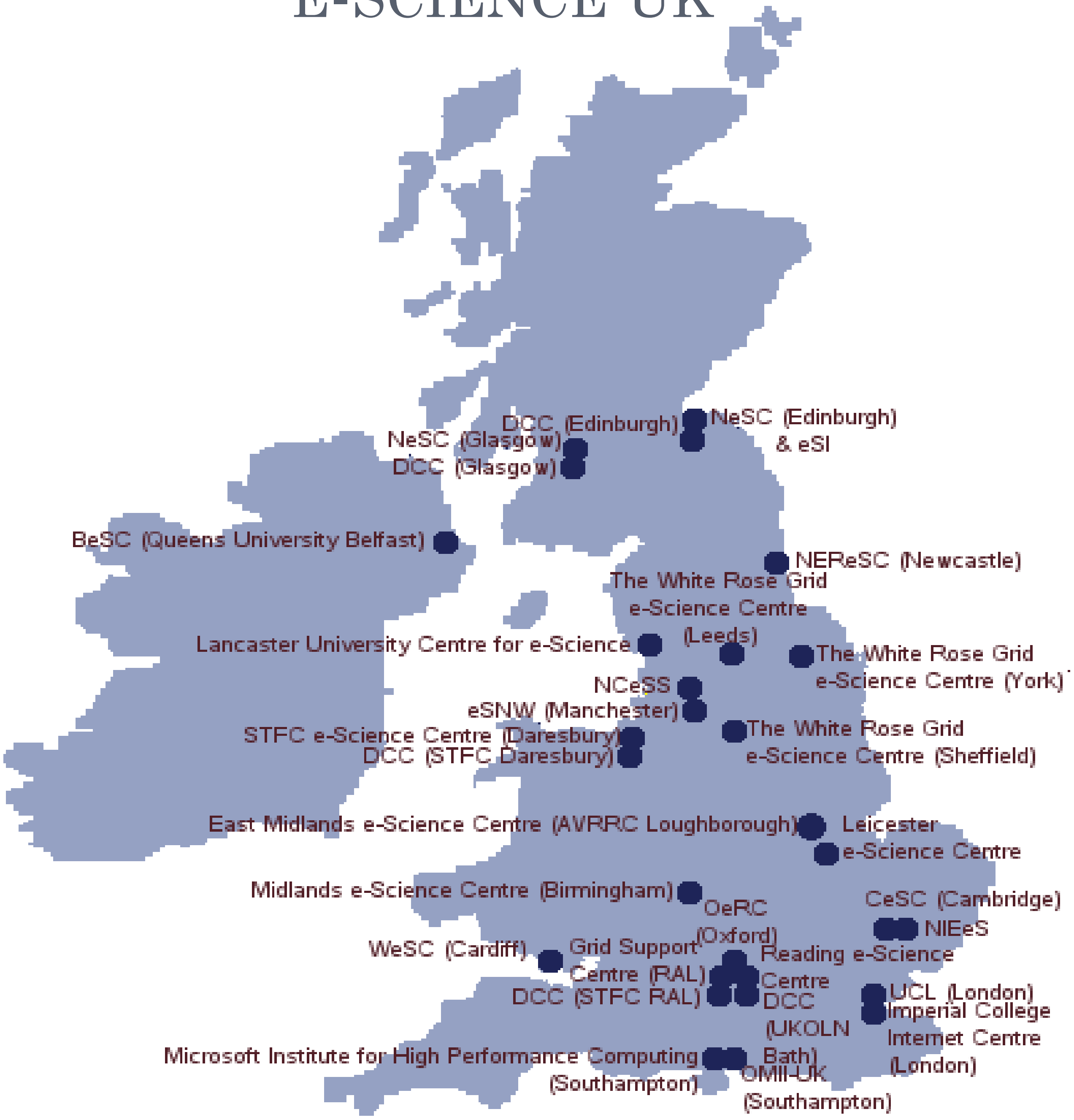
MAIN E-CIENCE GROUPS

- eScience UK
- eScience US
- eScience NL

E-SCIENCE UK

- Director General of the Research Councils, John Taylor started the term e-Science in 1999.
- In November 2000, National UK e-Science program had the initial fund of 98 million pounds.
- The UK e-Science programme comprised a wide range of resources, centres and people including the National e-Science Centre (NeSC) which is managed by the Universities of Glasgow and Edinburgh, with facilities in both cities.

E-SCIENCE UK


DCC (Edinburgh)
NeSC (Edinburgh)
& eSI
NeSC (Glasgow)
DCC (Glasgow)
BeSC (Queens University Belfast)
NEReSC (Newcastle)
The White Rose Grid
e-Science Centre
(Leeds)
Lancaster University Centre for e-Science
The White Rose Grid
e-Science Centre (York)
NCeSS
eSNW (Manchester)
STFC e-Science Centre (Daresbury)
DCC (STFC Daresbury)
The White Rose Grid
e-Science Centre (Sheffield)
East Midlands e-Science Centre (AVRRC Loughborough)
Leicester
e-Science Centre
Midlands e-Science Centre (Birmingham)
OeRC
(Oxford)
CeSC (Cambridge)
NIEeS
WeSC (Cardiff)
Grid Support
Centre (RAL)
DCC (STFC RAL)
Reading e-Science
Centre
DCC
(UKOLN
Bath)
UCL (London)
Imperial College
Internet Centre
(London)
Microsoft Institute for High Performance Computing
(Southampton)
OMII-UK
(Southampton)

E-SCIENCE UK MEMBERS

- White Rose Grid e-Science Centre (WRGeSC)
- Belfast e-Science Centre (BeSC)
- Cambridge e-Science Centre (CeSC)
- STFC e-Science Centre (STFCeSC)
- e-Science North West (eSNW)
- National Grid Service (NGS)
- OMII-UK
- Lancaster University Centre for e-Science
- London e-Science Centre (LeSC)
- North East Regional e-Science Centre (NEReSC)
- Oxford e-Science Centre (OeSC)
- Southampton e-Science Centre (SeSC)
- Welsh e-Science Centre (WeSC)

E-SCIENCE UK PROJECTS

- GRIDPP (PPARC)
- ASTROGRID (PPARC)
- Comb-e-Chem (EPSRC)
- DAME (EPSRC)
- DiscoveryNet (EPSRC)
- GEODISE (EPSRC)
- myGrid (EPSRC)
- RealityGrid (EPSRC)
- Climateprediction.com (NERC)
- Oceanographic Grid (NERC)
- Molecular Environmental Grid (NERC)
- NERC DataGrid (NERC + OST-CP)
- Biomolecular Grid (BBSRC)

E-SCIENCE US

- The term **CyberInfrastructure** is typically used to define e-Science projects.
- The term is first used by the US National Science Foundation (NSF).
- Funded by the National Science Foundation Office of CyberInfrastructure (NSF OCI) and Department of Energy.

E-SCIENCE US

- “**Cyberinfrastructure** is the coordinated aggregate of software, hardware and other technologies, as well as human expertise, required to support current and future discoveries in science and engineering.” - *Francine Berman, San Diego Supercomputer Center.*
- “**Cyberinfrastructure** consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible.” - *Craig Stewart, Indiana University.*

E-SCIENCE US PROJECTS

- TeraGrid
  - combining resources at eleven partner sites.
  - started in 2001 and operated from 2004 through 2011 led by University of Chicago.
- nanoHUB
  - gateway comprise community-contributed resources and geared toward educational applications, professional networking, and interactive simulation tools for nanotechnology.
  - product of the Network for Computational Nanotechnology (NCN), a multi-university initiative of eight member institutions.
- iPlant Collaborative
  - create virtual organization for the plant sciences.

E-SCIENCE US PROJECTS

- Open Science Grid Consortium
- Datanet
- National Center for Supercomputing Applications
- National LambdaRail and Internet2
- ICME cyberinfrastructure

E-SCIENCE NL

- Netherlands eScience Center in Amsterdam, founded by NWO and SURF.
- Stimulate creative data-driven research across all scientific disciplines
- Develop and apply tools to enable data-intensive scientific research
- Promote knowledge-based collaboration between cross-disciplinary researchers.

E-SCIENCE NL PROJECTS

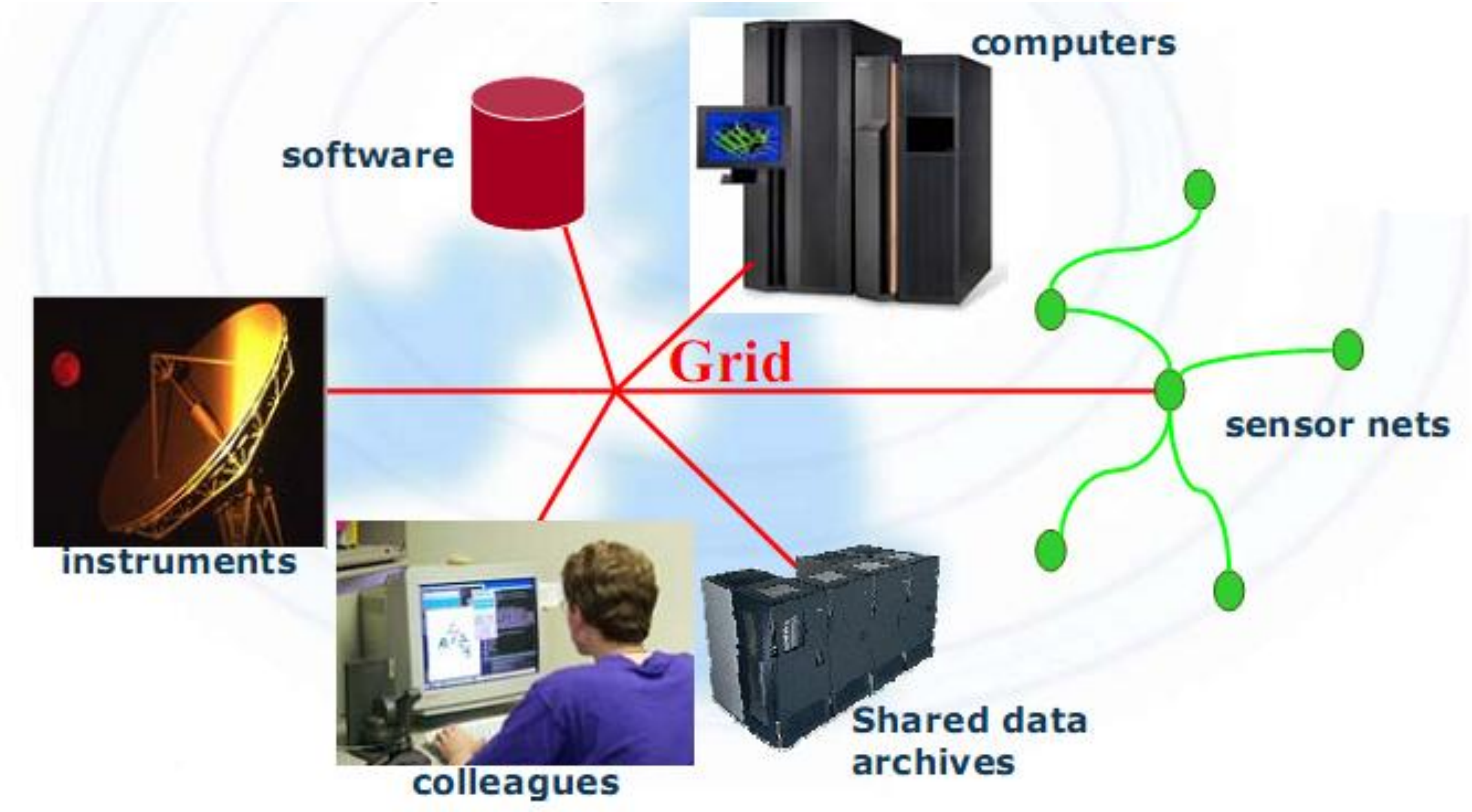
- **Astronomy**: eScience Infrastructure for Huge Interferometric Datasets
- **Climate Research**: eScience Approach to Determine Regional Sea-Level Variability
- **Cognition**: Biomarker Boosting through eScience Infrastructure
- **eChemistry**: Computational Metabolite Identification and Biochemical Network Reconstruction for Integrative Metabolomics Data Analysis
- **eEcology**: Virtual Labs eEcology
- **eHumanities**: BiographyNED

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

E-SCIENCE INFRASTRUCTURE

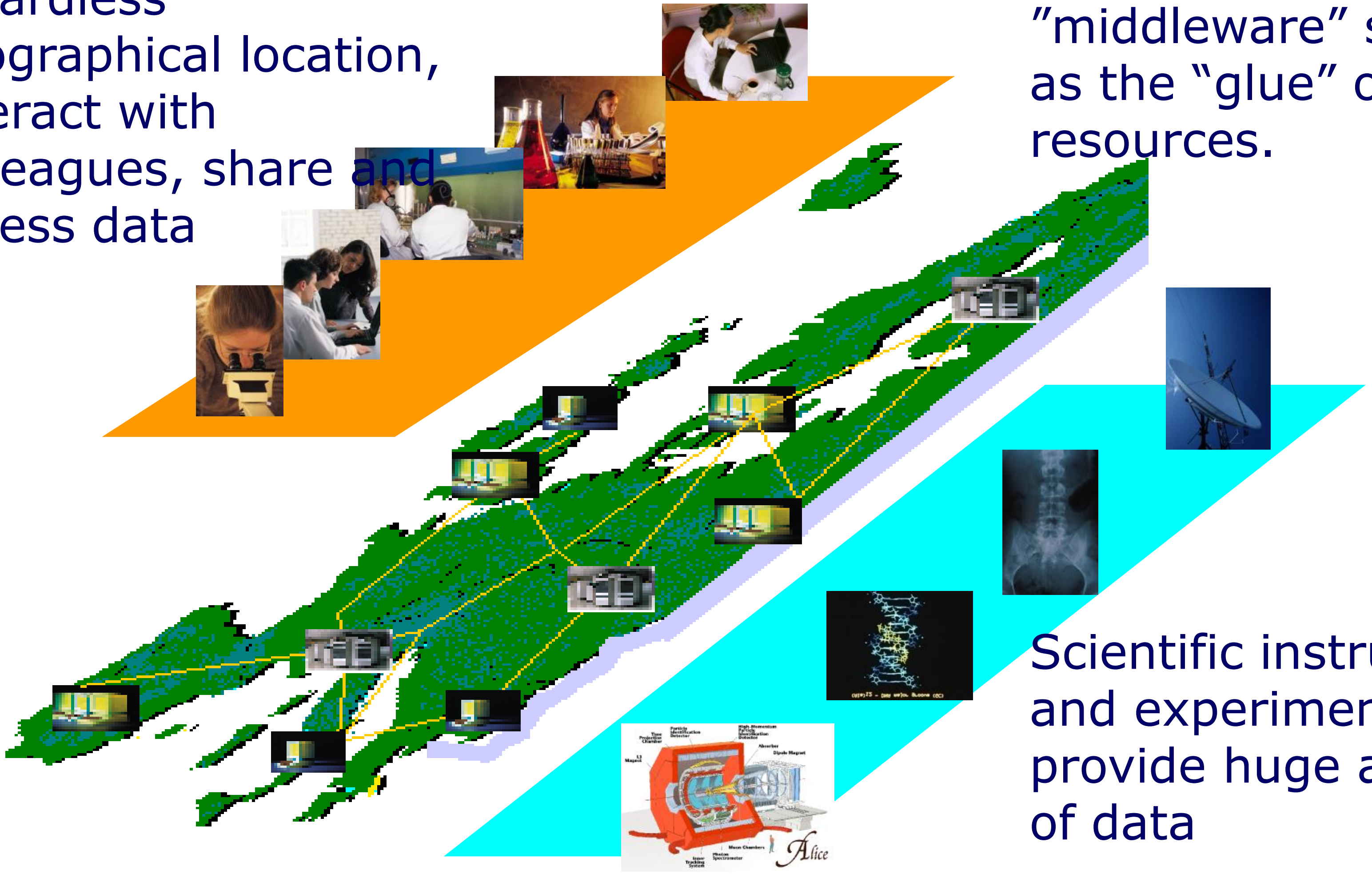
- Grids: a foundation for e-Science



E-SCIENCE INFRASTRUCTURE

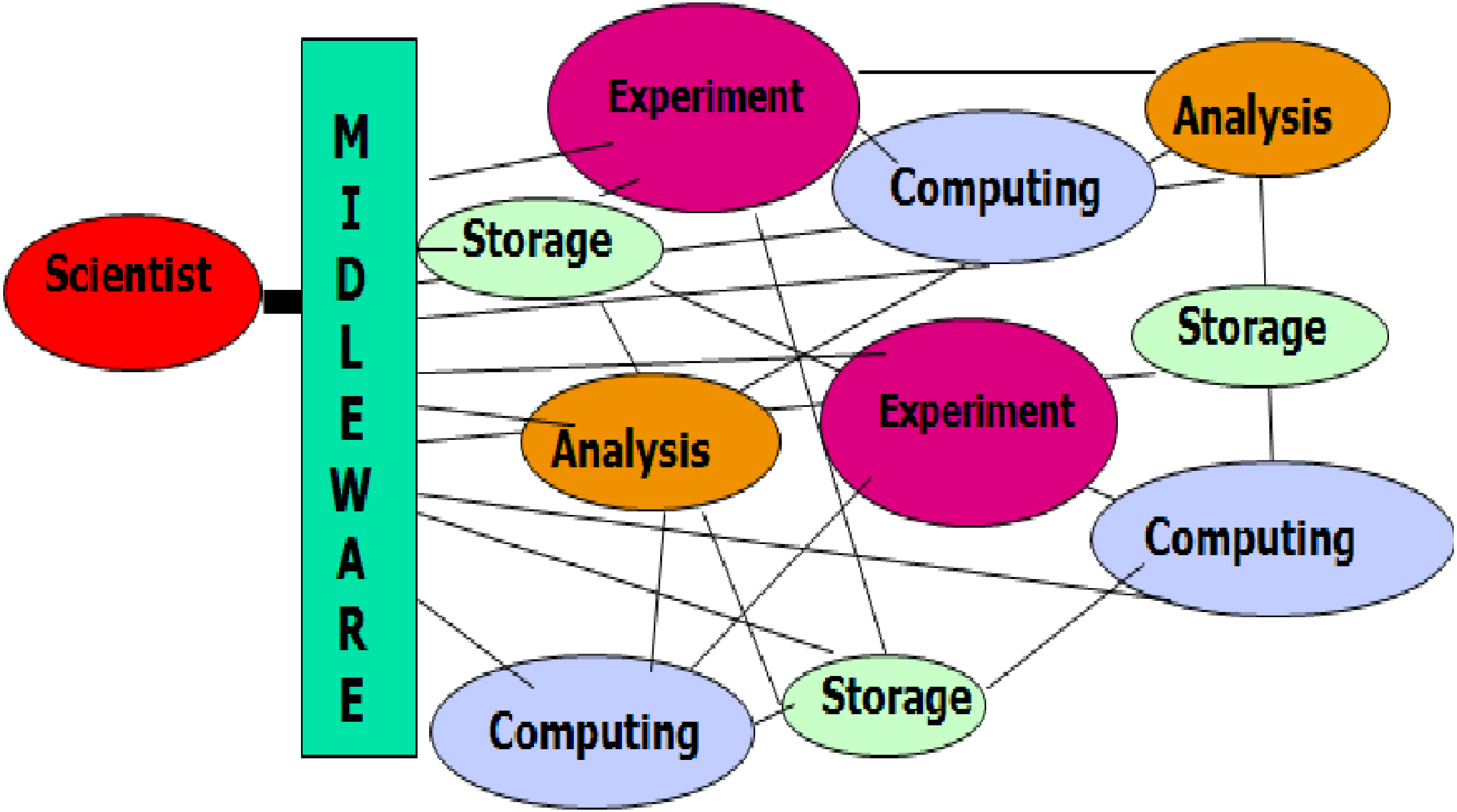
Researchers perform their activities regardless geographical location, interact with colleagues, share and access data

The Grid: networked data processing centres and ”middleware” software as the “glue” of resources.



Scientific instruments and experiments provide huge amount of data

E-SCIENCE INFRASTRUCTURE



KEY COMPONENTS: DATA

- Probably the most important component, as many applications involve huge amounts of data spread across multiple sites
  - 100TB to 1PB+ in high-end applications
  - often too big to be stored in one place, must be distributed
  - user wants to treat as a single filesystem or database irrespective of actual location
  - raises issues of file distribution, replication, caching and archiving
  - implications for networking requirements

KEY COMPONENTS: COMPUTE RESOURCES

- In the Grid computing model, the “system” keeps track of available compute resources. The user specifies the task to be performed, and the “system” decides where best to carry it out.
- The usual analogy is to electricity provision – the user doesn’t know/care which power station is actually generating the electricity.
- eScience projects will know exactly which resources they are using, using distributed resource managers like Grid Engine or Condor to handle clusters.

KEY COMPONENTS: NETWORKING

- The co-scheduling requirements may extend to reserving network bandwidth.
- This requires networking which offers QoS (Quality of Service) guarantees.
- Also need network instrumentation to make intelligent decisions (move data to a fast machine which is "far away" or a slower machine which is "closer"?)
- In practice, networking is often the bottleneck.

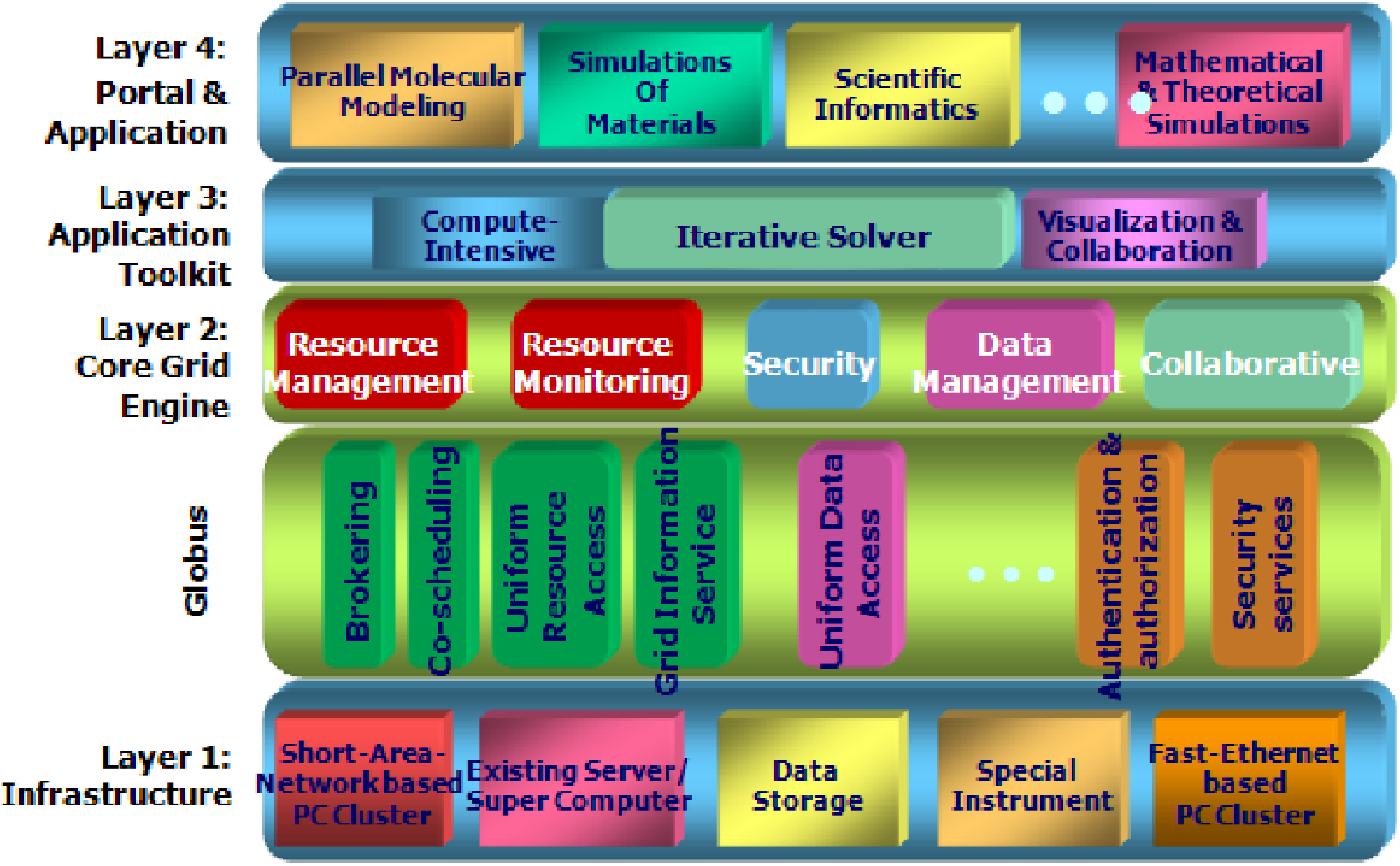
KEY COMPONENTS: SECURITY

- The three main components of security are:
  - authentication – verifying identity of users and machines
  - authorization – deciding what they are allowed to do
  - encryption – providing secure communication

KEY COMPONENTS: COLLABORATIVE WORKING

- Collaborative working is fundamental to eScience
  - video-conferencing.
  - remote/distributed visualisation

E-SCIENCE GRID BASED FRAMEWORK



Layer 4: Portal & Application

Parallel Molecular Modeling

Simulations Of Materials

Scientific Informatics

• • •

Mathematical & Theoretical Simulations

Layer 3: Application Toolkit

Compute-Intensive

Iterative Solver

Visualization & Collaboration

Layer 2: Core Grid Engine

Resource Management

Resource Monitoring

Security

Data Management

Collaborative

Globus

Brokering

Co-scheduling

Uniform Resource Access

Grid Information Service

Uniform Data Access

• • •

Authentication & authorization

Security services

Layer 1: Infrastructure

Short-Area-Network based PC Cluster

Existing Server/ Super Computer

Data Storage

Special Instrument

Fast-Ethernet based PC Cluster

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

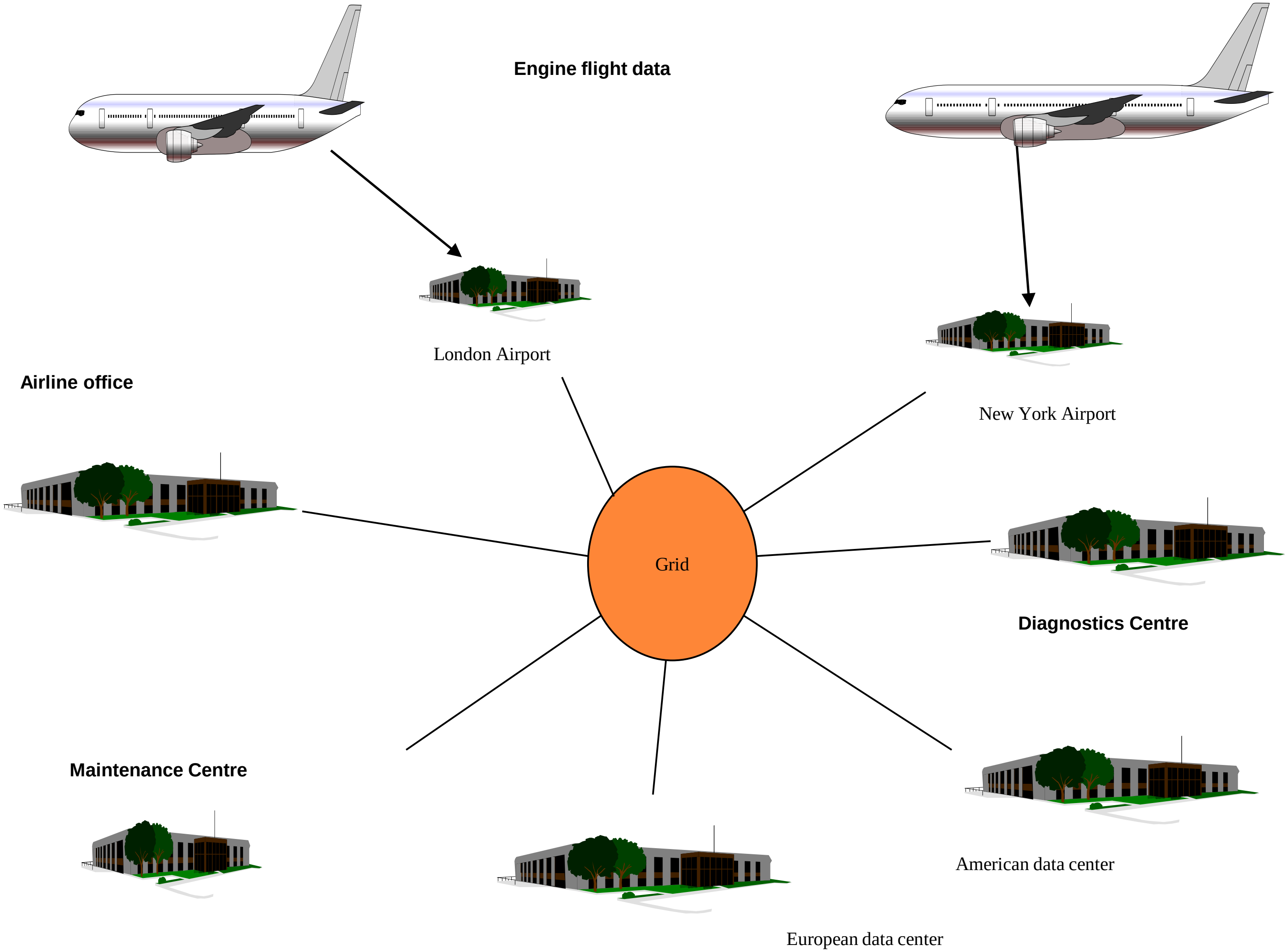
DAME

- DAME - Distributed Aircraft Maintenance Environment
- an e-Science pilot project, demonstrating the use of the GRID to implement a distributed decision support system for deployment in maintenance applications and environments.
- collaborative project by Oxford, Leeds, York and Sheffield with active support from main companies: Rolls-Royce, DS&S, Cybula.
- aim is to gather engine diagnostic data and make (real-time) decisions about engine maintenance.

DAME

- data flows from aircraft to airline to Rolls-Royce
- advice/decisions flow in reverse direction
- unusual symptoms can be queried against vast database
- timely preventive maintenance can greatly
- reduce costs and flight cancellations

DAME – Distributed Aircraft Maintenance Environment



Engine flight data

London Airport

New York Airport

Airline office

Grid

Diagnostics Centre

Maintenance Centre

American data center

European data center

EDIAMOND

Aiming to prove the benefits of grid technology to breast Imaging in the UK.

EDIAMOND

- “One of the pilot e-science projects is to develop a digital mammographic archive, together with an intelligent medical decision support system for breast cancer diagnosis and treatment. An individual hospital will not have supercomputing facilities, but through the grid it could buy the time it needs. So the surgeon in the operating room will be able to pull up a high-resolution mammogram to identify exactly where the tumour can be found." - *Tony Blair*

EDIAMOND

- eDiamond – developing a national database of mammographic images.
- collaboration between Oxford, Mirada (spin-off company) and IBM
- mentioned in Tony Blair's speech
- again a prototype, but with huge potential

EDIAMOND

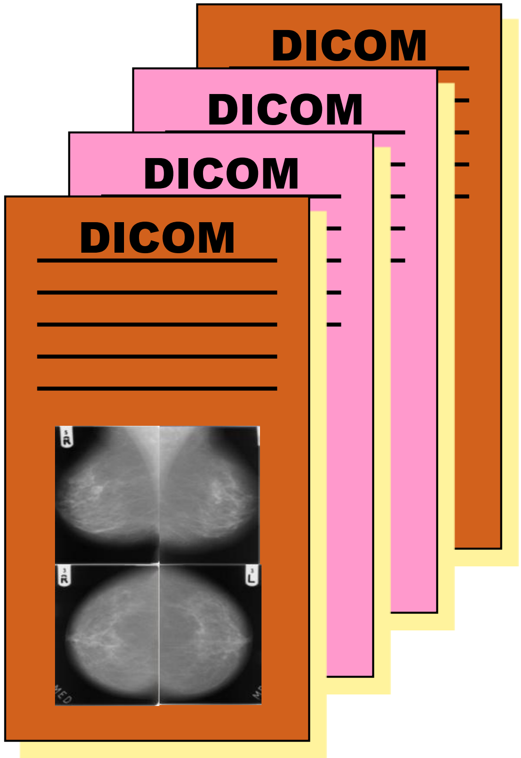
- a huge distributed database will hold images for which the outcome (cancerous or not) is known
- new images can be compared against others ("find one like this") to improve diagnosis
- can also be used to train radiologists, and possibly even develop expert systems to handle the simplest diagnoses

EDIAMOND: DATA

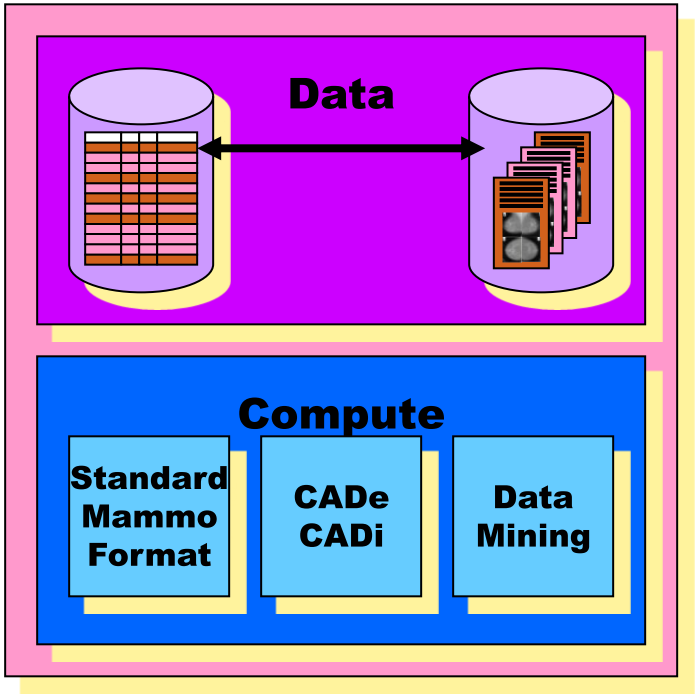
Data

Patient	Age	…	Image
107258	55	…	1.dcm
236008	62	…	2.dcm
700266	59	…	3.dcm
895301	58	…	4.dcm
………	…	…	……..
………	…	…	……..
………	…	…	……..
………	…	…	……..
………	…	…	……..
………	…	…	……..
………	…	…	……..
………	…	…	……..

Images



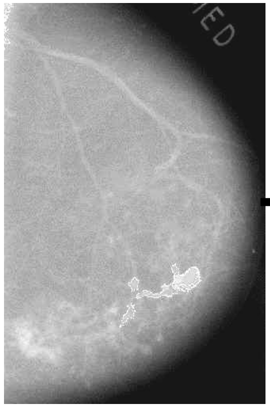
Grid



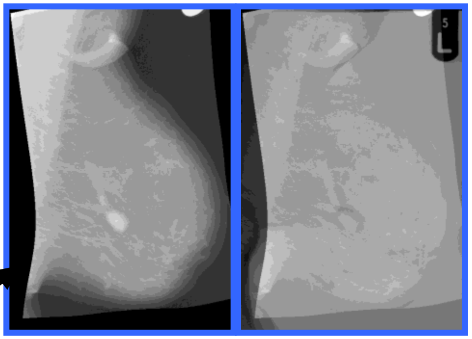
The Logical View of this information is as a Single Resource

EDIAMOND: COMPUTATION

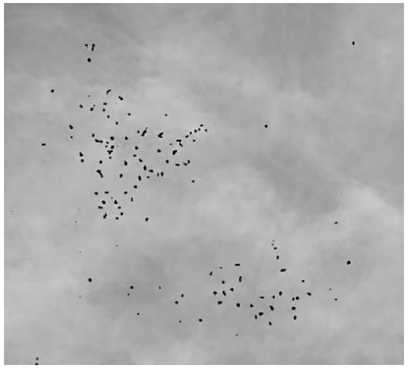
Mammograms have different appearances, depending on image settings and acquisition systems



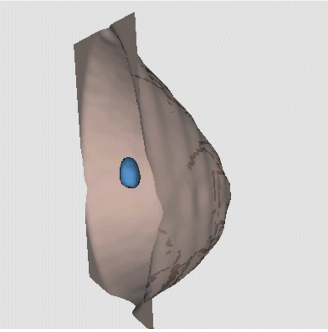
Standard Mammo Format



Temporal mammography



Computer Aided Detection



3D View

Compute power can address several issues

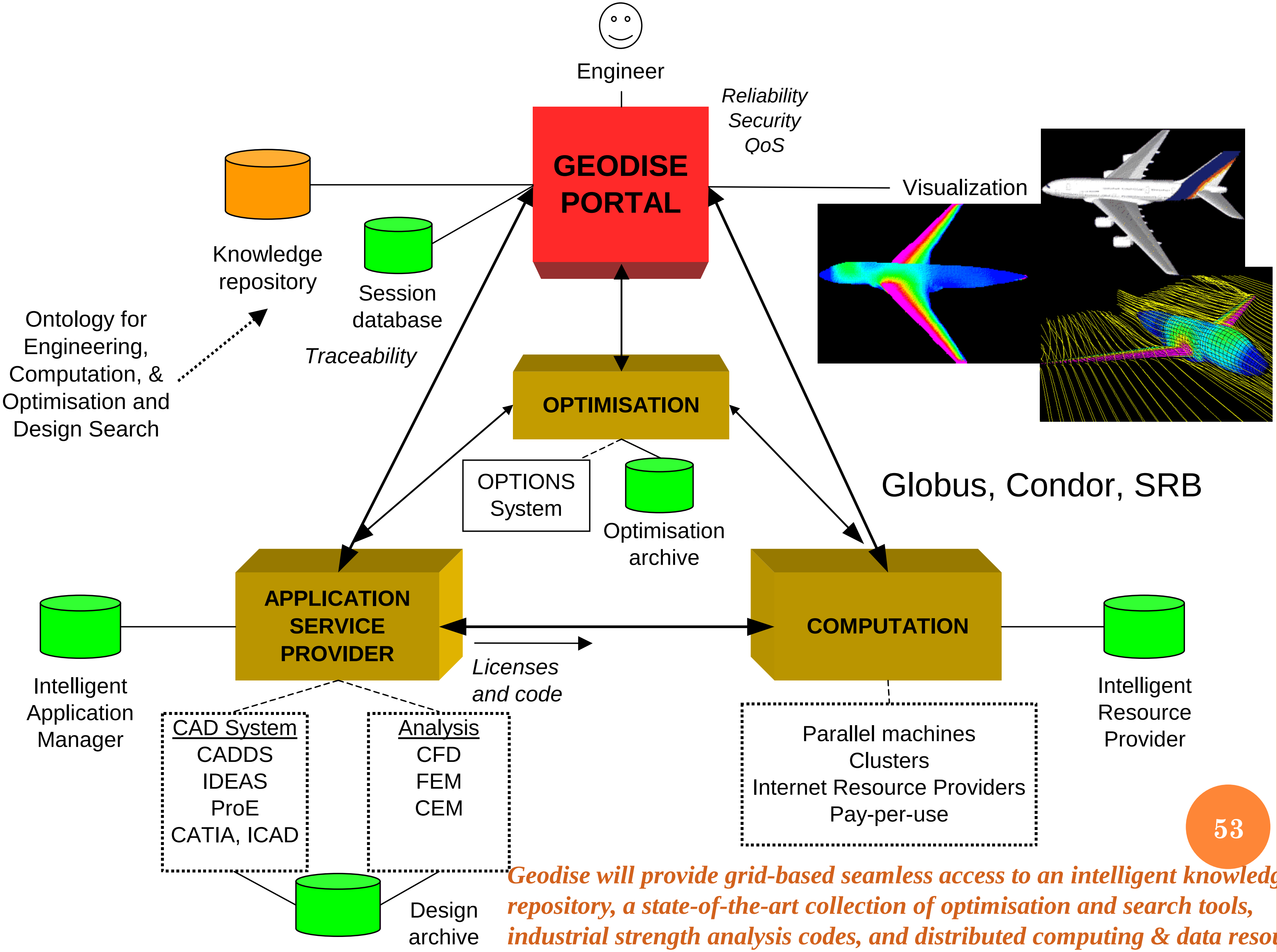
GEODISE

- Grid-Enabled Optimisation and Design Search for Engineering
- collaboration between Oxford, Southampton and Manchester, with support from Rolls-Royce, BAESystems, Intel, Microsoft and others
- exemplar application is aerodynamic inlet design to reduce noise emission, but relevant to any large-scale engineering design, especially involving multiple companies

GEODISE

- using grid technologies with distributed databases and computations
- modular workflow management is a critical component
- also a strong emphasis on knowledge capture and management

GEODISE



Geodise will provide grid-based seamless access to an intelligent knowledge repository, a state-of-the-art collection of optimisation and search tools, industrial strength analysis codes, and distributed computing & data resources

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

THE FUTURE HOLDS FOR E-SCIENCE

- Innovation
  - It is the general consensus that the technology of tomorrow must be ready to meet the inspirational thinking of scientists.
- Business
  - There is a desire not only to make the technology of e-Science available to scientists, but also commercial entities, such as engineers.
- Collaboration
  - Partnership is a vital element to the development of better storage facilities and the enhancement of Grid infrastructures.

THE FUTURE HOLDS FOR E-SCIENCE

- Complex ideas
  - Scientists are trying to research things that they have never even touched upon before. E-Science and its conglomerates are well aware that advances in information technology are the only way forward for the advancement of science.
- Education
  - If e-Science is to improve its image and further its impact upon science, then it is essential that the students of tomorrow are trained in the use of advanced computing technology.
- International development
  - It is essential to the future success of e-Science that its methods and technology are used across the globe, not just within UK.

OUTLINE

- What is e-Science?
- The aim of e-Science
- Main e-Science groups
- e-Science infrastructure
- Some e-Science projects
- The future holds for e-Science
- Conclusion

CONCLUSION

- "eScience will change the dynamic of the way science is undertaken" - *John Taylor, Director General of the Research Councils (2000)*
- "[The eScience grid] intends to make access to computing power, scientific data repositories and experimental facilities as easy as the web makes access to information. " - *Tony Blair, October 2002*

REFERENCES

- eScience and Grid Computing, *Prof. Mike Giles, Oxford Centre for Computational Finance City Seminar, 12/02/03*
- What is Grid Computing?, *Richard Hopkins, NGS Induction – Rutherford Appleton Laboratory, 2nd / 3rd November 2005*
- What is e-Science and Grid computing? *Dave Berry, NeSC*
- The Encyclopedia Wikipedia, http://en.wikipedia.org/wiki/
- NSF office of Cyberinfrastructure, http://www.nsf.gov/
- E-Science Grid, http://www.escience-grid.org.uk/
- Distributed Aircraft Maintenance Environment , http://www.cs.york.ac.uk/dame/
- eDiaMoND, http://www.ediamond.ox.ac.uk/
- Grid Enabled Optimisation and Design Search for Engineering (GEODISE), http://www.geodise.org

THANK YOU!