NCESS
International Conference
Manchester
Progress with e-Science?

Malcolm Atkinson
e-Science Institute
& e-Science Envoy

www.nesc.ac.uk
29th June 2006
Overview

- **A Context**
  - E-Science, Service-oriented Architectures & Services
  - International & UK Scene
  - Projects, Activities & Technologies

- **A Wealth of Opportunities & Challenges**
  - Causes of Data Growth
  - Interpretational challenges
  - Software challenges

- **Crucial Issues**
  - Usability & Abstraction
  - Interoperation & Federations
  - Cultural challenges of e-Science
Principles & Global Context
What is e-Science?

- **Goal:** to enable better research in *all* disciplines
- **Method:** Develop *collaboration* supported by advanced distributed computation
  - to generate, curate and analyse rich data resources
    - From experiments, observations and simulations
    - Quality management, preservation and reliable evidence
  - to develop and explore models and simulations
    - Computation and data at all scales
    - Trustworthy, economic, timely and relevant results
  - to enable *dynamic* distributed collaboration
    - Facilitating collaboration with information and resource sharing
    - Security, trust, reliability, accountability, manageability and *agility*

Our challenge is to develop an integrated approach to *all* three
Are we making any progress?

- **General purpose e-Infrastructure**
  - Supports most research requirements
  - Convenient & reliable
  - Amortises costs over wide communities

- **General purpose e-Science methods**
  - Conduct of multi-disciplinary collaboration
  - Distributed collaboration
  - Use of diverse information sources
  - Use of advanced computational methods
Web Services

- Generic & Industry Supported
- Mechanism for presenting any resource so it can be used remotely
  - Data resources
  - Computation resources
  - Instruments
  - Research processes & procedures
- Wide range of choices as to how it may be done
  - Easy for provider to make localised decisions
  - Quick deployment
  - Onus on users to compose services
  - Little opportunity to amortise costs across services
- An obvious first step & components of e-Infrastructure
A Grid Computing Timeline

1995
- I-Way: SuperComputing '95

'96
- US Grid Forum forms at SC '98

'97
- European & AP Grid Forums

'98
- Grid Forums merge, form GGF

'99
- "Anatomy" paper

2000
- "Physiology" paper

'01
- OGSA-WG formed

'02
- GGF & EGA form OGF

'03
- Japan government funds:
  - Business Grid project
  - NAREGI project

'04
- DARPA funds Globus Toolkit & Legion

'05
- EU funds UNICORE project

2006
- US DoE pioneers grids for scientific research

- NSF funds National Technology Grid

- NASA starts Information Power Grid

Today:
- Grid solutions are common for HPC
- Grid-based business solutions are becoming common
- Required technologies & standards are evolving

Source: Hiro Kishimoto GGF17 Keynote May 2006
What is a Grid?

A grid is a system consisting of
- Distributed but connected resources and
- Software and/or hardware that provides and manages logically seamless access to those resources to meet desired objectives.

Source: Hiro Kishimoto GGF17 Keynote May 2006
Grid & Related Paradigms

**Distributed Computing**
- Loosely coupled
- Heterogeneous
- Single Administration

**Cluster**
- Tightly coupled
- Homogeneous
- Cooperative working

**Grid Computing**
- Large scale
- Cross-organizational
- Geographical distribution
- Distributed Management

**Utility Computing**
- Computing “services”
- No knowledge of provider
- Enabled by grid technology

Source: Hiro Kishimoto GGF17 Keynote May 2006
How Are Grids Used?

- High-performance computing
- Collaborative design
- Drug discovery
- Data center automation
- Financial modeling
- Collaborative data-sharing
- High-energy physics
- Life sciences

Source: Hiro Kishimoto GGF17 Keynote May 2006
Grids In Use: E-Science Examples

• Data sharing and integration
  - Life sciences, sharing standard data-sets, combining collaborative data-sets
  - Medical informatics, integrating hospital information systems for better care and better science
  - Sciences, high-energy physics

• Simulation-based science and engineering
  - Earthquake simulation

• Capability computing
  - Life sciences, molecular modeling, tomography
  - Engineering, materials science
  - Sciences, astronomy, physics

• High-throughput, capacity computing for
  - Life sciences: BLAST, CHARMM, drug screening
  - Engineering: aircraft design, materials, biomedical
  - Sciences: high-energy physics, economic modeling

Source: Hiro Kishimoto GGF17 Keynote May 2006
Grids

- (Potentially) Generic & Industry Supported
- Mechanism for presenting *many heterogeneous* resources so they can be used remotely
  - Data resources
  - Computation resources
  - Instruments
  - Research processes & procedures
- *Restricting* choices as to how it may be done
  - *Harder* for provider to make localised decisions
  - Deployment can be challenging
- Providing more homogeneity through virtualisation
  - *Should* be easier to compose services
  - *More* opportunity to amortise costs
- A component of e-Infrastructure
How much do Grids help a researcher?

- The grid *per se* doesn’t provide
  - Supported e-Science methods
  - Supported data & information resources
  - Computations
  - Convenient access

- Grids help providers of these
  - International & national secure e-Infrastructure
  - Standards for interoperation
  - Standard APIs to promote re-use

- But Research Support must be built
  - What is needed?
  - Who should do it?
UK Context
Commitment to e-Infrastructure

- A shared resource
  - That enables science, research, engineering, medicine, industry, ...
  - It will improve UK / European / ... productivity
    - Lisbon Accord 2000
- Commitment by UK government
  - Sections 2.23-2.25
- Always there
  - c.f. telephones, transport, power

Science & innovation investment framework
2004 - 2014

July 2004

Gordon Brown  Charles Clarke  Patricia Hewitt
Chancellor of the Exchequer  Secretary of State for Education and Skills  Secretary of State for Trade and Industry
The e-Science On The Map Today

- Globus Apache Project & CDIG
- National Centre for e-Social Science
- Digital Curation Centre
- NERC e-Science Centre
- OMII-UK
- e-Science Institute
- Funded centres
- National Grid Service
- NGS Support Centre
- National Institute for Environmental e-Science
- EGEE-II
- NGS
And much more besides

- Supported data & information services
  - E.g. EBI, MIMAS, EDINA, Zoo-DB, ...
- Community efforts
  - E.g. GridPP, AstroGrid, NERC DataGrid
- Access to new scientific facilities
  - E.g. Satellite observations, Met-Office data
  - E.g. Diamond
  - E.g. CEASAR, HPCx & HECTOR
- Communities developing code & tools
- Pilot projects in most research areas
A Complex multi-provider System

- **Strengths**
  - Resilience - no over-dependence on one set of decisions
  - Dynamic & rapid response to need
  - Resources gathered through many channels
  - Specific user requirements understood

- **Weaknesses**
  - Complexity & possibly unnecessary duplication
  - Reinforces old silos
  - Composition and amortisation harder

- **Is more coordination**
  - Desirable?
  - Feasible?
OMII-UK & National Grid Service: Life Sciences Gateway

Talk to us about other ‘gateways’
- Computational Chemistry
- Engineering
- Image Analysis
- …
The National Grid Service

Core sites
- White Rose (Leeds)
- Manchester
- Oxford
- CCLRC

Partner sites
- Bristol
- Cardiff
- Lancaster

Access to HPC facilities
- HPCx
- CSAR

Capacity
- 300 + CPUs
- 30+ Tera Bytes

Specialist facilities
- Cardiff 4x16 proc SGI
- Bristol: Intel
- Lancaster SUN Cluster

Services:
- Job submission (GT2)
- Storage service (SRB)
- Oracle Service
- OGSA-DAI
Related projects & collaborations are where the future expansion of resources will come from.

<table>
<thead>
<tr>
<th>Project</th>
<th>Anticipated resources (initial estimates)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Related Infrastructure projects</strong></td>
<td></td>
</tr>
<tr>
<td>SEE-grid</td>
<td>6 countries, 17 sites, 150 cpu</td>
</tr>
<tr>
<td>EELA</td>
<td>5 countries, 8 sites, 300 cpu</td>
</tr>
<tr>
<td>EUMedGrid</td>
<td>6 countries</td>
</tr>
<tr>
<td>BalticGrid</td>
<td>3 countries, fewx100 cpu</td>
</tr>
<tr>
<td>EUChinaGrid</td>
<td>TBC</td>
</tr>
<tr>
<td><strong>Collaborations</strong></td>
<td></td>
</tr>
<tr>
<td>OSG</td>
<td>30 sites, 10000 cpu</td>
</tr>
<tr>
<td>ARC</td>
<td>15 sites, 5000 cpu</td>
</tr>
<tr>
<td>DEISA</td>
<td>Supercomputing resources</td>
</tr>
</tbody>
</table>
Use of the infrastructure

Massive data transfers > 1.5 GB/s

Sustained & regular workloads of >30K jobs/day
- spread across full infrastructure
- doubling/tripling in last 6 months – no effect on operations

Statistics:
- Submitted: 617
- Waiting: 05
- Ready: 046
- Scheduled: 9566
- Running: 12166
- Done: 6427
- Aborted: 5104
- Cancelled: 88
- Active Sites: 156 : 33903
A Knowledge Transfer Network project funded by the DTI Technology Programme aimed at transferring knowledge about Grid Computing Technologies to Public and Private Sectors in the UK.

Partnership between Intellect, the UK Hi-Tech Trade Association; National e-Science Centre, a world leader in Grid Computing research; and CNR Ltd, a consultancy focused on SME organisations and business intermediaries.

Substantial number of industrial, business and academic partners

Website
- Background Information
- Industry News/Events
- User Case Studies

Events programme
- Technical Overviews
- Multiple vendor perspectives
- User Case Studies

Sector Agenda
- Healthcare; Government;
  Telecoms; Services; etc..

User Community
- Network with peers
- Find useful contacts
- Contribute experience

www.gridcomputingnow.org

Funded by Government, Regional Development Agencies, Devolved Administrations and Research Councils
Invest in People

- **Training**
  - Targeted
  - Immediate goals
  - Specific skills
  - Building a workforce

- **Education**
  - Pervasive
  - Long term and sustained
  - Generic conceptual models
  - Developing a culture

- **Both are needed**
Collaboration is the key to e-Science
Biomedical Research Informatics Delivered by Grid Enabled Services

Portal

Syteny Grid Service

http://www.brc.dcs.gla.ac.uk/projects/bridges/
eDiaMoND: Screening for Breast Cancer

1 Trust → Many Trusts
Collaborative Working
Audit capability
Epidemiology

Provided by eDiamond project: Prof. Sir Mike Brady et al.
climateprediction.net and GENIE

- Largest climate model ensemble
- >45,000 users, >1,000,000 model years

Response of Atlantic circulation to freshwater forcing
Integrative Biology

Tackling two Grand Challenge research questions:

• What causes heart disease?
• How does a cancer form and grow?

Together these diseases cause 61% of all UK deaths

Will build a powerful, fault-tolerant Grid infrastructure for biomedical science

Enabling biomedical researchers to use distributed resources such as high-performance computers, databases and visualisation tools to develop complex models of how these killer diseases develop.

Courtesy of David Gavaghan & IB Team
IB Partners

Courtesy of David Gavaghan & IB Team
Foundations of Collaboration

- Strong commitment by individuals
  - To work together
  - To take on communication challenges
  - Mutual respect & mutual trust
- Distributed technology
  - To support information interchange
  - To support resource sharing
  - To support data integration
  - To support trust building
- Sufficient time
- Common goals
- Complementary knowledge, skills & data

Can we predict when it will work? Can we find remedies when it doesn’t?
New projects in e-Science
Understanding the brain may be the greatest informatics challenge of the 21st century.

- determining ion channel contribution to the timing of action potentials
- resolving the ‘neural code’ from the timing of action potential activity
- examining integration within networks of differing dimensions
CARMEN Consortium

Leadership & Infrastructure

Colin Ingram

Paul Watson

Leslie Smith

Jim Austin
CARMEN Consortium

International Partners

Ad Aertsen
(Freiburg)
Neural network modelling
and large-scale simulations

George Gerstein
(Pennsylvania)
Analysis of spike pattern trains

Sten Grillner
(Karolinska Institute)
Chairman of the OECD,
International Neuroinformatics
Coordinating Facility

Shiro Usui
(RIKEN Brain Science Institute)
Lead for the Japan Node of the
International Neuroinformatics
Coordinating Facility

Daniel Gardner
(Cornell)
Lead for the US NIH,
Neuroscience Information
Framework and Brain ML
CARMEN Consortium

Commercial Partners

- applications in the pharmaceutical sector
- interfacing of data acquisition software
- application of infrastructure
- commercialisation of tools
The Challenge

International Technology Roadmap for Semiconductors

<table>
<thead>
<tr>
<th>Year</th>
<th>2005</th>
<th>2010</th>
<th>2015</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPU Half Pitch (nm)</td>
<td>90</td>
<td>45</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>MPU Gate Length (nm)</td>
<td>32</td>
<td>18</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

Device diversification

90nm: HP, LOP, LSTP

45nm: UTB SOI

32nm: Double gate

Device diversification

90nm: HP, LOP, LSTP

45nm: UTB SOI

32nm: Double gate

25 nm

FinFET
UTB SOI
FD SOI
Bulk MOSFET
LSTP
LOP
HP(MPU)
Stat. Sets

2005 edition Toshiba 04
University Partners

Advanced Processor Technologies Group (APTGUM)
Device Modelling Group (DMGUG)
Electronic Systems Design Group (ESDGUS)
Intelligent Systems Group (ISGUY)
National e-Science Centre (NeSC)
Microsystems Technology Group (MSTGUG)
Mixed-Mode Design Group in IMNS (MMDGUE)
e-Science NorthWest Centre (eSNW)

16th March 2006
Industrial Partners

Global EDS vendor and world TCAD leader
600 licences of grid implementation, model implementation

UK fabless design company and world microprocessor leader
Core IP, simulation tools, staff time

UK fabless design company and world mixed mode leader
Additional PhD studentship for mixed mode design

Global semiconductor player with strong UK presence
Access to technology, device data, processing

Global semiconductor player with strong UK presence
Access to technology, device data, processing

Global semiconductor player with UK presence
CASE studentship, interconnects

Trade association of the microelectronics industry in the UK
Recruiting new industrial partners and dissemination
Data Deluge
Compound Causes of (Geo)Data Growth

- Faster devices
- Cheaper devices
- Higher-resolution
  - all ~ Moore’s law
- Increased processor throughput
  - ⇒ more derived data
- Cheaper & higher-volume storage
- Remote data more accessible
  - Public policy to make research data available
  - Bandwidth increases
  - Latency doesn’t get less though
Interpretational Challenges

- Finding & Accessing data
  - Variety of mechanisms & policies

- Interpreting data
  - Variety of forms, value systems & ontologies

- Independent provision & ownership
  - Autonomous changes in availability, form, policy, ...

- Processing data
  - Understanding how it may be related
  - Devising models that expose the relationships

- Presenting results
  - Humans need either
    - Derived small volumes of statistics
    - Visualisations
Interpretational Challenges

- **Finding & Accessing data**
  - Variety of mechanisms & policies

- **Interpreting data**
  - Variety of forms, value systems & ontologies

- **Independent provision & ownership**
  - Autonomous changes in availability, form, policy, ...

- **Processing data**
  - Understanding how data may be related
  - Devising models that expose the relationships

- **Presenting results**
  - Humans need either
    - Derived small volumes of statistics
    - Visualisations
Interpretational Challenges

- Finding & Accessing data
  - Variety of mechanisms & policies
- Interpreting data
  - Variety of forms, value systems & ontologies
- Independent provision & ownership
  - Autonomous changes in availability, form, policy, ...
- Processing data
  - Understanding how it may be related
  - Devising models that expose the relationships
- Presenting results
  - Humans need either
    - small volumes of statistics
    - visualisations
Next steps in e-Science
Collaboration

- Essential to assemble experts
  - Multi-discipline, Multi-organisation, Multi-national, Multi-scale
- Hard to achieve
  - Instinctive competition
  - Trust slow to build
  - Communication is difficult
- Requirements
  - Leadership
  - Investment
  - New culture
  - Technology
  - Cross commercial - academic boundaries

Address these issues

Focus here
Towards Accessible e-Science

- High-level tools
  - Abstraction
  - Metadata-driven interpretation & guidance
  - Well-defined semantics
  - Automated data management is key

- Convenient user-controlled composition
  - Lego-brick convenience + Precision control
  - Understood by scientists, engineers, modellers, diagnosticians, ...

- Responsibility & Credit
  - Provenance tracking automated & understood
  - Culture developed for consortia and collaboration
Built on Dependable Infrastructure

- Global Federations
  - Dependable and persistent facilities
  - Always there & always on
  - Consistent for mobile users
  - Consistent for mobile code & mobile information

- Affordable e-Infrastructure
  - Based on well-established standards
  - Based on well-honed operating procedures
  - Investment preserved through stability
  - Utility improved through agile development
  - Amortised across disciplines and nations
  - Consistent framework for multiple provisions

- Trustworthy management of information
Collaboration

- Collaboration is a Key Issue
  - Multi-disciplinary
  - Multi-national
  - Academia & industry

- Trustworthy data sharing key for collaboration
  - Plenty of opportunities for research and innovation
  - Establish common frameworks where possible
    - Islands of stability - reference points for data integration
  - Establish international standards and cooperative behaviour
    - Extend incrementally

- Trustworthy code & service sharing also key
Federation

- Federation is a Key Issue
  - Multi-organisation
  - Multi-purpose
  - Multi-national
  - Academia & industry
- Build shared standards and ontologies
  - Requires immense effort
  - Requires critical mass of adoption
  - Requires sufficient time
- Trustworthy code & e-Infrastructure sharing
  - Economic & social necessity
  - Balance independence & self-motivation
  - With social & technical constraints
Major Intellectual Challenges

- Require
  - Many approaches to be integrated
  - Many minds engaged
  - Many years of effort

- Using the Systems
  - Requires well-tuned models
  - Well-tuned relationships between systems & people
  - Flexibility, adaptability & agility

- Enabling this
  - Is itself a major intellectual challenge
Matched Systems

- Communities of Researchers
- E-Infrastructure Facilities
- Each will evolve

**e-Social Science is key to progress in all e-Science**
Questions & Comments
Please