Towards an e-Infrastructure for Research and Innovation:
A Progress Report on e-Science

Tony Hey
Director, UK e-Science Core Programme
Outline

1. Licklider’s vision
2. A status report on UK e-Science
3. Web Services and Grids
4. Building a National e-Infrastructure
5. Dual support: the role of JISC
1. Licklider’s Vision

“Lick had this concept – all of the stuff linked together throughout the world, that you can use a remote computer, get data from a remote computer, or use lots of computers in your job.”

Larry Roberts – Principal Architect of the ARPANET
The e-Science Paradigm

- The Integrative Biology Project involves seven UK Universities lead by Oxford and the University of Auckland in New Zealand
  - Models of electrical behaviour of heart cells developed by Denis Noble’s team in Oxford
  - Mechanical models of beating heart developed by Peter Hunter’s group in Auckland

- Researchers need to be able to easily build a secure ‘Virtual Organisation’ providing an international “collaboratory”
  - Will enable new science research
Multiscale modelling of the heart

Multiscale modelling of cancer

Man

Cell
An e-Infrastructure for e-Research

The invention and exploitation of advanced IT to build an information infrastructure to support multidisciplinary and collaborative Research and Innovation:

- to generate, curate and analyse research data
- to develop and explore models and simulations
- to enable the formation of dynamic distributed virtual organisations
The ‘Grid’ is a set of core middleware services running on top of high performance global networks to support research and innovation.
2. A Status Report on UK e-Science

• An exciting portfolio of Research Council e-Science projects
  – Beginning to see e-Science infrastructure deliver some early ‘wins’ in several areas
  – DiscoveryNet success at SC02
  – TeraGyroid success at SC03: ‘heroic’ achievement
  – Astronomy, Chemistry, Bioinformatics, Engineering, Environment, Healthcare ….

• The UK is unique in having a strong collaborative industrial component
  – Nearly 80 UK companies contributing over £30M
  – Engineering, Pharmaceutical, Petrochemical, IT companies, Commerce, Media, …
GridPP prototype Grid

- 1,000 CPUs
  - 500 CPUs at the Tier-1 at RAL
  - 500 CPUs at 11 sites across UK organised in 4 Regional Tier-2s
- 500 TB of storage
- 800 simultaneous jobs

- Integrated with international LHC Computing Grid (LCG)
  - 5,000 CPUs
  - 4,000 TB of storage
  - 70 sites around the world
  - 4,000 simultaneous jobs
- monitored via Grid Operations Centre (RAL)
DAME: Grid based tools and Infer-structure for Aero-Engine Diagnosis and Prognosis

“A Significant factor in the success of the Rolls-Royce campaign to power the Boeing 7E7 with the Trent 1000 was the emphasis on the new aftermarket support service for the engines provided via DS&S. Boeing personnel were shown DAME as an example of the new ways of gathering and processing the large amounts of data that could be retrieved from an advanced aircraft such as the 7E7, and they were very impressed”, DS&S 2004

Companies:
- Rolls-Royce
- DS&S
- Cybula

Universities:
- York,
- Leeds,
- Sheffield, Oxford

Engine Model

Case Based Reasoning

Signal Data Explorer

Engage徒弟
**Why Workflows and Services?**

Workflow = general technique for describing and enacting a process  
Workflow = describes *what* you want to do, not *how* you want to do it  
Web Service = *how* you want to do it  
Web Service = automated programmatic internet access to applications

- **Automation**  
  - Capturing processes in an explicit manner  
  - Tedium! Computers don’t get bored/distracted/hungry/impatient!  
  - Saves repeated time and effort  
- **Modification, maintenance, substitution and personalisation**  
- **Easy to share, explain, relocate, reuse and build**  
- **Available to wider audience: don’t need to be a coder, just need to know how to do Bioinformatics**  
- **Releases Scientists/Bioinformaticians to do other work**  
- **Record**  
  - Provenance: what the data is like, where it came from, its quality  
  - Management of data (LSID - Life Science IDentifiers)
**Workflow Components**

- **Scufl** Simple Conceptual Unified Flow Language
- **Taverna** Writing, running workflows & examining results
- **SOAPLAB** Makes applications available

**Freefluo**

Workflow engine to run workflows

**Web Service**

- e.g. DDBJ BLAST

**SOAPLAB Web Service**

Any Application
The Williams Workflows

A: Identification of overlapping sequence
B: Characterisation of nucleotide sequence
C: Characterisation of protein sequence
The Workflow Experience

Have workflows delivered on their promise? **YES!**

- **Correct and Biologically meaningful results**
- **Automation**
  - Saved time, increased productivity
  - Process split into three, you still require humans!
- **Sharing**
  - Other people have used and want to develop the workflows
- **Change of work practises**
  - *Post hoc* analysis. Don’t analyse data piece by piece receive all data all at once
  - Data stored and collected in a more standardised manner
  - Results amplification
  - Results management and visualisation
3. Web Services and Grids

- Computing models developed for sequential machines led to the distributed object model of distributed computing represented by Java and CORBA.

- Experience has shown that the distributed object model ties distributed entities together too tightly.

- Resulted in ‘fragile’ distributed software systems when going from LANs to WANs.

  - Replace distributed objects by ‘services’ connected by “one-way” messages and not by request-response messages.
  - IT industry has ‘united’ around Web Services.
The Web Services ‘Magic Bullet’

Company A (J2EE)

Company B (ONE)

Company C (.Net)
Web Service Grids: An Evolutionary Approach

Specifications that have/will enter a standardisation process but are not stable and are still experimental

‘WS-I+’ profile

Specifications that are emerging from standardisation process and are recognised as being ‘useful’

Standards that have broad industry support and multiple interoperable implementations
WS-I+ Grid Interoperability Profile

- WS-I identifies XSD, WSDL, SOAP, UDDI
- WS-I+ adds minimum additional capabilities to WS-I to allow development of Grid Services
  - BPEL and extensions for scientific workflows
  - WS-Addressing for virtualization of messaging
  - WS-ReliableMessaging/Reliability to provide basis for fault tolerant and efficient Grid services
- Expect progress in
  - WS-ResourceFramework – submitted to OASIS
  - Notification – dialogue between IT companies
  - Security – need to understand better relationship of Web Services and Grid approaches
  - Portlets – generic toolkit for portal construction
Important Higher Level Services

Many services associated with particular applications but also some services of broad applicability such as:

- Accounting
- Data movement
  - such as GridFTP and GridRPC
- Metadata
  - semantics of services are important
- Data Repositories
  - OGSA DAI with database and file access
- Computing services
  - Job Submittal, Status
  - Scheduling as in Condor, PBS, Sun Grid Engine
  - Links to MPI
Grids of Grids of Simple Services

Methods → Services → Functional Grids

CPUs → Clusters → Compute Resource Grids

Databases → Federated Databases → Data Resource Grids

Sensor → Sensor Nets

Overlay and Compose Grids of Grids
4. Building a National e-Infrastructure

Three major new activities for Phase 2 of the Core Programme:

1. Deployment of National Grid Service (NGS) and establishment of a Grid Operation Support Centre
2. Establish Open Middleware Infrastructure Institute (OMII) for testing, software engineering and UK repository
3. Set up Digital Curation Centre (DCC) to lead on long-term data preservation issues
NGS “Today”

Projects
- e-Minerals
- e-Materials
- Orbital Dynamics of Galaxies
- Bioinformatics (using BLAST)
- GEODISE project
- UKQCD Singlet meson project
- Census data analysis
- MIAKT project
- e-HTPX project
- RealityGrid (chemistry)

Users
- Leeds
- Oxford
- UCL
- Cardiff
- Southampton
- Imperial
- Liverpool
- Sheffield
- Cambridge
- Edinburgh
- QUB
- BBSRC
- CCLRC.

Interfaces
- OGSI::Lite

Core Node Software Stack

Example Applications
- DL_POLY Molecular Simulation Package
- Ab Initio
- NCBI BLAST
- MATLAB
- RealityGrid
NGS “Tomorrow”

NGS Production Service

Q2 Q4 Q2 Q3 Q1 Q4 Q3 Q2 Q1 Q3

2004 2006 2005

EGEE gLite alpha release

OMII release

GOC

Grid Operation Support Centre

NGS Expansion (Bristol, Cardiff...)

OGSA-DAI

WS plan

EGEE gLite release

OMII Release

NGS WS Service

NGS WS Service 2

WS2 plan

EGEE gLite release 1

OMII release

E-SCIENCE CA

Grid Engineering Task Force [ETF]

UK e-Science Grid Support Centre

HELPING REDUCE RISK AND IMPROVE GRID SUPPORT QUICKLY, EASILY AND RELIABLY

Web Services based National Grid Infrastructure
OMII Vision

• To be the national provider of reliable, interoperable, open source grid middleware
• Provide one-stop portal and software repository for grid middleware
• Provide quality assured software engineering, testing, packaging and maintenance for our products
• Lead the evolution of Grid middleware through a managed programme and wide reaching collaboration with industry
OMII Distribution 1 – Oct 2004

• Collection of tested, documented and integrated software components for Web Service Grids
• A base built from off-the-shelf Web Services technology
• A package of extensions that can be enabled as required
• An initial set of Web Services for building file-compute collaborative grids
• Technical preview of Web Service version of OGSA-DAI database middleware
• Sample applications
OMII future distributions

• Include the services in previous distributions +…
• OMII managed programme contributions
  – Database service
  – Workflow service
  – Registry service
  – Reliable messaging service
  – Notification service
• Interoperability with other grids
Digital Curation Centre

• Actions needed to maintain and utilise digital data and research results over entire life-cycle
  – For current and future generations of users
• Digital Preservation
  – Long-run technological/legal accessibility and usability
• Data curation in science
  – Maintenance of body of trusted data to represent current state of knowledge in area of research
• Research in tools and technologies
  – Integration, annotation, provenance, metadata, security…..
5. Dual Support and the Role of JISC

Provides two streams of public funding for university research:

– Funding provided by the DfES and HEFCs for research infrastructure – salaries of permanent academic staff, premises, libraries and IT services

– Funding from the DTI and OST for specific projects – in response to proposals submitted & approved through peer review

➢ A national eInfrastructure to support collaborative and multidisciplinary research and innovation is the joint responsibility of RCUK (OST) and JISC (HEFCs)
The JISC Communities

- Portals
- Applications
- Content
- Meta Data & Delivery tools
- Finding /Access tools
SuperJANET4/5
JISC £6.5M for UKLight ‘Lambda’ Network
RCUK Funding for Research using the UKLight network

Three major research projects funded:

• ESLEA (EPSRC, e-Science, PPARC and MRC)
  – Network protocols and Quality of Service research for four e-Science application areas - £1M

• MASTS (EPSRC and e-Science)
  – Probes and tools to record, analyse and control full, sampled and compressed network traffic - £650k

• 46PaQ (EPSRC)
  – IPv4 and IPv6 Performance and Quality of Service - £1.2M
eBank Project

Virtual Learning Environment

E-Scientists

Undergraduate Students

Graduate Students

E-Experimentation

E-Scientists

Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning
JISC £3M Programme for a Virtual Research Environment (VRE)
JCSR ‘e-Science for Schools’ Projects

Funded 3 demonstrators:

• e-Star
  - provides remote control of telescopes and access to astronomical databases

• e-Malaria
  - uses drug screening and remote use of crystallographic grid service

• e-Environment
  - with remote sensors and data collection and analysis

Major Components of the UK Vision

• Multidisciplinary Working
  – Creation of a multidisciplinary research environment
  – Links between Funding Councils and RCUK
  – Uses e-Science exemplars from Earth Systems Science and Systems Biology

• Information Infrastructure
  – Access to experimental data sets and publications
  – Collection and preservation of digital information
  – Importance of National e-Infrastructure
  – Tie into international efforts
  – OST to take the lead

- Capital Infrastructure – Large Facilities
  - Diamond Synchroton to open 2007
  - Second Target Station for ISIS Neutron Source from 2008
  - Large Hadron Collider operative from 2007

- Plus
  - Hector HPC Facility
  - ITER Fusion Machine
UK e-Infrastructure

Users get common access, tools, information, Nationally supported services, through NGS

GOSC

Integrated internationally

LHC
ISIS TS2
HPCx + HECtoR

Regional and Campus grids
Pasteur’s Quadrant

• Innovation not restricted to the classic ‘linear’ model
  Basic research → Applied Research → Development → Product

• Stokes classified innovation in 2-D model
  In addition to innovation arising from linear approach, innovation and fundamental research can result from application-inspired R&D programmes

• Paradigm is Louis Pasteur’s development of immunology
• UK e-Science Programme follows Stokes’ model
DTI Innovation Strategy

Technology Strategy Board

<table>
<thead>
<tr>
<th>ICT</th>
<th>BIO</th>
<th>Advanced Materials</th>
<th>Advanced Manufacturing</th>
<th>Energy &amp; Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Manager</td>
<td>Technology Manager</td>
<td>Technology Manager</td>
<td>Technology Manager</td>
<td>Technology Manager</td>
</tr>
<tr>
<td>Computing and Communications networks</td>
<td>Genomics</td>
<td>Structural Materials</td>
<td>Product and process technology</td>
<td>Fuel cells</td>
</tr>
<tr>
<td>Bio-informatics</td>
<td>Nanotechnology</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Priority selection will be defined through a set of key criteria

Knowledge Transfer Networks Collaborative R & D

DELIVERY
through working with...

- EU Framework
- RDAs
- BR Sector teams
- Innovation Group
- Business Link
- FCO
- Eureka
- Research Councils

Customer Solution
DTI Technology Fund: Inter-Enterprise Computing Theme

• 63 R&D proposals submitted at outline stage
  ➢ 18 invited to submit full proposals, competing for £6M+ funding

• 10 managed network proposals submitted at outline stage:
  ➢ 4 invited to submit full proposals

• e-Science stakeholders in 17 of the 18 invited R&D proposals

• All the network bidders have recognised that they must engage the e-Science community
e-Infrastructure for Research and Innovation

• Ten-year investment framework is collaboration between the Treasury, the DfES and the DTI
• The RCUK e-Science Programme with the Core Programme have made a good start at building the UK e-Infrastructure
• Need continuing collaboration between RCUK, JISC and the DTI

- Essential to continue ring-fenced funding for e-Infrastructure and e-Research in SR2004 settlement
Acknowledgements

With special thanks to Jim Austin, Ray Browne, Peter Burnhill, Tony Doyle, Alistair Dunlop, Geoffrey Fox, Peter Freeman, Jeremy Frey, David Gavaghan, Neil Geddes, Carole Goble, Sharon Lloyd, Hannah Tipney, Anne Trefethen and Lee Vousden