Grids & Life Sciences

Dr Richard Sinnott

Technical Director National e-Science Centre
Deputy Director Technical Bioinformatics Research Centre
University of Glasgow

26th March 2004
Outline

- What is e-Science?
- UK e-Science
  - UK e-Science: Roles and Resources
  - UK e-Science: Projects
  - e-Science Institute
- Grids and Life Sciences
  - Data volumes and heterogeneity
  - Computational demands
- Overview of Bridges Project
- Conclusions
What is e-Science?
e-Science and the Grid

'e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.'

'e-Science will change the dynamic of the way science is undertaken.'

John Taylor
Director General of Research Councils
Office of Science and Technology

Grid is infrastructure used for e-Science

• Metaphor of Power Grid: compute and data resources on demand ... revisited later
Foundation for e-Science

- e-Science methodologies will **rapidly transform** science, engineering, medicine and business
  - driven by exponential growth ($\times 1000$/decade)
    - enabling a whole-system approach
Data Grids for High Energy Physics

There is a "bunch crossing" every 25 nsecs.
There are 100 "triggers" per second
Each triggered event is ~1 MByte in size

Physicists work on analysis "channels".
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server.
global in-flight engine diagnostics

- 100,000 engines
- 2-5 Gbytes/flight
- 5 flights/day = 2.5 PB/day

Distributed Aircraft Maintenance Environment: Universities of Leeds, Oxford, Sheffield & York
Virtual Observatories

- Huge data sets
  - AstroGrid over 15TB data first week online
- Huge computations
  - Cross referencing data
  - Remove all junk from data sets
    - satellites, aeroplanes…

Supernova Remnant Cassiopeia-A – a 300 year old Supernova

Shocks seen in the IR using e-s Grid

Heavy element seen in the optical by Radio data

Images from Chandra Science Centre
e-Science in the UK
UK e-Science Funding

First Phase: 2001 -2004
- Application Projects
  - £74M
  - All areas of science and engineering
  - >60 Projects
  - 340 at first All Hands Mtg
- Core Programme
  - £35M
  - Collaborative industrial projects
  - ~80 Companies
  - > £30 Million

Second Phase: 2003 -2006
- Application Projects
  - £96M
  - All areas of science and engineering
- Core Programme
  - £16M + £25M
  - £150M (non-ringfenced) DTI
  + EU money!
  + €40M
  + Janet upgrade
  + HPC(x) £55M
# e-Science and SR2002

<table>
<thead>
<tr>
<th>Funding Body</th>
<th>2004-6</th>
<th>2001-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRC</td>
<td>£13.1M</td>
<td>(£8M)</td>
</tr>
<tr>
<td>BBSRC</td>
<td>£10.0M</td>
<td>(£8M)</td>
</tr>
<tr>
<td>NERC</td>
<td>£8.0M</td>
<td>(£7M)</td>
</tr>
<tr>
<td>EPSRC</td>
<td>£18.0M</td>
<td>(£17M)</td>
</tr>
<tr>
<td>HPC</td>
<td>£2.5M</td>
<td>(£9M)</td>
</tr>
<tr>
<td>Core Prog.</td>
<td>£16.2M</td>
<td>£20M</td>
</tr>
<tr>
<td>PPARC</td>
<td>£31.6M</td>
<td>(£26M)</td>
</tr>
<tr>
<td>ESRC</td>
<td>£10.6M</td>
<td>(£3M)</td>
</tr>
<tr>
<td>CLRC</td>
<td>£5.0M</td>
<td>(£5M)</td>
</tr>
</tbody>
</table>
# E-Science Projects

## Projects in Bio-Sciences

<table>
<thead>
<tr>
<th>Title (Abbreviation)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integration of Sequence and Structural Family Data (ISSFD)</td>
<td>MyGRID: Directly Supporting the E-Scientist (myGrid)</td>
</tr>
<tr>
<td>3D OPT: Microscopy Grid: Bringing the Grid to the Biomedical Workbench (3D OPT)</td>
<td>Biography of ageing e-science integration and simulation system (BASIS)</td>
</tr>
<tr>
<td>e-Science Technologies in the Simulation of Complex Materials (I/NA13)</td>
<td>Bioinformatics Summer Schools for Researcher in Biological Sciences (Biolf-summer School)</td>
</tr>
<tr>
<td>Bioinformatics Summer Schools for Researchers in Biological Sciences</td>
<td>BiodiversityWorld (Biodiversity)</td>
</tr>
<tr>
<td>A problem solving environment for global biodiversity prototype and demonstrator</td>
<td>e-PROTEIN: A distributed pipeline for structural-based proteome annotation using GRID technology (e-PROTEIN)</td>
</tr>
<tr>
<td>e-PROTEIN: A distributed pipeline for structural-based proteome annotation using GRID technology</td>
<td>A GRID Database for biomolecular simulations (BiomGRID)</td>
</tr>
<tr>
<td>A GRID Database for biomolecular simulations (BiomGRID)</td>
<td>UTOPIA (UTOPIA)</td>
</tr>
<tr>
<td>UTOPIA (UTOPIA)</td>
<td>The CoG GridBio Annotation Tool (COAT)</td>
</tr>
<tr>
<td>BioMedical Research Information Delivered by Grid Enabled Services (BRIDGES)</td>
<td>Cardiovascular Functional Genomics - An Integrated Distributed Database (COSMOS GRID)</td>
</tr>
<tr>
<td>Cardiovascular Functional Genomics - An Integrated Distributed Database (COSMOS GRID)</td>
<td>COSMOS GRID (COSMICS GRID)</td>
</tr>
<tr>
<td>COSMIC GRID (COSMIC GRID)</td>
<td>Profiling for Breast and Colorectal Cancer (Profilingfor Breast and Colorectal Cancer)</td>
</tr>
<tr>
<td>Profiling for Breast and Colorectal Cancer (Profilingfor Breast and Colorectal Cancer)</td>
<td>E-HTPX (E-HTPX)</td>
</tr>
<tr>
<td>E-HTPX (E-HTPX)</td>
<td>BioQuery (BioQuery)</td>
</tr>
<tr>
<td>BioQuery (BioQuery)</td>
<td>LOWT (LOWT)</td>
</tr>
<tr>
<td>LOWT (LOWT)</td>
<td>Development and Distribution of genetic circuit models for seasonal reproduction (Development and Distribution of Genetic Circuit Models)</td>
</tr>
<tr>
<td>Development and Distribution of genetic circuit models for seasonal reproduction</td>
<td>Development of an Information Repository for Microbial Genome Comparison Using a Parallel, Object Based Architecture (Development of an Information Repository for Microbial Genome Comparison)</td>
</tr>
<tr>
<td>Development of an Information Repository for Microbial Genome Comparison Using a Parallel, Object Based Architecture</td>
<td>Development of generic software for analysis, archiving and internet dissemination of brain and systems physiological data (Development of Generic Software for Analysis, Archiving and Internet Dissemination of Brain and Systems Physiological Data)</td>
</tr>
<tr>
<td>Development of generic software for analysis, archiving and internet dissemination of brain and systems physiological data</td>
<td>Microarray Analysis Summer School (Microarray Analysis Summer School)</td>
</tr>
<tr>
<td>Microarray Analysis Summer School (Microarray Analysis Summer School)</td>
<td>Model Sharing and Co-simulation Standards for System Biology (Model Sharing and Co-simulation Standards for System Biology)</td>
</tr>
<tr>
<td>Model Sharing and Co-simulation Standards for System Biology (Model Sharing and Co-simulation Standards for System Biology)</td>
<td>MultiFlora I: Combining Information Extraction and Knowledge Representation for Biodiversity Informatics (MultiFlora I: Combining Information Extraction and Knowledge Representation for Biodiversity Informatics)</td>
</tr>
<tr>
<td>MultiFlora I: Combining Information Extraction and Knowledge Representation for Biodiversity Informatics</td>
<td>Novel approaches to the analysis of genetic diversity in germplasm collections (Novel Approaches to the Analysis of Genetic Diversity in Germplasm Collections)</td>
</tr>
<tr>
<td>Novel approaches to the analysis of genetic diversity in germplasm collections</td>
<td>Proposal to provide an extended GRID support service for BBSRC (Proposal to Provide an Extended GRID Support Service for BBSRC)</td>
</tr>
<tr>
<td>Proposal to provide an extended GRID support service for BBSRC (Proposal to Provide an Extended GRID Support Service for BBSRC)</td>
<td>Structure-based function prediction using peer-to-peer distributed computing (Structure-based Function Prediction Using Peer-to-Peer Distributed Computing)</td>
</tr>
<tr>
<td>Structure-based function prediction using peer-to-peer distributed computing (Structure-based Function Prediction Using Peer-to-Peer Distributed Computing)</td>
<td></td>
</tr>
</tbody>
</table>
E-Science Projects

- Many others
- Physical/engineering sciences
- Social sciences
- Middleware development
- ...
NeSC in the UK

Directors’ Forum
- Helped build a community

Engineering Task Force

Grid Support Centre

Architecture Task Force
- UK Adoption of OGSA
- OGSA Grid Market
- Workflow Management

Database Task Force
- OGSA-DAI
- GGF DAIS-WG

e-SI Programme
- Training, coordination, community building, workshops, pioneering

GridNet, e-Storm
UK E-Science Centres and Grid

- Currently Level-2 Grid based on Globus Toolkit 2
  - Particle Physics Grids in UK
    - GridPP
    - ScotGrid
  - (and more)
  - Demonstrated broad set of applications running across it
  - International Grids
    - Earth Systems BLAST jobs
    - European Data Grid
    - Grid Physics Network (GriPhyN)
    - NASA Information Power Grid (IPG)
- Transition to Grid based upon Open Grid Service Architecture
  - Based upon Grid (web) services
  - Based upon Grid (web) services
  - Engineering Simulation Grid
Glasgow e-Science Hub

- E-Science Hub
  - Externally
    - Glasgow end of NeSC
      - Involved in UK wide activities
        » ETF: In May 2003 became first UK e-Science Centre to run integration tests across every site of the UK (Level 2) Grid. Therefore 100% access to UK Grid resources at this time
      - Public visibility of NeSC
        » responsible for NeSC web site
  - Internally
    - Focal point for e-Science research/activities at Glasgow
    - Work closely with foundation departments
      - Department of Computing Science
      - Department of Physics & Astronomy
    - Also working closely with other groups including
      - Bioinformatics Research Centre
      - Electronics and Electrical Engineering
      - Biostatistics
      - ...
Glasgow e-Science Investment

- Major investment by university
  - 230m² of newly renovated floor space in Kelvin Building
    - offices
    - access grid facility
    - training room
      - equipped with 20PCs/server for training courses

- Funding Technical Director
Glasgow e-Science Activities

- Consolidating resources
  - Building around ScotGrid
    - Providing shared Grid resource for wide variety of scientists inside/outside Glasgow
      - Particle physicists, computer scientists, bioinformaticians, ...
      - Target shares established
      - Focal point for e-Science at Glasgow

Hardware
- 59 IBM X Series 330 dual 1 GHz Pentium III with 2GB memory
- 2 IBM X Series 340 dual 1 GHz Pentium III with 2GB memory
- 3 IBM X Series 340 dual 1 GHz Pentium III with 2GB memory and 100 + 1000 Mbit/s ethernet
- 1TB disk
- LTO/Ultrium Tape Library
- Cisco ethernet switches

New...
- IBM X Series 370 PIII Xeon with 32 x 512 MB RAM
- 5TB FastT500 disk 70 x 73.4 GB IBM FC Hot-Swap HDD
- eDIKT 28 IBM blades dual 2.4 GHz Xeon with 1.5GB memory
- eDIKT 6 IBM X Series 335 dual 2.4 GHz Xeon with 1.5GB memory
- CDF 10 Dell PowerEdge 2650 2.4 GHz Xeon with 1.5GB memory
- CDF 7.5TB Raid disk

Shared Resources:
- Disk ~15TB
- CPU ~ 330 1GHz
e-Science Institute
e-Science Institute
Past Programme of Events

- Planned 6 * two-week research workshops / year
- Actually ran 48 events in first 12 months!
  - GGF5, HPDC11 and a cluster of workshops
  - Protein Science, Neuroinformatics, ...
  - Major training events
    - Web Services, DiscoveryLink, Relational DB design,
- e-SI Clientele and Outreach (year 1)
  - > 2600 individuals
  - From > 500 organisations
  - 236 speakers
  - Many participants return frequently
- Event Diversity
  - Conferences
  - Summer schools
  - Workshops
  - Training
  - Community building
    - Company Sponsorship
    &
    - International Engagement

Event Diversity Chart:
- GGF1, GGF2, GGF3, GGF4, GGF5, GGF6, GGF7
Grids and Life Sciences
Biology & Medicine

- Extensive Research Community
  - >1000 per research university

- Extensive Applications
  - Many people care about them
    - Health, Food, Environment

- Interacts with virtually every discipline
  - Physics, Chemistry, Nano-engineering, ...

- 450 Databases relevant to bioinformatics
  - Heterogeneity, Interdependence, Complexity, Change, ...

- Wonderful Scientific Questions
  - How does a cell work?
  - How does a brain work?
  - How does an organism develop?
  - What happens to the biosphere when the earth warms up?
  - ...
Database Growth

- DBs growing exponentially!!!
  - Biobliographic (MedLine, ...)
  - Amino Acid Seq (SWISS-PROT, ...)
  - 3D Molecular Structure (PDB, ...)
  - Nucleotide Seq (GenBank, EMBL, ...)
  - Biochemical Pathways (KEGG, WIT...)
  - Molecular Classifications (SCOP, CATH,...)
  - Motif Libraries (PROSITE, Blocks, ...)

EMBL Database Growth (total record number in millions)

PDB Content Growth (number of structures for the year, total available structures)
More genomes...
Distributed and Heterogeneous data

**Sequence**

LPSYVDWRSA GAVVDIKSQG
ECGGCWAFSA IATVEGINKI
TSGSLISLSE QELIDCGRTQ
NTRGCDGGYI TDGFQFIIND
GGINTEENYP YTAQDGDCDV

**Structure**

**Function**

Gene expression

Morphology

pregnancy
Complexity of Biological Data

- Nucleotide sequences
- Nucleotide structures
- Gene expressions
- Protein structures
- Protein functions
- Protein-protein interaction (pathways)
- Cell
- Cell signalling
- Tissues
- Organs
- Physiology
- Organisms
- Populations

Mechanism of enzyme activity

National e-Science Centre

University of Glasgow
Bioinformatics Grid Needs

Compute Needs

- Whole genome-genome \((2 \times 3 \times 10^9 \text{ bp})\) comparisons between two species
  - Current strategy essentially chops up one genome and fires searches for those fragments in the other then re-assembles results
    - messy approximate matching - re-assembly difficult
    - important correlations can be lost
      - to make this tractable so called junk DNA ignored
      - chopping may introduce artefacts or hide phenomena

- Better to put both full genomes in memory and perform a useful complete comparison
- Only possible with very high-end machines (available via grids)
Bioinformatics Grid Needs

Workflow / Virtual Organisation Needs

OGSA_DAI/DAIT, IBM DiscoveryLink, ...

WSDL descriptions, Semantic grid, ...

BioInf community, Database schemas, ...

UDDI repositories, BioInf portals, ...

Single sign on authentication, Granularity of authorisation

National Data Curation Centre (GU,EU,UKOLN, CCLRC)

Grid engineering (scheduling, resource reservation, workflow enactment, ...)

Goble myGrid presentation
Overview of BRIDGES

- Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES)
  - NeSC (Edinburgh and Glasgow) and IBM
- Supporting project for CFG project
  - Generating data on hypertension
  - Rat, Mouse, Human genome databases
- Variety of tools used
  - BLAST, BLAT, Gene Prediction, visualisation, ...
- Variety of data sources and formats
  - Microarray data, genome DBs, project partner research data, medical records, ...
- Aim is integrated infrastructure supporting
  - Data federation
  - Security
CFG Partner Distribution

- Glasgow
  - Private data
- Oxford
  - Private data
- London
  - Private data
- Edinburgh
  - Private data
- Leicester
  - Private data
- Netherlands
  - Private data
- Public curated data

Shared data
Problems to be addressed

- BRIDGES will address the following problems facing CFG biologists

  - How to integrate data with multiple levels of security including public data, project only data and private data?
  - How to search multiple distributed databases through single optimised queries?
  - How to use multiple tools in a coordinated (and automated) manner, e.g. how to develop re-useable workflows for the CFG scientists?
  - Integration of a range of bioinformatics analysis and visualisation tools, e.g. BLAST, genome browsers, etc.
  - How to deal with inconsistencies of online databases and possible “dirty data”?
  - How to get more “up to date” data?

- Make it all user friendly...
  - portals,
  - hidden infrastructure, e.g. security authorisation
Planned Approach

- BRIDGES will address these problems through
  - Development of re-useable Grid services based upon GT3 technologies
  - Virtualisation of multiple distributed data sets to provide a single virtual data set for use by the biologists - exploiting IBM’s Information Integrator
  - Developing a collection of data on a well-managed platform, including copies of extracts of relevant public data, all project data, and the required software tools (administered using DB2 and Information Integrator)
  - Access to and integration of multiple distributed data sets in a Grid environment using results from the OGSA_DAI/DAIT projects
  - A secure environment offering authentication and authorisation
    - will build on results of the PERMIS security authorisation project
Security

- **OGSA security**
  - Single sign-on based on (X.509) digital certificates
    - establish credentials
      - Certification authority based
  - Services (and clients) have APIs for fine grained security
    - Based on GSS-API
  - Provides for authentication but need authorisation
    - No standardised way of doing this right now!
  - Collaborating with PrivilEge and Role Management Infrastructure Standards Validation (PERMIS) team
Security Authentication

- **PERMIS allows to**
  - Define roles for who can do what
  - Based on Privilege Management Infrastructure certificates
  - Users get certificate for what they are allowed to do
    - Defines dynamic attributes for privileges
    - Can user X access/change data Y?
  - Will investigate scalability and generality of PERMIS for biologists!
Sort of things that Bridges is developing...
DRILL-DOWN FUNCTIONS

To tabular summaries

To sequence

To multiple alignment

Status/Vision of Tools via Portal
Conclusions

- The GRID is happening and will influence life sciences
  - Government SR allocated £115M for eScience
  - Life sciences are key target areas for eScience

- Globus Toolkit version 3 recently released
  - Non-trivial learning curve
  - Immaturity of technology
  - Standards perpetually evolving...

- Courses offered at NeSC
  - Web services, grid services, XML, Java, DB technologies, bioinformatics, others…?
  - [www.nesc.ac.uk](http://www.nesc.ac.uk) for registration and suggesting events

- Grid Computing modules under development at Glasgow as part of advanced MSc
Conclusions

- Main take home message:
  - need coordination of
    - GRID software engineering/standards
    - bioinformatics needs/standards
  - Without this, metaphor of electric power GRID doomed
Questions?

www.nesc.ac.uk