UK e-Science: data, data everywhere, ...

Sunlabs, Boston
Malcolm Atkinson
Director
www.nesc.ac.uk
7th October 2005

Data, Data Everywhere & …

- Growing volumes
- Growing diversity
- Growing complexity
- How do we mine its riches for nuggets of information?
- Find & Access
- Understand
- Extract, Combine & Digest
- Test hypotheses
- Bingo!

⇒ Rich resource

⇒ Use OGSA-DAI
Story Line

- The UK e-Science Context
- Example Projects
  - Their demand for Data Integration
- OGSA-DAI
  - An Extensible Framework
  - A kit of parts
  - A way of life
- Summary & Take Home Messages

What is e-Science?

- Goal: to enable better research in all disciplines
What is e-Science?

- **Method: Invention and exploitation of advanced computational methods**
  - to generate, curate and analyse research data
    - From experiments, observations and simulations
    - Quality management, preservation and reliable evidence
  - to develop and explore models and simulations
    - Computation and data at extreme scales
    - Trustworthy, economic, timely and relevant results
  - to enable dynamic distributed virtual organisations
    - Facilitating collaboration with information and resource sharing
    - Security, reliability, accountability, manageability and agility

Why use / build Grids?

- **Research Arguments**
  - Enables new ways of working
  - New distributed & collaborative research
  - Unprecedented scale and resources
- **Economic Arguments**
  - Reduced system management costs
  - Shared resources ⇒ better utilisation
  - Pooled resources ⇒ increased capacity
  - Load sharing & utility computing
  - Cheaper disaster recovery
Grids: a foundation for e-Science

- e-Science methodologies will rapidly transform science, engineering, medicine and business
  - driven by exponential growth (×1000/decade)
    - enabling a whole-system approach

Diagram derived from
Ian Foster’s slide

Why use / build Grids?

- **Computer Science Arguments**
  - New attempt at an old hard problem
  - Frustrating ignorance of existing results
  - New scale, new dynamics, new scope

- **Engineering Arguments**
  - Enable autonomous organisations to
    - Write complementary software components
    - Set up run & use complementary services
    - Share operational responsibility
Why use / build Grids?

- Political & Management Arguments
  - Stimulate innovation
  - Promote intra-organisation collaboration
  - Promote inter-enterprise collaboration

What is e-Infrastructure - Political view

- A shared resource
  - That enables science, research, engineering, medicine, industry, ...
  - It will improve UK / European / ... productivity
    - Lisbon Accord 2000
    - E-Science Vision SR2000
      - John Taylor
  - 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
Chancellor of the Exchequer

Charles Clarke
Secretary of State for Education and Skills

Patricia Hewitt
Secretary of State for Trade and Industry
UK e-Science Budget (2001-2006)

Total: £213M + £100M via JISC

EPSRC Breakdown

- Applied (£35M) 45%
- Core (£31.2M) 40%
- HPC (£11.5M) 15%

+ Industrial Contributions £25M

Source: Science Budget 2003/4 – 2005/6, DTI(OST)

Slide from Steve Newhouse

The e-Science Centres

- Globus Alliance
- National Centre for e-Social Science
- Open Middleware Infrastructure Institute
- OMII-UK
- EGEE
- e-Science Institute
- Digital Curation Centre
- Grid Operations Support Centre
- National Institute for Environmental e-Science
- OMII-UK
The Primary Requirement ...
Building people grids

Enabling People to Work Together on Challenging Projects: Science, Engineering & Medicine

Service-Oriented Architecture

Registry

Discovery → Registration

Invocation

Client ↔ Service
Story so Far

- UK commits to invest
  - To stimulate academia & industry
- Differentiated by Coordination
  - Strong leadership from Tony Hey
- Initially committed to US Grid M/W
- Invested in Web Services
- Focus on Data & Information Grids
  - DAI
  - Semantic Grids
  - Application driven development

Story Line

- The UK e-Science Context
- Example Projects
  - Their demand for Data Integration
- OGSA-DAI
  - An Extensible Framework
  - A kit of parts
  - A way of life
- Summary & Take Home Messages
Motivation

- Entering an age of data
  - Data Explosion
    - CERN: LHC will generate 1GB/s = 10PB/y
    - VLBA (NRAO) generates 1GB/s today
    - Pixar generate 100 TB/Movie
  - Storage getting cheaper
- Data stored in many different ways
  - Data resources
    - Relational databases
    - XML databases
    - Flat files
- Need ways to facilitate
  - Data discovery
  - Data access
  - Data integration
- Empower e-Business and e-Science
  - The Grid is a vehicle for achieving this

Composing Observations in Astronomy

No. & sizes of data sets as of mid-2002, grouped by wavelength
- 12 waveband coverage of large areas of the sky
- Total about 200 TB data
- Doubling every 12 months
- Largest catalogues near 1B objects

Data and images courtesy Alex Szalay, John Hopkins
eDiaMoND: Screening for Breast Cancer

1 Trust → Many Trusts
Collaborative Working
Audit capability
Epidemiology

Screening

- Digital Reading
- Secondary Capture

Case Information

- X-Rays and Case Information
- CAD Temporal Comparison

Biopsy

Assessment/ Symptomatic
- Other Modalities
- MRI
- PET
- Ultrasonic
- Better access to Case information and digital tools

Training

- Manage Training Cases
- Perform Training

Case and Reading Information

- SMF
- CAD

Provided by eDiamond project: Prof. Sir Mike Brady et al.
**Marketing department identifies likely buyers of new product**

- **Company wants real-time integrated view of customer buying behavior**
- **Data resides in various distributed CRM & ERP systems**
- **Grid allows developers and apps to access and integrate customer data sources together in real time—across many distributed databases**

**Providing Data to Cluster-Based Analytical Application**

- **Company has centralized HPC cluster running compute-intensive applications**
- **Source data for analysis distributed among 3 global sites, one of them an external partner**
- **Manual data-sharing processes increase costs/errors, and hinder time-to-results**
- **Grid enables secure, automatic provisioning of remote data to HPC cluster—feeding CPUs more data faster**
Story so far

- There is a lot of data
  - growing in every dimension
  - Distributed
  - Many different producers & owners
  - Heterogeneous
  - High value resource
  - Combined it is more valuable

- There are many requirements for data integration
  - Takes many forms
  - Driven by insights
  - Enable conversion of insight to tested hypothesis

- There are many data owners
  - Their autonomy and policies must be respected

Generic Repeatable Solutions Required

Grid Construction Principles

- **Dynamic coupling based on SOA**
- **Respect Autonomy – local policies**
- **Independent construction & provision**
  - Requires adherence to agreed protocols
  - Interoperability
  - Preferably with widely adopted standards
- **Algorithms & Information structures**
  - Must scale well
  - Must be tolerant to partial failure
  - Must be tolerant to partial system change
- **Mechanisms to build trust are essential**

Providing *mutually consistent* services with compatible policies
Story Line

- The UK e-Science Context
- Example Projects
  - Their demand for Data Integration
- OGSA-DAI
  - Architectural Context
  - An Extensible Framework
  - A kit of parts
  - A way of life
- Summary & Take Home Messages

Three communities

- Users: Individual & Organisations
- Data, Information & Knowledge Providers
- Compute Storage & Communications Providers
Three communities

Users: Individual & Organisations

Initiate & Steer work
Provide Requirements
Use knowledge & insight
Want easy & reliable facility
Expect agility, reliability, stability
& latest techniques
Pay the bills

Data, Information & Knowledge Providers

Compute Storage & Communications Providers

Three communities

Users: Individual & Organisations

Provide & operate resources
Storage centres, Data centres, DBMS & File Systems
Computation environment, Processing & Communications
Need to change facilities & policies
Prefer consolidated requirements
Must be paid

Data, Information & Knowledge Providers

Compute Storage & Communications Providers
OGSA-DAI

Three communities

Users: Individual & Organisations

Create & Collect data
Organise & Structure data
Provide, organise & maintain metadata
Offer access and domain specific services
Establish use policies
Will change structure, services & policies
May pay or be paid

OGSA-DAI

a framework for a lasting & productive relationship

Data, Information & Knowledge Providers

Compute Storage & Communications Providers
OGSA-DAI In One Slide

- An extensible framework for data access and integration.
- Expose heterogeneous data resources to a grid through web services.
- Interact with data resources:
  - Queries and updates.
  - Data transformation / compression
  - Data delivery.
- Customise for your project using
  - Additional Activities
  - Client Toolkit APIs
  - Data Resource handlers
- A base for higher-level services
  - federation, mining, visualisation,…

Connecting three communities

Users: Individual & Organisations

Portals & Applications

Client Library

OGSA-DAI

Data, Information & Knowledge Providers

Provided Interfaces & Services

Adaptive Interfaces

Adaptive Interfaces

Compute Storage & Communications Providers
International Collaboration & Use

USA:
- Globus Alliance
- IBM Corporation
- caBIG
- BIRN
- Indiana University
- GridSphere
- GEON
- LEAD
- MCS
- NCSA
- Secure Data Grid
- UNC

Japan:
- AIST
- BioGrid
- NAREGI

Europe:
- CERN
- DataMiningGrid
- GridSphere
- GridMiner
- GridSphere
- inteligrid
- N2Grid
- OntoGrid
- Provenance
- SIMDAT
- OMII-EU

UK:
- OMII
- OMII-UK
- NAG
- NAGSS
- NixiSS
- AsimGrid
- BRIDGES
- CancerGrid
- ConvertGrid
- eDiaMonD
- EDINA
- First Group plc
- Fujitsu Labs Europe
- GEODAM
- GeneGrid
- Genomic Technology and Informatics
- GOLD
- Human Genetics Unit
- IBM UK
- i^3 Grid
- Oracle UK

China:
- CAS
- ChinaGrid
- cnGrid
- INWA
- NMI-China

Australia:
- Curtin Business School
- INWA

Tutorials
- Boston
- CERN
- Edinburgh
- San Francisco
- Seoul
- Tokyo
- Cambridge
- Chicago
- London
- Seattle
- Singapore
- ISSGC 03 to 05

Meeting User Requirements

FirstDIG
ConvertGrid
GeneGrid
eDiaMoND
BRIDGES
LEAD
OGSA-DQP
OGSA-WebDB
caBIG

1485 registered users
5250+ downloads

United Kingdom: 40%
China: 30%
United States: 11%
Others: 20%

Austria: 1%
France: 3%
Italy: 2%
Japan: 5%
Germany: 3%
United States: 11%

1485 registered users
5250+ downloads
Inside a Data Service (1/2)

- Extensible document-based interface:
  - Change in behaviour ≠ interface change.
  - Reduce number of operation calls.
- Perform document:
  - Data resource-related activities – query, update, create, compress, transform, deliver.
  - Data flows between activities.
- Response documents:
  - Status of executed activities, result data.
DAIS Data Access

OGSA-DAI

Consumer

Database Data Service

SQLExecute ( SQLExpression )

SQLResponse

SQLDescription:
  Readable
  Writeable
  ConcurrentAccess
  TransactionInitiation
  TransactionIsolation
  Etc.

SQLAccess

Relational Database

DAIS Derived Data Access

OGSA-DAI

Consumer

Database Data Service

SQLExecuteFactory ( SQLExpression BehaviouralProperties )

Reference to SQLResponse Data Service

SQLFactory

SQLDescription:
  Readable
  Writeable
  ConcurrentAccess
  TransactionInitiation
  TransactionIsolation
  Etc.

SQLResponseDescription

SQLResponseAccess

Row Set

RDBMS specific mechanism for generating result set

Rowset

GetRowset ( rowsetnumber )
Inside a Data Service (2/2)
OGSA-DAI Deck of Activities

OGSA-DAI Architecture
Disciplines organise specialised extension

Users: Individual & Organisations

Portals & Applications

[Diagram showing OGSA-DAI, Data, Information & Knowledge Providers, Compute Storage & Communications Providers, Portals & Applications, Activities, Provided Interfaces & Services, Stable interfaces for users, Stable interfaces for application developers, Stable interfaces for activity developers, Stable interfaces for data providers, Stable interfaces for resource providers, Adaptive Interfaces, Adaptive & Extensible Interfaces, Provided Interfaces & Services, Medical image specialisation, Astro specialisation]

Providing stability

Users: Individual & Organisations

Portals & Applications

[Diagram showing OGSA-DAI, Data, Information & Knowledge Providers, Compute Storage & Communications Providers, Portals & Applications, Activities, Provided Interfaces & Services, Stable interfaces for users, Stable interfaces for application developers, Stable interfaces for activity developers, Stable interfaces for data providers, Stable interfaces for resource providers, Adaptive Interfaces, Adaptive & Extensible Interfaces, Provided Interfaces & Services, Medical image specialisation, Astro specialisation]
Extending capability & interfaces

Users: Individual & Organisations

- Add new operations, more abstraction & composition tools
- Add new functions & higher abstraction
- Expose new capability offered by data providers
- Always preserving old interfaces exposed to data, developer & user communities

OGSA-DAI

Data, Information & Knowledge Providers

- Provided Interfaces & Services
- Adaptive Interfaces

Portals & Applications

- Extending capability & interfaces

Adaptive Interfaces

Computational Storage & Communications Providers

OGSA-DAI Client Library

Path towards improvement

Users: Individual & Organisations

- Drive abstraction & composition from metadata
- Automate adaptation to sustain delivered functions & extension
- Exploit metadata

- Negotiate SLAs with providers & obtain better description of resources

OGSA-DAI

Data, Information & Knowledge Providers

- Provided Interfaces & Services
- Adaptive Interfaces

Portals & Applications

- Extending capability & interfaces

Adaptive Interfaces

Computational Storage & Communications Providers

Exploit metadata

Automate adaptation Optimise parallel, resilient & pipelined evaluators Control system & extension behaviour

Activities
Story Line

- The UK e-Science Context
- Example Projects
  - Their demand for Data Integration
- OGSA-DAI
  - Architectural Context
  - An Extensible Framework
  - A kit of parts
  - A way of life
- Summary & Take Home Messages

Challenges versus Goals

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heterogeneity &amp; Variety</td>
<td>Simple operational model</td>
</tr>
<tr>
<td>Complex platform behaviour</td>
<td>Simple application model</td>
</tr>
<tr>
<td>Partial failures</td>
<td>Simple user model</td>
</tr>
<tr>
<td>Partial failures + large tasks</td>
<td>Minimal resource wastage</td>
</tr>
<tr>
<td>Autonomy – ownership</td>
<td>Stability &amp; uniformity</td>
</tr>
<tr>
<td>Independent provision</td>
<td>Simple resource access</td>
</tr>
<tr>
<td>Scale, costs, latency</td>
<td>Good performance</td>
</tr>
<tr>
<td>Vulnerable to misuse</td>
<td>Dependable protection</td>
</tr>
<tr>
<td>Diverse &amp; evolving</td>
<td>Flexible &amp; agile</td>
</tr>
<tr>
<td>Various assets</td>
<td>IPR &amp; assets well protected</td>
</tr>
</tbody>
</table>

- Reputation, equipment, teams, data, algorithms, working practices
Data Services: challenges

- Scale
  - Many sites, large collections, many uses

- Longevity
  - Research requirements outlive technical decisions

- Diversity
  - No “one size fits all” solutions will work
    - Primary Data, Data Products, Meta Data, Administrative data, …

- Many Data Resources
  - Independently owned & managed
    - No common goals
    - No common design
    - Work hard for agreements on foundation types and ontologies
    - Autonomous decisions change data, structure, policy, …
  - Geographically distributed

- and I haven’t even mentioned security yet!

Core features of OGSA-DAI – I

- A framework for building applications
  - Supports data access, insert and update
    - Relational: MySQL, Oracle, DB2, SQL Server, Postgres
    - XML: Xindice, eXist
    - Files – CSV, BinX, EMBL, OMIM, SWISSPROT,…
  - Supports data delivery
    - SOAP over HTTP
    - FTP; GridFTP
    - E-mail
    - Inter-service
  - Supports data transformation
    - XSLT
    - ZIP; GZIP
  - Supports security
    - X.509 certificate based security
Core features of OGSA-DAI – II

- A framework for building data clients
  - Client toolkit library for application developers
- A framework for developing functionality
  - Extend existing activities, or implement your own
  - Mix and match activities to provide functionality you need
- Highly-extensible
  - Customise our out-of-the-box product
  - Provide your own services, client-side support and data-related functionality
- Comprehensive documentation and tutorials
- Latest release supports GT3.2 (to be deprecated), GT4.0, and Axis 1.2 / OMII_2 using Java 1.4

E-Infrastructure: the path to Nirvana 1

- Users: Individual & Organisations
- Discipline Specific Toolsets Languages & Portals
- User Mobility
- Stability Extended with Improved Abstractions
- Integration, Sharing, Trust building, Resource Managing Distributed System Infrastructure
- Data Information & Knowledge Providers
- Compute Storage & Communications Providers
- New technology
E-Infrastructure: the path to Nirvana 2

Integration, Sharing, Trust building, Resource Managing, Distributed System Infrastructure

Users: Individual & Organisations

Data Information & Knowledge Providers

User Adaptation Layer

Stability + Higher-order Operations + Improved (semantic) Meta data

Compute Storage & Communications Providers

E-Infrastructure: the path to Nirvana 3

Integration, Sharing, Trust building, Resource Managing, Distributed System Infrastructure

Users: Individual & Organisations

Data Information & Knowledge Providers

User Adaptation Layer

Stability + Higher-order Operations + Improved (semantic) Meta data

Compute Storage & Communications Providers

Increasing Self-organisation & Autonomic Behaviour

Automated Generation Of Stability Adapters & Translators
Summary: Take home message

- E-Infrastructure is arriving
  - Built on Grids & Web Services
  - Data and Information grow in importance
- There is a dramatic rate of change
- An opportunity for everyone

Can you ride the wave?

Questions & Comments
Please
Distributed Systems to Grids

Bespoke pioneering distributed systems, e.g. SABRE & SAGE

Linklater proposes shared multi-site computing

ARPA net
IBM CICS
Ethernet
TCP

Dozens of academic networks
Two dominant protocols

CORBA & DCOM

IP-based Internet
Academic & Research
WWW


Distributed Systems to Grids

Bespoke pioneering distributed systems

Linklater proposes shared multi-site computing

IBM CICS

ARPA net

Ethernet
TCP

Dozens of academic networks
Two dominant protocols

CORBA & DCOM

IP-based Internet
Academic & Research
WWW

Many research grids
Using many protocols & M/W stacks
Web Services

Unicore
Globus
I-way

Collaboration via shared bio/chem/medical “DBs”

Condor

D-Grid