UK e-Science

Grid Infrastructure meets Biological Research Challenges

Malcolm Atkinson

Director of National e-Science Centre

www.nesc.ac.uk

2nd October 2002

The UK Biological Grid — Data and Computation

The Wellcome Trust Genome Campus

Hinxton, Cambridgeshire
Overview

- UK e-Science
  - Reminder of Investment and Infrastructure
- International e-Science
  - Examples and Collaboration
- Data Access and Integration
  - Lego Bricks for Scientific Application Developers
- A Computer Scientist’s View of Biology
  - Diversity and Opportunity
- The Way Ahead
e-Science

- Fundamentally about Collaboration
  - Sharing
    - Ideas
    - Thought processes and Stimuli
    - Effort
    - Resources
  - Requires
    - Communication
    - Common understanding & Framework
    - Mechanisms for sharing fairly
    - Organisation and Infrastructure

Scientists (Biologists) have done this for Centuries
Fundamentally about Collaboration

- Sharing
  - Ideas
  - Thought processes and Stimuli
  - Effort
  - Resources
  - Communication
  - Common understanding & Framework
  - Mechanisms for sharing fairly
  - Organisation and Infrastructure

Text, digital media, structured, organised & curated data, computable models, visualisation, shared instruments, shared systems, shared administration, …
Nationally & Internationally Distributed, …
Routine, Daily, Automated, …

That Requires very Significant Investment in Digital Systems and their Support
e-Science (take 3)

- Fundamentally about Collaboration
  - Sharing
    - Ideas
    - Thought processes and Stimuli
    - Effort
    - Resources
  - Requires
    - Communication
    - Common understanding & Framework
    - Mechanisms for sharing fairly
    - Organisation and Infrastructure

The Grid SHOULD make this much easier by providing a common, supported high-level of Software and Organisational infrastructure.

Digital networks, digital work-places, digital instruments, …

Metadata, ontologies, standards, shared curated data, shared codes, …

Common platforms, shared software, shared training, …

Authentication, Authorisation, Accounting, Provenance, Policies, …

Shared Provision of Platform,
Grid Expectations

- **Persistence**
  - Always there, Always Working, Always Supported

- **Stability**
  - You can build on foundations that don’t move

- **Trustworthy & Predictable**
  - Honours commitments
    - Digital policies, digital contracts, security, …
    - Data integrity, longevity and accessibility
    - Performance

- **High-level & Extensible**
  - The capabilities you need are already there

- **Ubiquitous**
  - Your collaborators use it
Grid Reality

- **Persistence**
  - Always there, Always Working, Always Supported

- **Stability**
  - You can build on foundations that don't move

- **Trustworthy & Predictable**
  - Honours commitments
  - Digital policies, digital contracts, security, …
  - Data integrity, longevity and accessibility
  - Performance

- **High-level & Extensible**
  - The capabilities you need are already there

- **Ubiquitous**
  - Your collaborators use it

Political, Economic & Technical issues to Solve

Early days but Open Grid Services link with Web Services + GGF standardisation

Not yet but very substantial global effort to achieve this

Good basis for extension Commitment to basic functionality WS + Community effort

Global & Industrial Rallying Cry Must work with Web Services
UK Grid Network

Access Grid
always-on video walls

National e-Science Centre

HPC(x)
National e-Science Centre

- **Events**
  - Workshops
  - Research Meetings
  - International Meetings

- **History of Events**
  - GGF5
  - HPDC11
  - Summer school
  - > 50 workshops held
  - > 1000 people in total
  - Many return often

- **Planned Events**
  - 25 workshops
  - Conferences to 2005

- **Visitors**
  - 3 arrived
  - 4 arranged

- **International collaboration, visits & visitors**
  - China
  - Argonne National Lab
  - SDSC
  - NCSA
  - ...

- **Centre Projects**
- Pilot Projects
- Regional Support
- Research Projects
  - EPSRC, MRC, WT, SHEFC
A day in the life of NeSC
Online Access to Scientific Instruments

DOE X-ray grand challenge: ANL, USC/ISI, NIST, U.Chicago

From Steve Tuecke 12 Oct. 01
DataGrid Testbed

- HEP sites
- ESA sites

- Testbed Sites

Francois.Etienne@in2p3.fr - Antonia.Ghiselli@cnaf.infn.it
A Simplified Grid Anatomy

Scientific Users

- Monitoring
- Diagnosis
- Logging
- Scheduling
- Accounting
- Authorisation

Scientific Application

Grid Plumbing & Security Infrastructure

- Owners
- Data & Compute Resources

Application Developers

Operations Team

Distributed
A Biological Grid Anatomy

Biological Users

Scientific Application

Monitoring  Diagnosis  Data Integration

Scheduling  Accounting  Data Access

Grid Plumbing & Security Infrastructure

Data & Compute Resources

Structured Data

Distributed
Database Growth

EMBL Database Growth
total record number (millions)

[Bar chart showing EMBL database growth over years]

PDB protein structures

[Bar chart showing PDB protein structures over years]
Scientific Data

- **Deluge of Data**
  - **Exponential growth**
    - Doubling times
      - Astronomy: 12 months
      - Bio-Sequences: 9 months
      - Functional Genomics: 6 months
      - Bytes/dollar: 12 to 18 months

- **Not How big it is but**
Scientific Data

- Deluge of Data
  - Exponential growth
    - Doubling times
      - Astronomy: 12 months
      - Bio-Sequences: 9 months
      - Functional Genomics: 6 months
      - Bytes/dollar: 12 to 18 months

- Not How big it is but
- What you do with it
  - Sharing
  - Curation
  - Metadata
  - Automated movement, access & integration
  - Computational Access
Scientific Data

- Deluge of Data
  - Exponential growth
    - Doubling times
      - Astronomy: 12 months
      - Bio-Sequences: 9 months
      - Functional Genomics: 6 months
  - Bytes/dollar: 12 to 18 months

Not how big it is but

- How you Embrace & Manage Change
  - The Database is a Knowledge chest
  - The Database is a Communication Hub
  - Autonomously Managed (Curated) change
  - An Essential part of e-BioMedical Science
Wellcome Trust: Cardiovascular Functional Genomics

Glasgow

Shared data

Edinburgh

Leicester

London

Oxford

Public curated data

Netherlands
Data Access & Integration

- Central to e-Science
  - Especially Earth Sciences, Ecology, Biology & Medicine

- Collaboration
  - Shared Databases
  - Curated Knowledge
  - Accumulated Observations
  - Accumulated Simulations

- Computation
  - Data mining
  - Input to models
  - Calibration of models

- Presentation
  - Publication of results
  - Visualisation
GGF DAIS WG

- **Chairs**
  - Norman Paton (Manchester Uni.)
  - Leanne Guy (CERN)
  - Dave Pearson (Oracle UK)

- **Activity**
  - BoF GGF4 Toronto
  - WG Meeting GGF5 Edinburgh
  - Papers for GGF6
  - Workshops & Mail lists

- **Goals**
  - Agree Standards for Database Access & Integration
  - Freely available reference implementations
    - OGSA-DAI one source & focus for discussions
OGSA-DAI project

- Lego kit for Data Access & Integration
  - Components for e-Science Applications
  - Accelerated Application Development
  - Multiple Data Models
  - Distributed Data
  - Access via Grid & Proxies
  - Integration, Translation & Transformation

- Open Source Reference Implementation
  - For DAIS-WG standard

- Trigger for Component Construction
  - Start a community
OGSA-DAI Partners

£3 million, 18 months, started February 2002

EPCC & NeSC
IBM UK
IBM USA
Manchester e-SC
Newcastle e-SC
Oracle
Primary Components
Advanced Components

Client

GDS:PerformScript

GDS

DB

Translation

GDT

Translation

Consumer
Composed Components

Translation

GDS

Client

GDS:performScript

GDS:performScript

GDS:performScript

GDT

Translation

GDT

Consumer

GDS:performScript

GDS:performScript

GDS:performScript
Distributed Query

DQP : Distributed Query Processor
GDT : Grid Data Transport
T : Translation
Q : Query
GDTV : Grid Data Transport Vehicle
F : Factory
QPM : Query Progress Monitor
PNM : Progress Notification Message
AM : Application Metadata
CRM : Computational Resource Metadata
NS : Notification Sink
OGSA-DAI Time Line

WS + GSI UK support (> 100 downloads)
XML + OGSA Prototypes for Early Adopters

Design Documents & Demos for DAIS WG @ GGF5

XML + OGSA Prototype Available
RDB + GT2 / OGSA Prototypes Available

GGF6 WG Papers & Prototypes
Ship Alpha Release for GT3 Integration
Presentation & Beta @ GGF7
Productisation, RAMPS & Extension

Feb '02 May '02 Jul '02 Sep '02 Dec '02 Feb '03 May '03 Sep '03
Phase 1 Starts Phase 2 Starts
OGSA-DAI Summary

- On Schedule & Going Well
- Contributions via DAIS-WG @ GGF5 & 6
- Releases with GT3 Releases scheduled
- Status: Early Days
  - Released prototypes
  - Tested Architectural Design
  - Using OGSA
  - Working with Early Adopter Pilot Projects
    - AstroGrid & MyGrid
- Influence OGSA-DAI direction
  - Via DAIS-WG & Direct messages to us
Biomedical e-Scientists

- Is this one species?
  - Understanding bird energy
  - Understanding a river / ocean interaction
  - Understanding a biochemical pathway
  - Understanding a cell
  - Understanding a Heart or Brain
  - Understanding Rhododendra
  - Understanding Evolution
  - ...

- No One-Size fits all solutions
  - But sharable re-usable components
Many, many …
  - More than we can address
  - Compute needs
  - Data management needs
  - Data integration needs
  - …

Must choose some pioneers
  - To meet a range of common requirements
  - To provoke rich & high-level platform
  - To generate re-usuable components

A Long-Term Commitment Needed
Advancing Biological Grid

Biological Users

Scientific Application

Biomedical (Grid) Application Component Library

Monitoring  Diagnosis  Data Integration
Scheduling  Accounting  Data Access

Grid Plumbing & Security Infrastructure

Data & Compute Resources  Structured Data

Distributed
Summary

- **e-Science**
  - Data as well as Compute Challenges
    - Needed to be put together
  - Need ubiquitous supported consistent platforms

- **Grid**
  - A (potentially) invaluable platform
  - Only show in town

- **Data Integration**
  - Hard \(\Rightarrow\) Develop & Use Standard kit of parts
  - Started to build the kit

- **Opportunities**
  - No one-size fits all, but re-usable subsystems
  - Invest in wider range of Problem driven pioneering
  - Strategic choices needed