Learning from e-Science: Grids and e-Infrastructure at GridKa School Karlsruhe

Malcolm Atkinson
Director

www.nesc.ac.uk
26th September 2005

Contents

- History & Motivation for e-Science & Grids
  - Linklatter to Foster & Kesselman to ...
- What are Grids & Web Services
  - Disruptive technology advances
  - Rallying cries
- Example Grids & Projects
  - National Grids, Open Science Grid, BIRN, GEON, ...
  - Data and Information – Biomedicine, Engineering & Finance
- Example Middleware Project
  - OGSA-DAI – international use and collaboration
- Reaping the Benefit of Grids
  - What is the impact on your research
- Future e-Infrastructure
  - Ambient Knowledge-based Computational Infrastructure
What is e-Science?

- **Goal:** to enable better research in *all* disciplines

- **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*
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


Collaboration via shared bio/chem/medical "DBs"

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

D-Grid starts

Condor

Unicore
Globus
I-way

Uniprotected Internet
Academic & Research
WWW

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

---

Why use / build Grids?

- **Computer Science Arguments**
  - New attempt at an old hard problem
  - Frustrating ignorance 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
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

Contents
- History & Motivation for e-Science & Grids
  - Linkletter to Foster & Kesselman to ...
- What are Grids & Web Services
  - Disruptive technology advances
  - Rallying cries
- Example Grids & Projects
  - National Grids, Open Science Grid, BIRN, GEON, ...
  - Data and Information – Biomedicine, Engineering & Finance
- Example Middleware Project
  - OGSA-DAI – international use and collaboration
- Reaping the Benefit of Grids
  - What is the impact on your research
- Future e-Infrastructure
  - Ambient Knowledge-based Computational Infrastructure
Grids: a 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

Diagram derived from Ian Foster’s slide
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
  - Much more than just providing services

Providing mutually consistent services with compatibly policies

Why are distributed systems hard

- They are necessary to our modern way of life
  - telephone, airline booking, internet, web, financial trading
    - Access to remotely curated data and remote / mobile resources
    - Integration of company-wide IT support for R&D
- Global, always on, single-purpose systems
  - Hard to build, finance and manage
  - Often based on sharing a common (may be replicated) database
Why are distributed systems hard

- Global, always on, multi-purpose systems
  - Never been done – can it be done?

Only by a huge international collaborative effort

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 operation</td>
<td>Simple resource access</td>
</tr>
<tr>
<td>Scale, cooperation</td>
<td>Good performance</td>
</tr>
<tr>
<td>Vulnerable to attack</td>
<td>Dependable protection</td>
</tr>
<tr>
<td>Diverse &amp; evolving</td>
<td>Flexible &amp; agile</td>
</tr>
<tr>
<td>Valuable assets</td>
<td>IPR &amp; assets well protected</td>
</tr>
</tbody>
</table>
  - Reputation, equipment, teams, data, algorithms, working practices
The Primary Requirement ... Building people grids

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

Contents

- History & Motivation for e-Science & Grids
  - Linklatter to Foster & Kesselman to ...
- What are Grids & Web Services
  - Disruptive technology advances
  - Rallying cries
- Example Grids & Projects
  - National Grids, Open Science Grid, BIRN, GEON, ...
  - Data and Information – Biomedicine, Engineering & Finance
- Example Middleware Project
  - OGSA-DAI – international use and collaboration
- Reaping the Benefit of Grids
  - What is the impact on your research
- Future e-Infrastructure
  - Ambient Knowledge-based Computational Infrastructure
How Grid Software Works: NSF Network for Earthquake Engineering Simulation (NEES)

Transform our ability to carry out research vital to reducing vulnerability to catastrophic earthquakes

Ian Foster
Grid2003: An Operational Grid
- 28 sites (2100-2800 CPUs) & growing
- 400-1300 concurrent jobs
- 8 substantial applications + CS experiments
- Running since October 2003

http://www.ivdgl.org/grid2003
Database Growth

EMBL Database Growth

PDB Content Growth

BRIDGES

Glasgow
Edinburgh
Leicester
Oxford
London
Netherlands

Publically Curated Data

Ensembl
MGI
HUGO
OMIM
SWISS-PROT

DATA HUB

CFG Virtual Organisation

Information Integrator

Synteny Grid Service

VO Authorisation

Glasgow
Edinburgh
Oxford
Leicester
London
Netherlands

Private data

Private data

Private data

Private data

Private data

Private data

Public data
Mammography

A prototype of a national database of mammographic images in support of the UK breast screening programme

Mammograms have different appearances, depending on image settings and acquisition systems

eDiaMoND: Screening for Breast Cancer

1 Trust \(\Rightarrow\) Many Trusts
Collaborative Working
Audit capability
Epidemiology

Provided by eDiamond project: Prof. Sir Mike Brady et al.
Virtual Observatories

Observations made across entire electromagnetic spectrum

⇒ e.g. different views of a local galaxy
Need all of them to understand physics fully
Databases are located throughout the world

Peter Clarke

Contents

- History & Motivation for e-Science & Grids
  - Linklatter to Foster & Kesselman to ...
- What are Grids & Web Services
  - Disruptive technology advances
  - Rallying cries
- Example Grids & Projects
  - National Grids, Open Science Grid, BIRN, GEON, ...
  - Data and Information – Biomedicine, Engineering & Finance
- Example Middleware Project
  - OGSA-DAI – international use and collaboration
- Reaping the Benefit of Grids
  - What is the impact on your research
- Future e-Infrastructure
  - Ambient Knowledge-based Computational Infrastructure
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!

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 you project using
  – Additional Activities
  – Client Toolkit APIs
  – Data Resource handlers

• A base for higher-level services
  – federation, mining, visualisation,…
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

Slide from Neil Chue Hong

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

Slide from Neil Chue Hong
International Collaboration & Use

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

Japan:
- AIST
- BioGrid
- NAREGI

Europe:
- CERN
- DataMiningGrid
- GridMiner
- GridSphere
- Intergrid
- NGrid
- OntoGrid
- Provenance
- SIMDAT
- OMII-EU

UK:
- OMII
- OMII-UK
- NGS
- NCeSS
- NIeeS
- AstroGrid
- BioSimGrid
- BRIDGES
- CancerGrid
- ConvertGrid
- eDaMoND
- EDINA
- First Group plc
- Fujitsu Labs Europe
- GEODA
- GeneGrid
- Genomic Technology and Informatics
- GOLD
- Human Genetics Unit
- IBM UK
- iCluster
- Oracle UK

China:
- CAS
- ChinaGrid
- cnGrid
- INWA
- China-MIT
- OMII-China

Australia:
- Curtin Business School
- INWA

Tutorials
- Boston
- CERN
- Edinburgh
- San Francisco
- Seoul
- Tokyo

Cambridge
Chicago
London
Seattle
Singapore
ISSGC 03 to 05

1485 registered users
5250+ downloads

Meeting User Requirements

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

- History & Motivation for e-Science & Grids
  - Linklatter to Foster & Kesselman to ...
- What are Grids & Web Services
  - Disruptive technology advances
  - Rallying cries
- Example Grids & Projects
  - National Grids, Open Science Grid, BIRN, GEON, ...
  - Data and Information – Biomedicine, Engineering & Finance
- Example Middleware Project
  - OGSA-DAI – international use and collaboration
- Reaping the Benefit of Grids
  - What is the impact on your research
- Future e-Infrastructure
  - Ambient Knowledge-based Computational Infrastructure

It’s a Bonanza - can you keep your head?

- Your insights in your subject
  - Foundation step in discovery
  - Directs your line of work
  - Selecting well is still a hard challenge
- A Jungle of Opportunities
  - Retain judgement & sense of direction
  - Don’t be distracted by technology, ...
- Be a Smart Tool User
  - Understand and exploit tools
  - Demand good & persistent tools
  - Early adopters may win or be mired in a techno-bog
Be a Smart Data User

- Choosing data sources
  - How do you find them?
  - How do they describe and advertise the data?
  - Is the equivalent of Google possible?
- Obtaining access to that data
  - Overcoming administrative barriers
  - Overcoming technical barriers
- Understanding that data
  - The parts you care about for your research
- Extracting nuggets from multiple sources
  - Pieces of your jigsaw puzzle
- Combing them using sophisticated models
  - The picture of reality in your head
- Analysis on scales required by statistics
  - Coupling data access with computation
- Repeated Processes
  - Examining variations, covering a set of candidates
  - Monitoring the emerging details
  - Coupling with scientific workflows

∴ Your model ≠ their model
⇒ Negotiation & patience needed from both sides

Be a Smart Collaborator

- Understand & Negotiate
  - Position, Role & Responsibilities
- Deliver on your promises
  - On time
- Record and State Provenance of Work
- Build international & national collaboration
  - Investment in communication and learning
- Expect good collaboration from others
CS: Take Home Message

- There are plenty of Research Challenges
  - Workflow & DB integration, co-optimised
  - Distributed Queries on a global scale
  - Heterogeneity on a global scale
  - Dynamic variability
    - Authorisation, Resources, Data & Schema
    - Performance
  - Some Massive Data
  - Metadata for discovery, automation, repetition, ...
  - Catalogues
  - Optimising Data replication & Data movement
  - Provenance tracking

- Grasp the theoretical & practical challenges
  - Working in Open & Dynamic systems
  - Incorporate all computation

Contents

- History & Motivation for e-Science & Grids
  - Linklatter to Foster & Kesselman to ...
- What are Grids & Web Services
  - Disruptive technology advances
  - Rallying cries
- Example Grids & Projects
  - National Grids, Open Science Grid, BIRN, GEON, ...
  - Data and Information – Biomedicine, Engineering & Finance
- Example Middleware Project
  - OGSA-DAI – international use and collaboration
- Reaping the Benefit of Grids
  - What is the impact on your research
- Future e-Infrastructure
  - Ambient Knowledge-based Computational Infrastructure
E-Infrastructure Nirvana

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

Users: Individual & Organisations

DIK Adaptation Layer

Data Information & Knowledge Providers

CSC Adaptation Layer

Compute Storage & Communications Providers

User Adaptation Layer

E-Infrastructure: the path to Nirvana

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

Users: Individual & Organisations

Stability Extended with Improved Abstractions

Discipline Specific Toolsets Languages & Portals User Mobility

Data Information & Knowledge Providers

New technology

User Adaptation Layer

Compute Storage & Communications Providers
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

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

Increasing Self-organisation & Autonomic Behaviour

Automated Generation Of Stability Adapters & Translators

Compute Storage & Communications Providers
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