

Informing Business & Regional Policy:
Grid-enabled fusion of global data &
'local' knowledge

INWA

Ally Hume

Terry Sloan

Adam Carter

EPCC

Ashley Lloyd

Curtin Business School &

The University of Edinburgh Management School

- ▶ The Grid vision
- ▶ The INWA project
- ▶ Data mining over the Grid
- ▶ Barriers encountered & Conclusions

The Grid Vision

“... flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources-what we refer to as virtual organisations.”

The Anatomy of the Grid: Enabling Scalable Virtual Organizations. I. Foster, C. Kesselman, S. Tuecke. *International J. Supercomputer Applications*, 15(3), 2001.

|epcc|

The INWA Project

- ▶ **Funded by UK Economic & Social Research Council (UK) in the Pilot Projects in E-Social Science**
 - Small scale projects to explore the potential of Grid technologies within the social sciences
- ▶ **Started November 2003, finishes August 2004**
- ▶ **10 partners involved from both academia & commerce (finance & telecoms)**
 - 6 located in UK (EPCC, UEMS, LUMS, Bank, ESRC, ESPC)
 - 4 located in Australia (Curtin, Telco, Property data provider, Sun Microsystems)

▶ The INWA vision

- To improve business decision-making by using a Grid infrastructure to establish secure communications between data-holders with market knowledge and experienced analytical scientists.

▶ The INWA Objectives

- Migrate the process of analysing corporate data to a secure Grid environment
- Deliver a Grid-enabled data mining tool set.
- Establish a grid portal in the Asia-Pacific region
- Demonstrate model building on real corporate data from global companies through interaction across the Grid
- Assess the benefits to decision-making of aggregation
- Assess the potential for fusing private and public data models at different levels of aggregation

▶ Resources

- UK mortgage data
- UK property data
- Australian telco data
- Australian property data
- Compute power at EPCC
- Compute power at Curtin

▶ Individuals and Organisations:

- Analyst at EPCC, UK
- Analyst at Curtin, Australia
- EPCC, UK – compute resource provider and host
- Curtin, Australia – compute resource host
- Sun Microsystems, Aus – compute resource provider
- Bank, UK – data provider
- ESPC, UK – data provider
- Telco, Aus – data provider
- VGO, WA, Aus – data provider

- ▶ Can existing grid technologies fulfill this vision?
 - TOG from Sun DCG provides access to remote HPC resource
 - OGSA-DAI provides access control and discovery of distributed heterogeneous data resources
 - FirstDIG grid data service browser provides SQL access to OGSA-DAI enabled resources
 - Globus Toolkit 2 and 3
- ▶ If not what are the barriers?
 - Technology?
 - Social factors?

Data Mining over the Grid

- ▶ A typical data mining project broadly involves
 1. Getting the data
 2. Cleaning it
 3. Mining it
- ▶ Iteration through steps 1 to 3 to refine models
- ▶ So where can the Grid help ?
 - ...

▶ Traditionally a file export

- But OGSA-DAI is available
 - Open Grid Services Architecture : Data Access and Integration
 - Assists with the access and integration of data from separate data sources via the Grid
- But organisations will not contemplate external access to operational/sensitive data
- So back to a file export

▶ UK Land registry

- Public data source but no OGSA-DAI interface
- Appropriate mechanisms need to be in place before data sharing can take place

▶ So simulated this access over the Grid

- But some security issues

▶ Fusing commercial data with public property data

<i>Account ID</i>	<i>Address</i>	<i>Loan</i>	<i>Date</i>	...
2289738	10 Downing Street, ...	200,000	10/2/2002	...
2672623	20 My Street, ...	100,000	14/8/1980	...

<i>Address</i>	<i>#Bedrooms</i>	<i>#Garages</i>	...
10 Downing Street, ...	4	3	...
20 My Street, ...	3	0	...

<i>Account ID</i>	<i>Address</i>	<i>Loan</i>	<i>Date</i>	<i>#Bedrooms</i>	<i>#Garages</i>	...
2289738	10 Downing ...	200,000	10/2/2002	4	3	...
2672623	20 My Street, ...	100,000	14/8/1980	3	0	...

- ▶ Why do it ?
 - Prospect of better models/predictions
 - Added value
- ▶ But
 - need a distributed-aggregated approach to preserve anonymity
- ▶ So simulated this over the Grid
 - Using a less specific join key
 - Not a 1-1 join but a 1-n so averaging necessary
 - Limited the potential gains from fusion
- ▶ Fuzzy joins
 - e.g. postcode formats, addresses (St=Street, flat numbers)

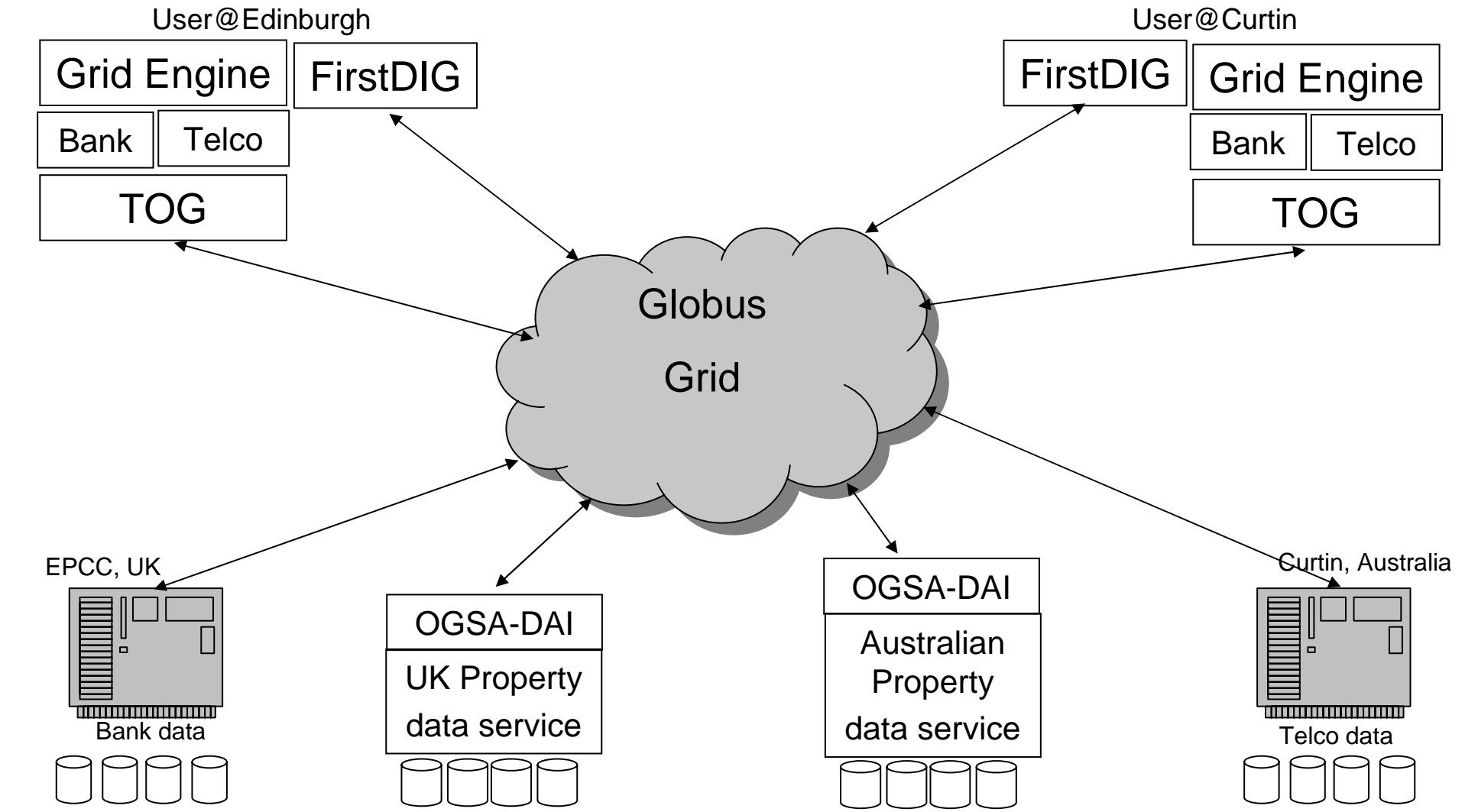
- ▶ Little real support for data integration over the Grid
 - OGSA-DQP is limited (Linux, OQL,..)
- ▶ Used FirstDIG browser
 - Relevant data pulled over
 - Data joined locally
 - This works but obviously is not ideal
- ▶ So again limited success over the Grid

The screenshot shows the 'First Data Service Browser - demonstrator' application window. It features a menu bar with 'File', 'Database Activity', and 'Help'. The main interface is divided into several sections:

- Service Group Registries:** A text box containing the URL `http://localhost:8080/ogsa/services/ogsadai/DAIServiceGroupRegistry`. Buttons for 'Add Registry ...' and 'Remove Registry' are to the right.
- GDS Factory URLs (databases):** A list of URLs including `http://129.215.62.133:8080/ogsa/services/ogsadai/DatabaseC`, `http://129.215.62.133:8080/ogsa/services/ogsadai/GDSFFirstMileage` (highlighted), and `http://129.215.62.133:8080/ogsa/services/ogsadai/GDSFFirstCCS`.
- SQL Statement:** A text area containing the query: `SELECT DAY,SUM(ACTLVLMIS) AS lostmiles FROM history GROUP BY day`. Buttons for 'Run Select Query', 'Run Update Query', 'Save Query ...', and 'Load Query ...' are to the right.
- Query Results:** A separate window displaying a table with two columns: 'DAY' and 'lostmiles'. The table contains data for dates from 2002-12-15 to 2003-01-10.

DAY	lostmiles
2002-12-15	12.06
2002-12-16	490.81000000...
2002-12-17	363.76000000...
2002-12-18	339.19
2002-12-19	297.84
2002-12-20	396.85
2002-12-21	51.57
2002-12-22	28.5600000000...
2002-12-23	441.06000000...
2002-12-24	23.39
2002-12-27	172.34
2002-12-28	101.11
2002-12-29	36.8700000000...
2002-12-30	604.36
2002-12-31	112.23
2003-01-02	158.63
2003-01-03	64.03
2003-01-04	69.79
2003-01-05	3.11
2003-01-06	202.77999999...
2003-01-07	236.35999999...
2003-01-08	282.80999999...
2003-01-09	203.86999999...
2003-01-10	255.46000000...

- ▶ Large data sets so, ...
- ▶ Cleaning and mining jobs sent to where data is resident (UK and Australia)
- ▶ Globus Toolkit V2.x (GT2), Grid Engine & Transfer-queue Over Globus (TOG) used
- ▶ But...
 - Installation issues with GT2
 - Security issues with GT2 & TOG
- ▶ All now works and is currently being used between UK and Australia



Conclusions

▶ **Trust**

- Dynamic, virtual organisation is simulated rather than created
- Organisations understandably wary about installation of software and the access it provides

▶ **Market**

- Not clear if data providers will publish data via web/grid service interfaces such as OGSA-DAI

▶ **Security, Security, Security**

- Not strong enough

▶ **Software**

- Not robust enough

- ▶ Simulation demonstrated the potential of a virtual organisation consisting of data providers and analytical scientists
- ▶ For this application, grid technologies not mature enough to support the concept of a dynamic, virtual organisation
 - Do not provide necessary security and robustness to instill trust
 - Is there a business benefit to outweigh the cost of addressing the risks ?
- ▶ Project contacts
 - <http://www.epcc.ed.ac.uk/~inwa>
 - inwa@epcc.ed.ac.uk