

## Linking e-Science capabilities for e-Social Science communities:

### *- extending the UK-Australia INWA project to the Chinese Academy of Sciences*

A.D. Lloyd<sup>1,2</sup> K. Nan<sup>3</sup> D. Qian<sup>4</sup> T.M. Sloan<sup>5</sup> Y. Sun<sup>2</sup> B. Yan<sup>3</sup>

<sup>1</sup>Edinburgh University Management School

<sup>2</sup>Curtin Business School

<sup>3</sup>Chinese Academy of Sciences

<sup>4</sup>Beihang University

<sup>5</sup>EPCC



- ▶ e-Science capabilities vs. e-Social Science collaboration
- ▶ INWA: large-scale observations of behavioural dynamics
- ▶ Extending the INWA Grid to China
- ▶ Concluding Remarks

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

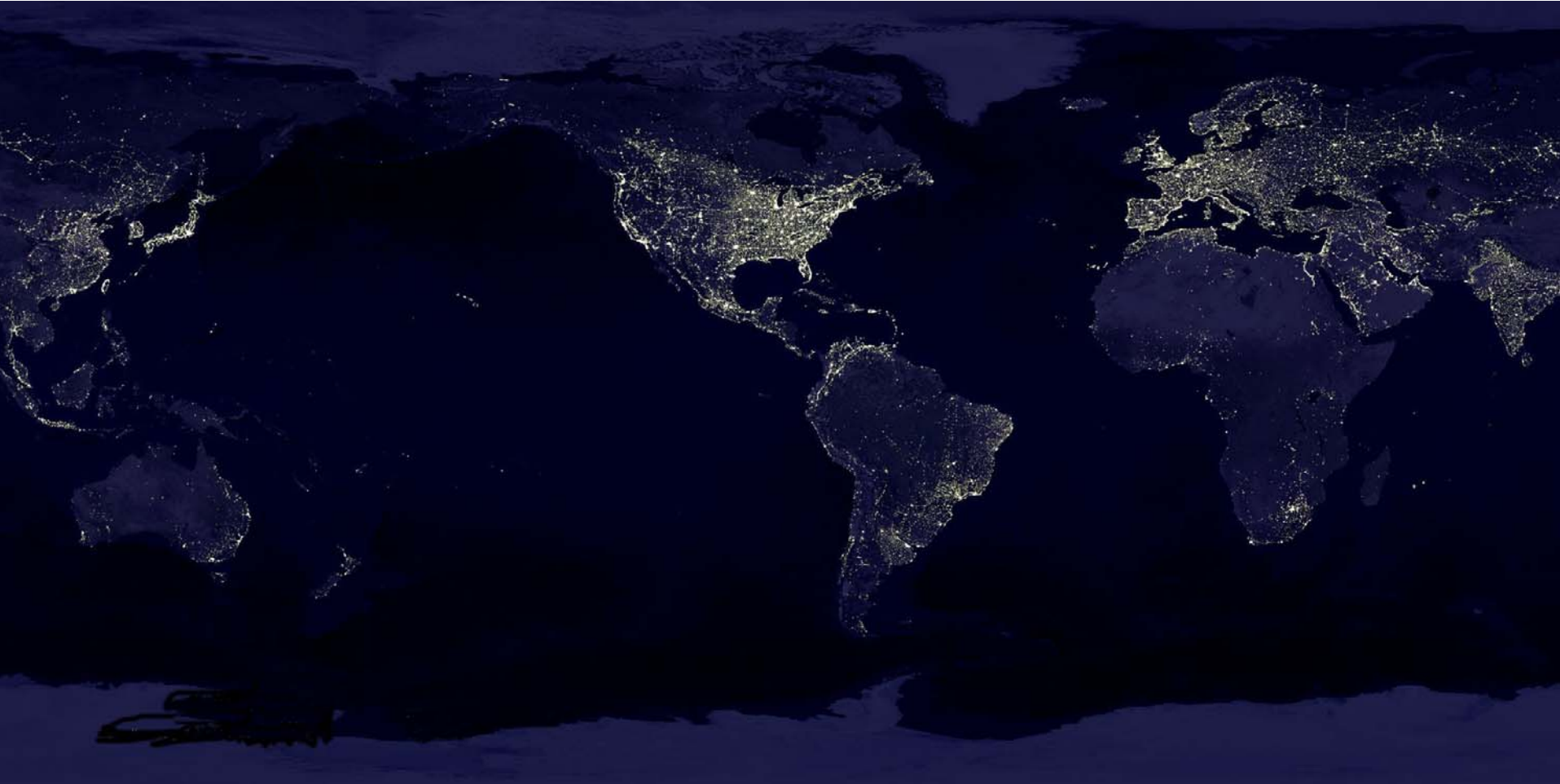
**The Anatomy of the Grid: Enabling Scalable Virtual Organizations.** I. Foster, C. Kesselman, S. Tuecke. *International J. Supercomputer Applications*, 15(3), 2001.

“...most social science research is done within national and local boundaries, and, most often by individual scholars rather than the research teams which populate medical and natural science research.”

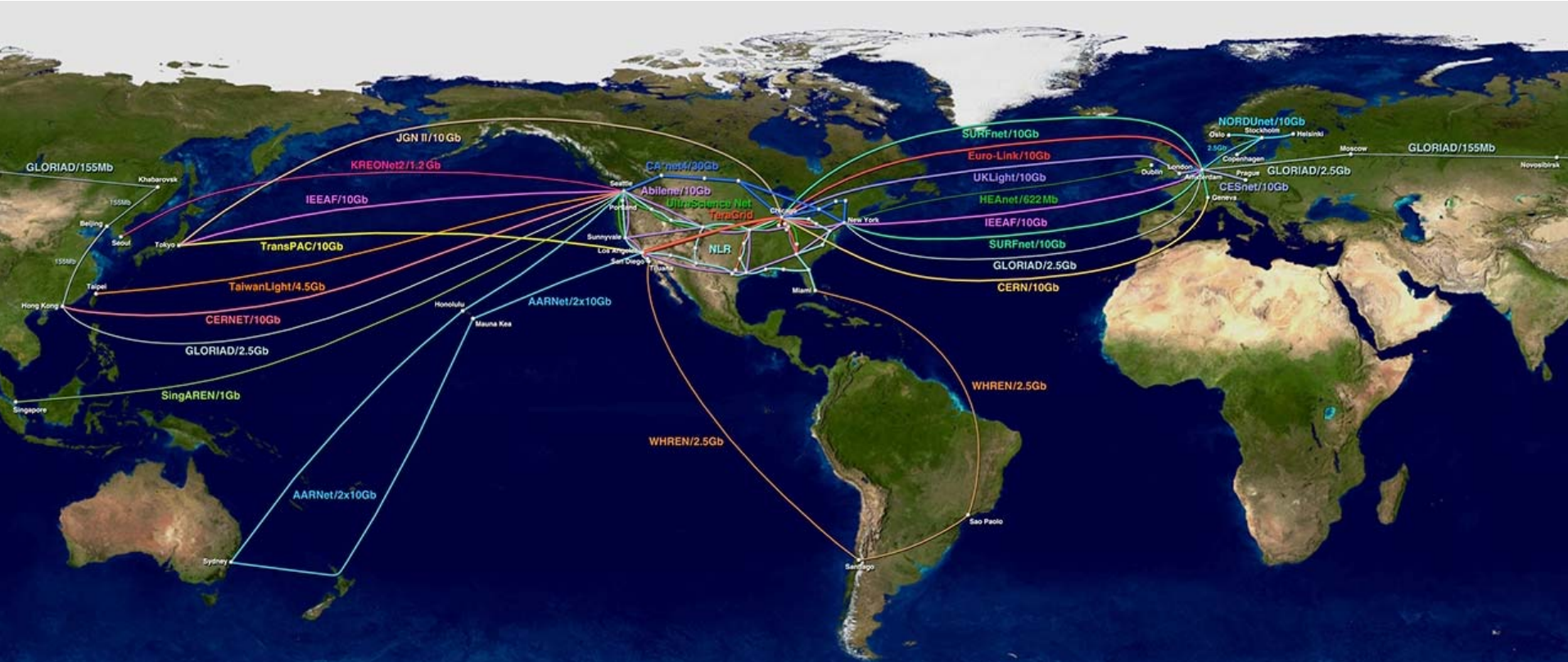
‘**International social science research: craft industry or baby behemoth?**’, Forbes, I. and Abrams, D. (2004), *International Social Science Journal*, vol. 56 no.180, pp. 189-192

## ▶ Forbes and Adam (2004)

- whilst innovative social science research often does make reference to global discourses, this kind of “cutting edge” research does not constitute the bulk of research reported in the social science major journals
- Conclude ideal conditions exist for ‘big social science’, but does not require the major infrastructure of an organization like CERN, rather “a much more distributed human infrastructure of researchers in multiple locations.... responding to big research questions by providing data about how very local phenomena articulate with more global phenomena,”
- question whether a global infrastructure for eScience can support international social science research.



eScience capabilities vs. Social Science collaboration

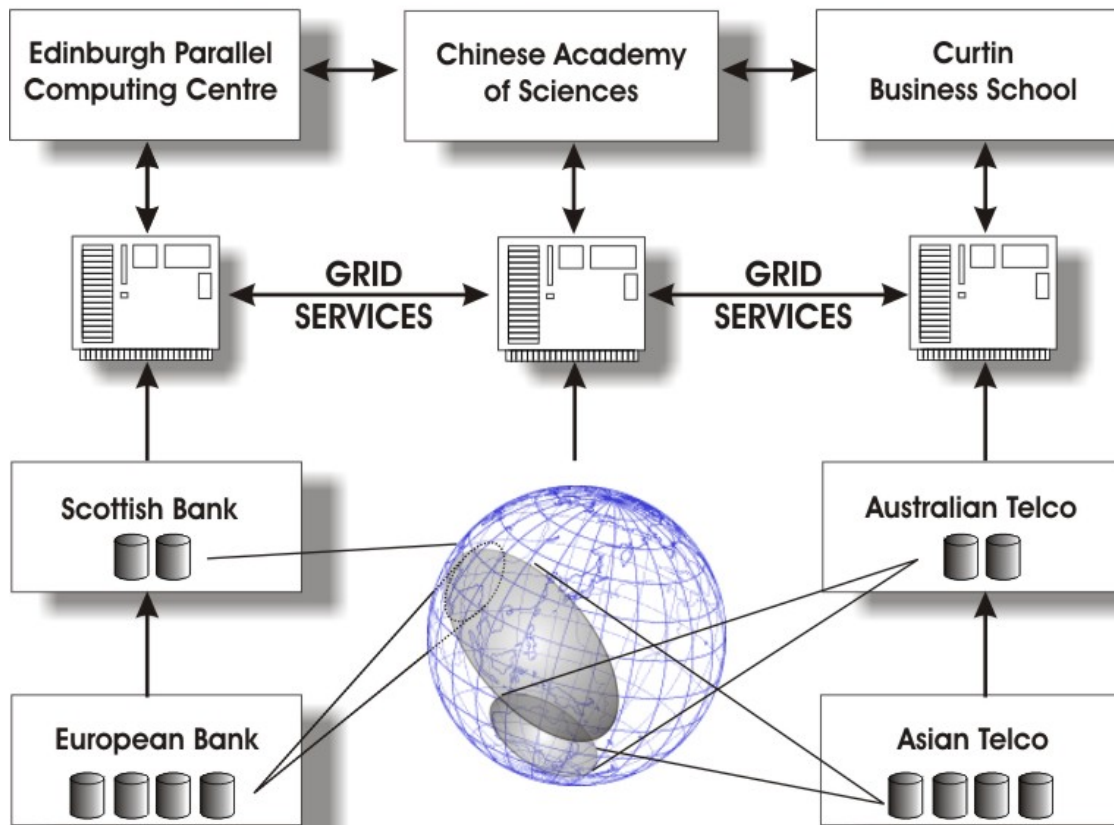


- are e-Social Science requirements for 'collaboratories' supported by a "robust, international Cyberinfrastructure" [Hey, 2005] established for eScientists?

- ▶ **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
  - “Informing Business & Regional Policy: Grid enabled fusion of global data & local knowledge”
  - INWA : Innovation Node Western Australia
- ▶ **Project Aims**
  - Evaluate the suitability of existing grid solutions for secure distributed data mining and analysis on commercially sensitive data to predict customer behaviour
  - Investigate the advantages of fusing public and private data enabled by a grid environment
- ▶ **Project phases**
  - 1<sup>st</sup> phase November 2003 to August 2004
    - Set up INWA grid between UK and Australia
    - Data mining of commercial data between sites over the grid
  - 2<sup>nd</sup> phase started November 2004, to April 2005
    - Addition of a node in China to the existing INWA grid
    - Basic data mining over the grid to China
    - Acquisition of Chinese commercial data
  - 3<sup>rd</sup> phase May 2005 onwards
    - Acquisition of commercial Chinese data
    - Data mining of Chinese commercial data over the grid



DATA - EXPERTISE - TOOLS - HPC



- large samples taken from telecommunications and financial services markets,
- analysis within distributed team setting supported by HPC-on demand, leading to
- predictive models of behaviour, with accuracies in excess of 80%
- similar studies in process with Chinese Academy of Sciences
- establishing this uncovered the social-shaping of some Grid technologies

INWA: large-scale observations of behavioural dynamics

**1999 – 2001: National High Performance Computing Environment (NHPCE) as part of Ministry of Science and Technology's (MoST) National 863 Program. 8 Major super-computing centres connected. Now part of 863 High Performance and Grid Computing project and China National Grid (CNGrid)**

**2003: ChinaGrid Project, aims to fully utilize the 15,000 GFLOPS of resources on CERNET (China Education and Research Network). CERNET already connects over 900 education and research institutions, 1.2 million PCs, 8 million users and will be extended to incorporate a further 200,000 schools with over 175 million users.**

**2004: Chinese Academy of Sciences (MoST) focuses on the Scientific Data Grid to share resources, and the China Science Grid to integrate instruments and resources within application grids for specific communities, e.g. the Virtual Observatory and Bioinformatics.**

**2005: China Next Generation Internet (CNGI) is the largest planned IPv6 network in the world, but...**

## Also ...

- ▶ Perth/BeiJing time zone
  - 2/3 of world GDP growth in 2003
  - highest proportional increase in investment of any region in the world
- ▶ Chinese Academy of Sciences (CAS) manage
  - entire .cn domain
  - traffic between government, industry and academia across a common network

## But ...

- ▶ 4 legal jurisdictions
  - Scotland
  - England
  - Australia
  - China
- ▶ Agreements in Mandarin and English
- ▶ Government control of access to commercial data

- ▶ The INWA project is one of the ESRC's Grid Pilot projects. In December 2003 it successfully linked Grid resources at EPCC and Curtin Business School, and in January 2005, it extended the infrastructure to include the Chinese Academy of Sciences.
- ▶ It has employed and contributed to the development of a wide range of core grid technologies and succeeded in integrating them to support distributed data mining.
- ▶ The facility for focusing expertise and delivering the HPC required for analysis, provides a window on behavioural dynamics in global markets and a model for e-Social Science collaborations to explore them
- ▶ The process of connecting three continents and maintaining operation through various generations of Grid middleware has highlighted a number of regional differences that are socially-shaped and may impact on future modalities of collaborative social science
- ▶ An abstraction of this work has been shown today – more details on the technology are given in the following talk and at AHM2005 and are available from [www.epcc.ed.ac.uk/inwa](http://www.epcc.ed.ac.uk/inwa)

## 5. Concluding Remarks

# Extending the INWA Grid: the technical challenges

T.M. Sloan<sup>3</sup>, A.D. Lloyd<sup>1,2</sup>

<sup>1</sup>Edinburgh University Management School

<sup>2</sup>Curtin Business School

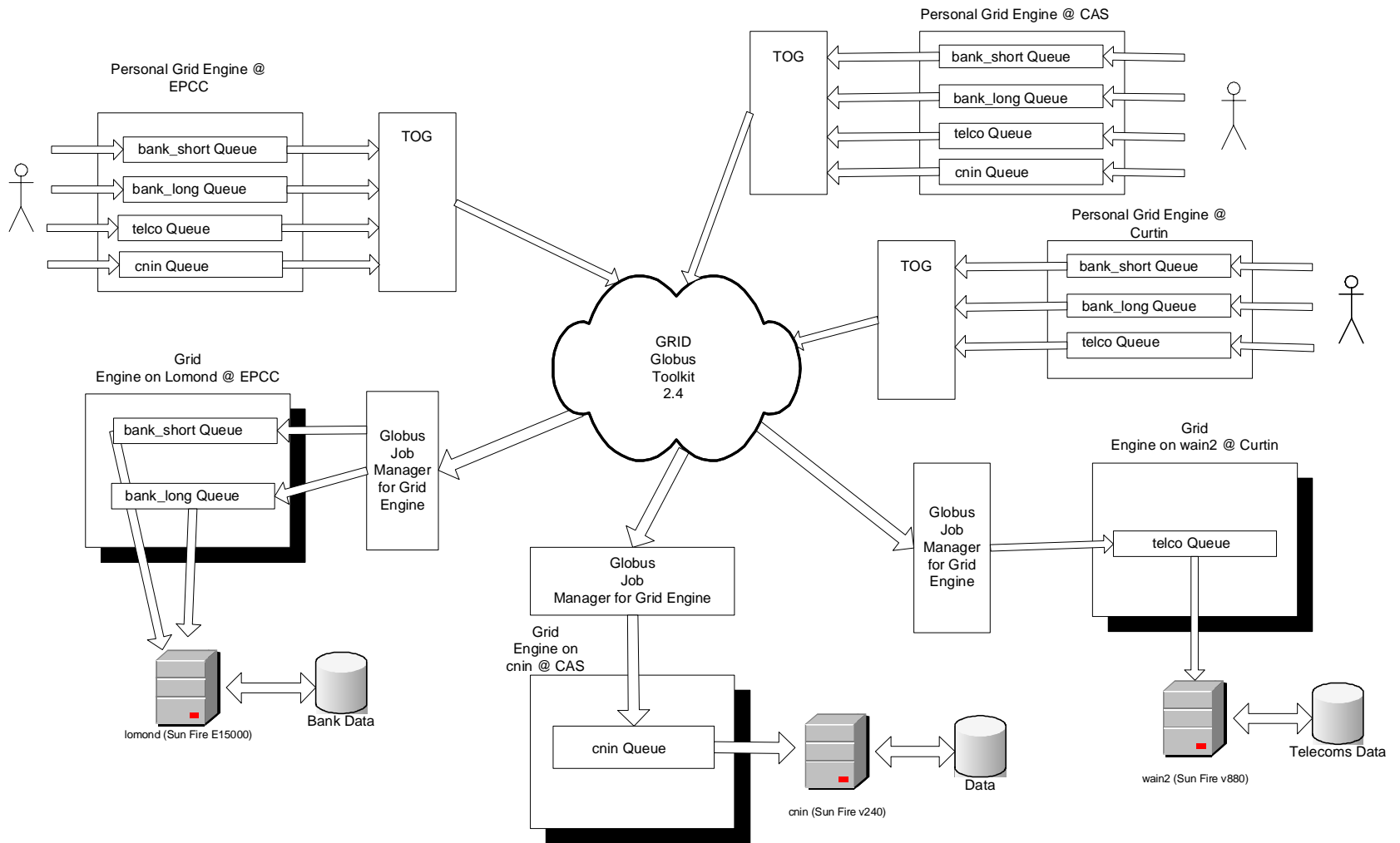
<sup>3</sup>EPCC

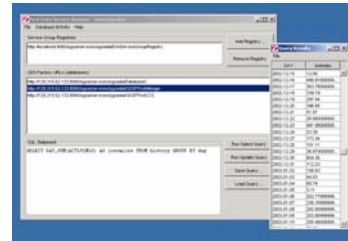


- ▶ INWA Grid technical infrastructure
- ▶ Technical Challenges

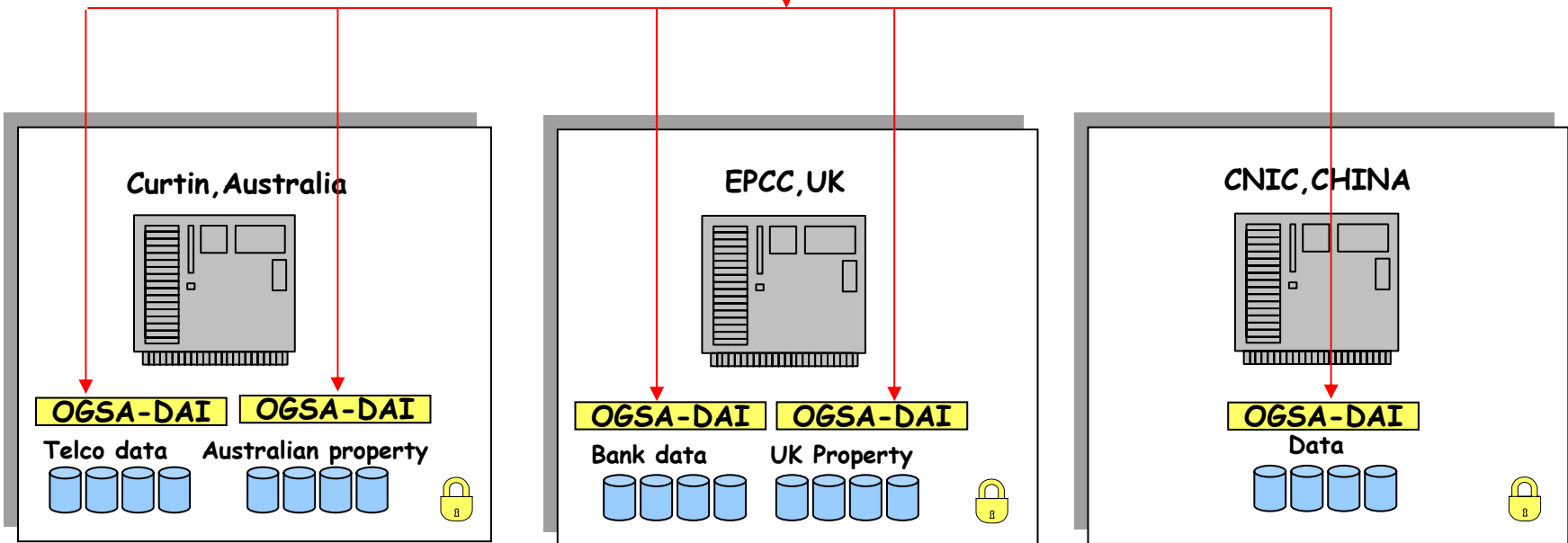
- ▶ Grid software employed to provide
  - Job submission infrastructure
    - Batch jobs for data preparation, cleaning, mining
  - Access to data resources
    - Data browsing, integration
- ▶ For analysis of customer behaviour
- ▶ Installed and operational across 3 sites, on 3 continents

- ▶ **Transfer-queue Over Globus (TOG) v1.1** from the UK e-Science Sun Data and Compute Grids project
  - provides access to remote HPC resource
  - Uses Globus Toolkit 2.4
- ▶ **Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) Release 3.1, Release 5.0**
  - provides access control and discovery of distributed heterogeneous data resources
  - Uses Globus Toolkit 3.0
- ▶ **First Data Investigation on the Grid (FirstDIG)**
  - grid data service browser provides SQL access to OGSA-DAI enabled resources
  - now part of OGSA-DAI R4.0/5.0
- ▶ **Globus Toolkit 2.4 and 3.0**
  - Grid middleware





Data Browser



- ▶ Network routing stability
- ▶ Reverse Domain Name Service (DNS) Lookup
- ▶ OGSA-DAI installation and performance
- ▶ Open Database Connectivity

- ▶ Initial deployment of GT 2.4 at CAS hampered by instabilities in Korea
- ▶ Required assistance from AARNet in Australia and KOREN in Korea
  - Australian Academic and Research Network
  - Korea Advanced Research Network
- ▶ This stabilised connections between Australia and China

- ▶ Reverse Domain Name Service (DNS) lookup

```
tms@e3500$ nslookup -sil 129.215.56.231
Server:          129.215.56.230
Address:         129.215.56.230#53
231.56.215.129.in-addr.arpa    name = e3500.epcc.ed.ac.uk
```

- ▶ Required by GT2: only for sustaining connections not establishing them

But ...

- ▶ In China, few IP addresses relative to demand
- ▶ Usually not possible to configure reverse DNS look-up at same DNS server that handles usual forward DNS lookup
- ▶ Had to explicitly configure INWA participating machines

## ▶ Release 3.1 installation

- Required database driver update
- On stabilisation of network testing was successful

## ▶ Release 5.0 installation

- To address 3.1 security and performance issues
- Minor software configuration
- Manual intervention for registering data services
- Some deficiencies in documentation since rectified in 6.0
- Internal data browser issues means custom client application needed for large result sets > 20000

- ▶ No general purpose statistics package supports OGSA-DAI interface
- ▶ Typically these packages support ODBC (Object Database Connectivity)
- ▶ Therefore built a prototype ODBC OGSA-DAI driver
- ▶ Tested successfully between UK and Australia with OGSA-DAI R3.1 data sources
- ▶ A production quality driver would enable e-Social Scientists to more easily benefit from the Grid

- ▶ The INWA Grid has demonstrated grid interoperation between three sites: Edinburgh – UK, Perth – Australia, and Beijing – China
- ▶ The INWA Grid has been used for analysis of real commercial data to understand customer behaviour
- ▶ The process of extending the grid infrastructure to China uncovered a number of new technical and socio-legal challenges that in part reflected differences in access to grid technologies in China and approaches to international collaborations between academia, industry and government
- ▶ Demonstrated that global e-Social Science collaboration based on e-Science infrastructure is possible.
- ▶ Creating, operating and exploiting a Grid for e-Social Science still requires significant levels of expertise and the supporting software could be improved further but the situation is improving all the time.