

Extending the INWA Grid

T. M. Sloan¹, A. D. Lloyd^{2,3}

¹EPCC, The University of Edinburgh, Scotland, UK

²Curtin Business School, Curtin University of Technology, Perth, Australia

³The University of Edinburgh Management School, Scotland, UK

t.sloan@epcc.ed.ac.uk

Abstract. The INWA project (Innovation Node: Western Australia) has previously investigated and reported on the use of existing Grid technologies for secure data mining of corporate data and the use of that data to build predictive models of consumer behaviour. The INWA grid infrastructure between the UK and Australia that was built and utilized during these investigations has recently been extended to include China. This paper introduces the INWA project, describes its technical grid infrastructure and reviews challenges encountered when upgrading the grid infrastructure and extending it to China.

Introduction

The INWA project was funded by the UK Economic and Social Research Council as part of their programme on ‘Pilot Projects in e-Social Science’. The project’s full title is ‘Informing Business and Regional Policy: Grid-enabled fusion of global data and local knowledge’. The project is a collaboration between various academic and commercial organisations. EPCC, the University of Edinburgh Management School and the Lancaster University Management School are the academic partners in the UK, with Curtin Business School from the Curtin University of Technology, the academic partner in Perth, Australia. The Chinese Academy of Sciences (CAS) in Beijing, China have now joined this collaboration. The commercial partners include Sun Microsystems along with partners from the financial and telecommunications sectors in the UK, Australia and now China.

Using Grid technologies, the INWA Grid allows a globally distributed research team to draw on local market knowledge when analysing commercial data drawn from commercial partners in the global Telecommunications and Financial Services sectors. Using data mining techniques, this data can then be converted into models to explain and predict the behaviour of consumers of a product or service, allowing companies to improve the management of customer relationships for existing products, and innovate in response to the emergent demand for new products and services that the analysis also targets.

The addition of a link with China provides an important insight into market dynamics under conditions of high growth, and since the Chinese Academy of Sciences manages the entire .cn domain and traffic between government, industry and academia across a common network infrastructure, the project also provides an interesting insight into a unique information infrastructure for innovation.

Following the description of this latest incarnation of the INWA Grid's technical infrastructure, the paper reviews some of the technical challenges encountered when upgrading the INWA Grid and extending it to China.

The INWA Grid

The INWA Grid is built from existing freely available Grid technology. A Job Submission Grid Infrastructure allows users to submit batch data processing jobs for execution on machines located at the various participating sites. This Job Submission Grid Infrastructure uses Globus Toolkit v2 for the Grid middleware, Grid Engine v5.3 as the compute resource manager and Transfer-queue Over Globus (TOG) to transfer jobs and results between local and remote sites (Hume et al, 2004).

Grid Engine is an open source distributed resource management system that allows the efficient use of compute resources within an organisation. The Globus Toolkit is essentially a Grid API for connecting distributed compute and instrument resources via the internet. TOG integrates Grid Engine v5.3 and Globus Toolkit v2 to allow access to remote resources at any collaborating enterprises. Building a suitable infrastructure for INWA from these components involves much more than merely plugging them together. It requires a significant breadth and depth of technical expertise. For example, the project staff worked with both TOG and Globus developers in order to ensure jobs and data are transferred securely. Systems and networking expertise were required to verify and validate the TOG and Globus updates.

The Job Submission Grid infrastructure now comprises three compute resources located at EPCC in the UK, Curtin in Australia and CAS in BeiJing, China. Researchers at EPCC are now able to submit jobs to process data located on remote resources in BeiJing as well as Edinburgh and Perth. The TOG software will transfer any output files produced by the job back to the researcher's local workstation in EPCC. Researchers at CAS in BeiJing are similarly able to submit jobs to process data located in Edinburgh or Perth. With a view to enhancing this job submission infrastructure in the future, JOSH (JOB Scheduling Hierarchically), the grid services based successor to TOG, has been installed and tested successfully between Australia and the UK. JOSH allows the easier addition of collaborating sites, provides access to third-party data sources and assigns a user's job for execution at a particular site based on data proximity, load and capability (Cawood et al, 2004).

OGSA-DAI (Open Grid Services Architecture – Data Access and Integration) provides a Grid service interface to data resources that allows for data access and integration (Hume et al, 2004). Initially in the INWA Grid, OGSA-DAI v3.1 has been used to provide access to the relational data sources at the various sites with the FirstDIG data browser. The FirstDIG data browser allows users to interact with OGSA-DAI enabled data sources via SQL queries. Data sources located at CAS in BeiJing have now been added to the existing list of data sources from EPCC in Edinburgh and Curtin in Perth that are accessible via this mechanism. There are known limitations with OGSA-DAI v3.1 most notably in security and when handling large datasets (Hume et al, 2004). The OGSA-DAI installations in China and Australia have therefore been subsequently updated to OGSA-DAI v5. This newer version of OGSA-DAI also includes an enhanced version of the FirstDIG data browser.

Challenges encountered when extending the INWA Grid: evidence of the Social Shaping of Grid Technologies

This section of the paper reviews some of the technical challenges encountered when extending the INWA Grid. Some technical challenges reflected the expected standardisation difficulties between carriers when running services between three continents and across multiple jurisdictions. Others provide examples of the social shaping of grid technologies.

Network Routing Instability

The initial deployment of Globus Toolkit v2.0 at CAS was hampered by network routing instabilities in Korea. With the assistance of AARNet in Australia and KOREN in Korea this was identified and network links between Australia and China were stabilised.

Reverse Domain Service (DNS) Lookup

Reverse Domain Name Service (DNS) Lookup is where an internet IP address of the form xxx.xxx.xxx.xxx is converted to the hostname by searching through domain name service tables (See Figure 1 for an example). This facility is required by Globus Toolkit v2 when connecting to remote machines and hence is necessary for the INWA Job Submission Grid Infrastructure.

```
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.
```

Figure 1. An example domain service lookup

In China there are very few IP addresses available relative to their demand. This means that organisations find it difficult to obtain a whole IP segment. In turn this means that it is typically not possible to configure reverse DNS lookup at the same DNS server that handles the normal forward DNS lookup. In China, therefore only forward DNS lookups tend to be used.

Diagnosing this problem was complicated by the fact that reverse DNS lookups are not required for establishing Grid connections, but are usually required for sustaining them. This created a connection fault that was difficult to separate from instability in network routing. Following the previously mentioned stabilisation of the network links with China, the issue was resolved by configuring the local hosts file on those UK and Australian machines participating in the INWA Grid to return the appropriate hostname for the participating Chinese machine. This allowed the INWA job submission infrastructure to be completed.

OGSA-DAI Installation and Performance

Initial problems with the installation of OGSA-DAI v3.1 appeared to be related to those hampering the Globus Toolkit installation, however on stabilisation of the network links and a database driver update, the installation and subsequent testing were successful. The subsequent installation of OGSA-DAI v5.0 required some minor alterations to the supporting software configuration. In particular, manual intervention was required when registering data

services for use in the INWA grid. In v5.0, services can now be deployed that enforce security. Whilst upgrading to this, deficiencies in the security documentation were identified and filtered back to the OGSA-DAI team for rectification in the forthcoming v6.0 documentation. Problems internal to the v5.0 data browser mean there is still no straightforward way to access large OGSA-DAI v5.0 enabled data resources without writing a custom client application that will support data streaming rather than bulk transfer, and consequently queries no larger than approximately 20000 entries are possible.

Open Database Connectivity

No general purpose software, such as interactive statistics packages support OGSA-DAI interface (Hume et al, 2004). The INWA project has therefore investigated whether data extraction and processing programs used in the area of social sciences, and which utilise ODBC data resources could use OGSA-DAI as their back-end data resource. Such programs typically focus on data extraction rather than data updates. Using an evaluation license for a commercial ODBC Software Development Kit (SDK), a prototype ODBC driver has been built. This allows an interactive SQL application to connect via ODBC with an OGSA-DAI v3.1-enabled data resource at Curtin in Australia. This prototype allows queries to be performed on the data resource. The provision of such a production-quality ODBC driver for OGSA-DAI that supported data streaming and hence grid access to large datasets would allow e-Social Scientists to more easily benefit from the grid.

Conclusions

The INWA Grid has now demonstrated full grid interoperation between three sites: Edinburgh – UK, Perth – Australia, and Beijing – China. 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.

References

- Cawood, G., Seed, T., Abrol, R. and Sloan, T. (2004): ‘TOG & JOSH: Grid scheduling with Grid Engine & Globus’, in S. J. Cox (ed.): *Proceedings of the UK e-Science All Hands Meeting 2004*, EPSRC, 2004, pp. 46-53.
- Hume, A. C., Lloyd, A. D., Sloan, T. M. and Carter, A. C. (2004): ‘Applying Grid Technologies to Distributed Data Mining’, in H. Jin, Y. Pan, N. Xiao and J. Sun (eds.): *GCC 2004, LNCS 3251*, Springer-Verlag Berlin Heidelberg, 2004, pp. 696-703.

Acknowledgments

The authors gratefully acknowledge the funding of this work by the UK’s Economic and Social Research Council under the award number RES-149-25-0005. In addition the authors gratefully acknowledge the assistance of K. D’Mellow, M. Jackson, L. Pottage and A. Hume in extending the INWA Grid, Sun Microsystems for providing hardware and support in Beijing and Perth and AARNet for network routing assistance.