

Grid-enabling Social Scientists: the FINGRID infrastructure

Lee Gillam, Khurshid Ahmad and Gary Dear

Department of Computing, School of Electronics and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom

{l.gillam, k.ahmad, g.dear}@surrey.ac.uk

Abstract. This paper discusses a Grid implementation for facilitating large-scale social sciences research. Grid-enabling resources, the computers, programs and data, necessitates technological planning, even in hard science disciplines like particle physics and astronomy where the resources are well identified. Grid middleware is inherently complex, covering a broad spectrum of functionality, and requires some efforts to understand and use effectively. This paper describes how emergent grid middleware (Globus Toolkit, OGSA-DAI, Condor, SRB) was used to create a 24-node grid for managing, retrieving and analysing “live” data. The application area was financial economics and econometrics, where time serial and textual data were analysed concurrently: speed-ups of upto 5-10 times were noted in both cases. The Grid is being used for capacity building in both research and teaching.

Introduction

The development of a ‘grid’ has to take into account software emerging from large consortia, for example the Globus Alliance, for providing access, sharing and visualisation, together with the configuration of complex hardware systems. Existing analytical techniques can then be refined, extended, Grid-enabled and integrated, for example for use in distributed aircraft management (DAME: Austin 2004). Such integration may involve attempting to provide what appears to be a homogeneous solution, based on heterogeneous resources, for example programs written for different operating systems. In the e-Social Science project **Financial Information Grid** (FINGRID: Ahmad et al 2004a) we developed a software prototype that enabled us to test the capability of Grid middleware for providing access to the computers, data and programs needed to undertake the analysis of financial information. Using Grid middleware, we refined, extended, Grid enabled and integrated programs for the analysis of live financial data: both quantitative in the form of time series, and qualitative in the form of financial news texts. As a result, FINGRID was able to use a 24 node computational environment for experimental and benchmarking purposes to quantify throughput capabilities for the analytical techniques involved, and hence to evaluate benefits of Grid technologies.

In this paper we provide an overview of the technologies involved in FINGRID: the computers, the Grid middleware, the data provision, and the legacy systems for qualitative and quantitative analysis. Grid-enabling resources, including computers, programs and data, necessitates technological planning, even in hard science disciplines like particle physics and astronomy where the resources are well identified. Grid middleware is inherently complex, covering a broad spectrum of functionality, and requires some efforts to understand and use effectively. We outline the provisions of FINGRID, how they were combined. Surrey is

using this provision to support Grid-based PhD research topics and the teaching of a Masters-level Grid computing module. These technologies and their combination may be relevant to other e-Social Science applications also.

Requirements

The aim of FINGRID was to develop a Grid-based demonstrator for sharing and analysing real-time financial data streams, time series of financial returns and financial news. This involved:

- qualitative analysis: content analysis of collections of text, specifically financial news, using an existing information extraction and text management system, System Quirk, written largely in Java with some legacy software in C, C++ and Prolog (Ahmad & Rogers 2001, Gillam 2004 and references therein)
- quantitative analysis: time series data, specifically of prices of financial instruments, using a Monte Carlo simulator (written in FORTRAN) and a wavelet-based approach to non-stochastic signals (facilitated by MATLAB)
- fusion of quantitative and qualitative data by correlating the time series with the results of information extraction (sentiment analysis, Gillam 2002, Ahmad et al 2004)

To provide a demonstrable Grid-based system, FINGRID needed to integrate Surrey's System Quirk software, signal processing tools from MATLAB, programs for bootstrapping, and the financial data from Reuters.

Building the Financial Information Grid

This section describes the hardware/software configuration of our grid. We then discuss issues related to grid middleware and issues related to the acquisition of *live* data. Security certification is outlined followed by the description of Surrey's Financial Information Grid.

Configuration

For our configuration, we initially used 5 Dell Optiplex GX150 machines (each with 60GB HDD – upgraded from original specification, 256MB RAM, 1GHz processors) and 3 Dell PowerEdge 2650 machines (each with 70GB HDD 1GB RAM, dual hyperthreading 2.4GHz processors). We created 2 bootable images that would provide configuration of Red Hat on each type of machine; each machine can be semi-automatically configured (within an hour), including provision of a variety of tried-and-tested software packages, particularly certain program compilers: the Java Development Kit (JDK), FORTRAN, C, C++ and SICStus Prolog. By the end of FINGRID, this provision included a further 16 2650's, bringing our total capacity to 24 machines providing 1.6 terabytes of disk space, 20GB memory and 96GHz processing. This effort was based on extensive prior "art" in hardware/software configuration; re-use of technology, Optiplexes that are now four years old and counting, provided the initial basis for this configuration, and the initial successes of this configuration enabled the rapid expansion of our setup.

Grid Middleware

Grid technologies initially considered for this provision were the Globus Toolkit, the Open Grid Services Architecture Data Access Infrastructure (OGSA-DAI), Condor, and the Storage Resource Broker (SRB). The Globus Toolkit is a software package that provides security, information infrastructure, resource management, data management and other facilities. It is a *de facto* standard for creating Grids. The resources, infrastructure and so forth, have to be “Globus-enabled” to allow them to be used in combination – generally this involves providing a “wrapper” of some form. Similar wrappers are required for these other Grid technologies, for example the creation of job submission files and “DAGs” in Condor. We had previously succeeded in experimental installations of Globus Toolkit version 2 (GT2) on Linux (Red Hat version 6.2), and since OGSA-DAI, Condor, and SRB all appear to be more easily supported and offer greater configurability in Unix-based systems, we decided to install GT3 and other Grid technologies under Linux (Red Hat version 7.3) also.

Each machine had been configured with a disk partition specifically for installation of the Grid software and to provide Grid experimental areas. Our Grid software installation incorporates:

- Globus Toolkit version 3.0.2, plus the Globus Simple Certification Authority (CA) package and the Java Commodity Grid (v1.1)
- OGSA-DAI version 3.0.2, with specific requirements for, knowledge of, and cross-configurations with multiple technologies including: Jakarta Log4j for debugging purposes; JUnit for testing purposes; MySQL and Xindice, optional database technologies; Apache Tomcat for providing Web Services; and Globus.
- Condor version 6.6.6 (superstition notwithstanding), providing a Condor pool containing 74 processors (Linux), numerically larger than that reported by CCLRC¹
- SRB version 3.2.1 using a database (Oracle) for storing metadata relating to files distributed across a virtualized file system.

System Quirk requires the Oracle database also, and we were able to use FORTRAN for compiling the bootstrap code, provided by the kickstart image, and had the MATLAB package installed by the image also. Replication across machines has so far proven to be a successful means of increasing the size of the Grid. The inexorable march towards the next version of the RedHat OS indicates that it will be necessary to recompile a number of these applications prior to further deployment. The National Grid Service (NGS) provides a RedHat ES 3.0 environment, so wider deployment may make demands on such local infrastructure upgrades. Currently, the stable release of Condor is 6.6.9; Globus is at 3.2.1, with development versions of GT4 available; OGSA-DAI is at release 5. Version compatibility/interoperability may be an additional concern for Grid developers necessitating additional resourcing.

Live Data

Reuters data is provided under contractual agreement via leased equipment comprising 2 dedicated telephone lines for current and historical data (128kbps and 64kbps), a PC feed

¹ The CCLRC reports a Condor pool running at Daresbury and Rutherford Appleton, containing 21 Linux processors - <http://tardis.dl.ac.uk/Condor/> (15 April 2005)

manager and a Sun workstation with presence on the network, but no analytical capability or Grid-enablement, for client request management. We interface to live and historic Reuters data via the Reuters Java Source-Sink Library Software Development Kit (SSL SDK). Others wishing to replicate such a setup should be aware of costs in the region of £25k per annum.

Security Issues

Once software installations were complete, security certificates were issued to enable full use of the Globus packages by both students and researchers, enabling single sign-on access. We currently maintain separation between teaching machines and research machines, based on groups of users determined at the OS level; researchers can make use of both. In lieu of a better solution, we created two separate instances of Globus *grid-mapfile*, used for authentication, according to the purpose of the machine. Duplicating entries for the *grid-mapfile* for researchers remains a manual task. We have yet to explore the implications of providing authorisation for external users – provisions under the e-Science programme require appointments and visits to centres for individual user registration.

With data distribution being restricted contractually, secure provision of Reuters data will be a further challenge if we were to attempt to migrate local service offerings such as those of FINGRID to provisions of the NGS. Additional layers of authorisation and accounting may be required: OS-mediated group memberships is, perhaps, a bare minimum, and some form of, perhaps micropayments-based, pay-per-use or pay-per-k (kilobyte) model for the data may be needed subject to negotiation. Assuring secure access to data and appropriate use and accounting will be an interesting challenge.

Integration: the FINGRID demonstrator

The FINGRID demonstrator used the Java CogKit to integrate (wrap):

- System Quirk components via the Quirk Java SDK
- Reuters data via the Reuters SSL SDK
- the MATLAB wavelet toolbox via JMatLink
- bootstrap simulation written in FORTRAN

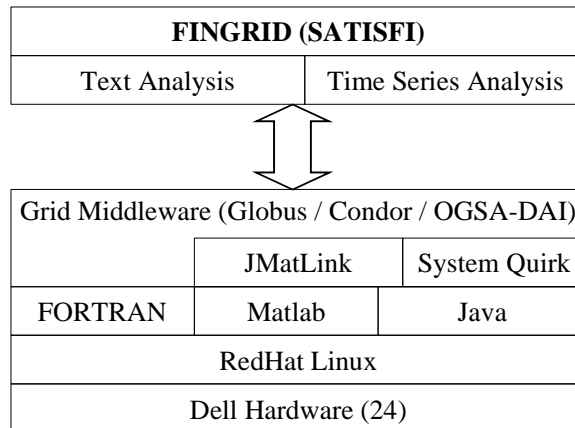


Figure 1: Snapshot of the hardware and software composition of the FINGRID demonstrator. The diagram does not include the Reuters components.

Primarily, FINGRID makes use of job management aspects of Globus, the Resource Specification Language (RSL), a wrapper syntax describing the environment required by a process, the Globus Resource Allocation Manager (GRAM), enabling the process to be executed and monitored, and the Global Access to Secondary Storage (GASS), to enable transfer of job output. There is some overlap between descriptions using RSL syntax under Globus job submission and description files under Condor: specification of the program to run, input/output requirements, environment requirements, and so forth. Both provide a means by which native programs, such as those compiled using FORTRAN or C, and more OS-independent programs, such as those compiled using Java, can be run on other available machines. This may necessitate code refactoring (changing internal structure without affecting behaviour) to produce fragments that can be submitted and run in parallel, with results being recombined from these parallel fragments. This refactoring was necessary for the bootstrap program. Job monitoring with Globus and Condor necessitates different considerations – notification of job completion being a good example: documentation on Condor generally indicates a fire-and-forget approach in which the system emails you when a job is complete. We are yet to explore the interface of Condor and Globus – CondorG – to get the best out of both of these technologies.

Since the data for the analysis are located on the local network, there has been little need to evaluate the capabilities of the GridFTP software for moving large text corpora between machines. With the ability to use SRB for data management, by which data stores distributed across machines can be virtualized anyway, further consideration of the relative merits of each technology remains a goal.

As reported elsewhere, our Grid-based text analysis and boot-strapping computations could be undertaken more efficiently using the power of the grid compared to use on a single machine (FINGRID project final report; Ahmad, Gillam and Cheng 2005). Such performance improvements suggest the potential for real-time analysis of “living” financial data, with possible industrial applications. Use of leading-edge techniques such as those identified above, both individually and in combination, to analyse various large data sets, both quantitative and qualitative, to distribute the processing needs of these tasks, and to fuse and visualise the results of these analyses, suggests potential benefits for other scientific activities also.

Capacity Building

Our Grid infrastructure has been used for supporting Grid-based PhD research topics and for teaching a Masters level degree module since January 2004². Surrey students can use locally issued Globus certificates to explore fully the features and provisions of Globus and other Grid technologies to help them understand Grid computing and to build their own Grid demonstrators. Students recreate some of the experiments undertaken in FINGRID to engender an understanding of the application of Grid technologies to problems in economics, with potentially wider application to other social sciences: *strengthening the UK's capacity for social science research*.

In our teaching configuration, we encountered problems incurred through a perceived requirement for multiple instance service support using OGSA-DAI such that individuals can experiment with, and have full control over, service provision and use both within and across machines. OGSA-DAI uses Tomcat, which requires 2 communication ports per instance, per machine. To enable each student to undertake service-interaction experiments, the "CATALINA_BASE" configuration was made for each student, with port settings generated from their numeric user ids automatically inserted into their server configuration files. OGSA-DAI files with reference to "localhost:8080" are then modified to refer to this user-port id. A common Tomcat deployment can be used, although similar small-scale efforts are needed here also, including provision of a shared data either under the installation, or in some other shared CATALINA_BASE directory such as the Grid experimental area identified above. The developers of OGSA-DAI have suggested that Globus may be more suited for teaching the basics of Grid service development and deployment.

We aim to continue exploring the benefits and drawbacks of Grid technologies by evaluating Globus, OGSA-DAI, Condor and SRB in contexts that will enable us to discuss the relative use, benefits and drawbacks of each technology, and their combinations, for addressing a variety of social sciences challenges. For example, we have undertaken initial experiments with Condor for distributing the execution of FINGRID's bootstrapping algorithm, which can be partially parallelised. Condor has enabled us to better distribute processes, but provides a challenge in job completion notification for these processes, and the necessary re-combination of the parallel results once all the parallel jobs have completed. For this latter problem, we have investigated use of DAGMan, a Condor-based application for handling interdependencies between submitted jobs, described using a Directed Acyclic Graph (DAG). DAGMan requires additional efforts to be made to provide a meta-level description of the interactions between the various processes, identifying dependencies at each stage. The bootstrap algorithm, written in FORTRAN (Lobato, Nankervis and Savin, 2001), has been decomposed into three separate phases, where the centre phase can be undertaken in parallel and so is suited for description as a DAG. Execution under Condor depends on the availability of the non-dedicated processors, and Condor's prioritisation algorithm needs to be taken into consideration when describing performance. Requirements for the "best" resources (memory, CPU) can be described in Condor's submission files; these jobs will generally be queued along with others with similar requirements until a processor is available (has finished its previously assigned job). We are considering the implications of teaching Condor, including the provision of separate teaching and research pools to demonstrate and evaluate Condor's *flocking* capabilities. Undertaking distributed computation is, of course, possible without using Grid technologies; providing the means to scale such efforts is where one major benefit of Grid technologies lies.

² See <http://www.computing.surrey.ac.uk/courses/csm23/>

Use of this Grid for the automatic annotation of video is being evaluated in the EPSRC-funded project Recovering Evidence from Video by Fusing Video Evidence Thesaurus & Video Meta-Data (REVEAL: GR/S98450/01). In addition, we plan to open talks with the NGS with respect to the possible migration of this localised system necessitating considerations of software licensing conditions for, for example, MATLAB. Work is planned that will explore the potential provision of an ISO-conformant metadata registry for social sciences, exploring the use of ISO-conformant metadata in Data Grids using the Storage Resource Broker (SRB) and its Metadata Catalog (MCAT) (Gillam 2005).

Acknowledgments

This work supported in part by the ESRC (FINGRID: RES-149-25-0028), EU (LIRICS: eContent-22236) and EPSRC (REVEAL: GR/S98450/01). The authors would like to acknowledge the contributions made to this effort by colleagues as identified in the FINGRID project Final Report. The authors are grateful to the anonymous reviewers for their helpful commentary on the original submission.

References

- Ahmad, K., Gillam, L. and Cheng, D. (2005). "Textual and Quantitative Analysis: Towards a new, e-mediated Social Science". First International Conference on e-Social Science, 22-24 June 2005, Manchester, UK
- Ahmad, K., and Rogers, M. A. (2001). "Corpus Linguistics and Terminology Extraction". In (Eds.) Sue-Ellen Wright and Gerhard Budin. *Handbook of Terminology Management (Volume 2)*. Amsterdam & Philadelphia: John Benjamins Publishing Company: 725-760.
- Ahmad, K., Taskaya-Temizel, T., Cheng D., Gillam, L., Ahmad, S., Traboulsi, H. and Nankervis, J. (2004). "Financial Information Grid – an ESRC e-Social Science Pilot". *Proceedings of the Third UK e-Science Programme All Hands Meeting (AHM 2004)*, Nottingham, United Kingdom. Swindon: EPSRC Sept 2004. ISBN 1-904425-21-6.
- Austin, J. (2004). "A grid based diagnostics and prognosis system for rolls royce aero engines: the DAME project". IEEE Proc. of the 2nd International Workshop on Challenges of Large Applications in Distributed Environments, 2004. 7 June 2004. p2.
- FINGRID Pilot Demonstrator Project (ESRC: RES-149-25-0028) Final Report. http://www.computing.surrey.ac.uk/grid/fingrid/papers_files/Reports/FINGRID_final.pdf
- Gillam, L (2005). "ISO standard Metadata Descriptors and Registries". From Metadata Standards to the Data GRID for Comparative Social Research, workshop at the First International Conference on e-Social Science, 22-24 June 2005, Manchester, UK
- Gillam, L. (2004). "Systems of concepts and their extraction from text". Unpublished PhD thesis, University of Surrey.
- Gillam, L. & Ahmad, K. (2002). "Sharing the knowledge of experts". *Fachsprache - The International Journal of LSP*, vol. 24(1-2): 2-19. ISBN: 0256-2510.

Gillam, L (Ed.) (2002) "Terminology and Knowledge Engineering: making money in the financial services industry". Proc. of Workshop at TKE 2002, Nancy, France. (Available at: <http://www.computing.surrey.ac.uk/ai/TKE>)

Lobato I., Nankervis, J. and Savin N.E. (2001) "Testing for Autocorrelation Using a Modified Box-Pierce Q Test" International Economic Review, Vol. 42, pp 187-205.