

Trading Grid Services Within the UK e-Science Grid

Steven Newhouse, John Darlington, Miqdad Asaria, Anthony Mayer & William Lee
London e-Science Centre,
Imperial College London,
London, UK.

email: `sjn5, jd, ma299, aem3, wwhl@doc.imperial.ac.uk`

Jon MacLaren
Manchester Computing,
The University of Manchester,
Manchester, UK.

Simon Cox & Kushan Nammuni
Southampton Regional e-Science Centre,
University of Southampton,
Southampton, UK, SO17 1BJ

email: `jon.maclaren@man.ac.uk` email: `S.J.Cox, K.Nammuni@soton.ac.uk`

Katarzyna Keahey
Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, Illinois, USA
email: `keahey@mcs.anl.gov`

August 22, 2003

Abstract

The Open Grid Services Architecture (OGSA) presents the Grid community with an opportunity to define standard interfaces to enable the construction of an interoperable Grid infrastructure. The provision of this infrastructure has to date come from the donation of time and effort from the research community primarily for their own use. The growing involvement of industry and commerce in Grid activity is accelerating the need to find business models that support their participation. It is therefore essential that an economic infrastructure be incorporated into the OGSA to support economic transactions between service providers and their clients. Within this paper we describe the current status of the 'Market for Computational Services' project funded by the UK e-Science Core Programme and the efforts taking place with the Global Grid Forum to build global interoperable standards in this area.

1 Introduction

The term 'Computational Grid' is so named because of an intended analogy to electrical power grids; a vision of computational 'power' available on tap, without the user needing to really care about precisely where and how the power was 'generated'. For this vision to become a reality, it must be possible for Grid users, or consumers, to access appropriate computational power; similarly, it must be possible for the resource provided to be able to receive payment for the use of their resources.

The recent moves within the Grid commu-

nity through the Open Grid Services Architecture (OGSA) [OGSA] to standardise on a framework specification as opposed to a service implementation, has provided a generic mechanism upon which an infrastructure to enable the purchase of Grid Services may be built. Without this capability there is no economic incentive to provide service implementations (be they software, data or compute services) as there is no standard mechanism to generate revenue for their provision. Existing on-line payment schemes could be used in the implementation of such a service interface, e.g. WorldPay [WORLDPAY]

With standardised schemes to describe elec-

tronic money and to virtualise the underlying resource as services through OGSA our only outstanding requirement is to provide standardised mechanisms to describe the protocols needed to set the cost of using the service, which is currently the focus of the Grid Economic Services Architecture Working Group (GESAWG) within the Global Grid Forum (GGF) [GESAWG] of which the authors are the chairs.

We describe the motivation for Computational Economics and then describe activity taking place within the UK's e-Science programme to build such an infrastructure using the OGSA. This work is motivated by a set of use cases, one of which relating to service provisioning through either direct invocation or indirectly through a resource broker, is described in this paper. We then examine how the demands of such an infrastructure could be met by the emerging Open Grid Services Architecture (OGSA) by extending its standard Grid Services with interfaces to support economic activity, before describing the current state of the implementation activity and future research directions.

2 Economy Based Grids

The marketing of computational services for economic reward has been the subject of much research activity over the last decade as the availability and power of distributed computing resources has evolved. One example of early work in exploiting distributed computing infrastructures was Spawn which demonstrated how different funding ratios could be used to guide resource allocation and usage [Waldspurger et al., 1992]. The growth of Grid infrastructures, such as Globus [GLOBUS], Unicore [UNICORE] and Condor [Litzkow et al., 1988] has promoted further discussion as to how economic paradigms may not only be used as an approach to resource allocation but as a means to making money. For instance, Nimrod/G has shown how historical execution times and heterogeneous resource costs can be used for the deadline scheduling of multiple tasks within a fixed budget [Abramson et al., 1995].

The key to trading in the 'real world' is a medium of exchange that is deemed to have value, and goods whose value can be assessed for exchange. Bringing an economic model into Grid computing presents two opportunities: using an economic paradigm to drive effective resource utilisation, and motivating service provisioning for real economic gain by third party service providers.

3 Building the UK's Computational Marketplace

As the global grid infrastructure started to emerge and its commercial adoption started to become a reality as opposed to a dream, the lack of any economic infrastructure to motivate the provision of grid services started to become a barrier to adoption. This has been recognised within the UK's Core programme and has led to the formation of the 'Computational Markets' project [MARKETS] to develop and explore the potential of such an infrastructure within the academic and commercial Grid communities. Its participants include the regional e-Science centres in London (lead site), the North West and Southampton, a variety of commercial partners including hardware vendors, application software vendors and service providers, and end users within the engineering and physics communities. The UK's Grid Support Centre will deploy the infrastructure developed through the project throughout the UK e-Science Grid.

There are two main goals of the project: to develop an OGSA based infrastructure that supports the trading of Grid Services, and to explore a variety of economic models within this infrastructure through its deployment across a testbed between the e-Science centres involved in the project. This will involve the development and instantiation of the Chargeable Grid Services, Resource Usage Service and Grid Banking Services across a distributed testbed.

A key goal within the UK's Computational Markets project, and within the wider UK programme, is that the project's activity should contribute to building international standards within the Grid community. It is envisaged that the project will produce a reference implementation of the architecture defined through activities within various Global Grid Forum (GGF) Working Groups.

4 Motivating Use Cases

The availability of flexible charging mechanisms that are fully integrated into the Grid infrastructure presents many commercial opportunities for independent service suppliers. One of the many architectural possibilities offered by OGSA is that of service provisioning through hierarchical encapsulation of service workflow and offering this as a service to the user. The infrastructure provided by OGSA, when coupled with economic mechanisms,

offers considerable scope for new service oriented markets. These have recently been explored in a series of use cases that are being developed within the GGF's GESA-WG [GESA-WG].

4.1 Coordination between services

Consider a simple scenario of a user wishing to use a commercial third party application to analyse a self-generated data set using a computational resource.

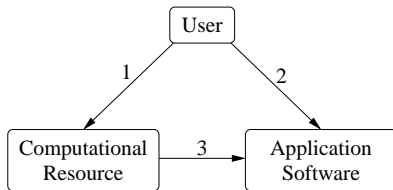


Figure 1: Coordinated use of an application software on hardware.

Setting aside for the moment the important factors that drive the selection of these services, it is necessary for the user to obtain a quotation and reservation on the *computational resource provider* (1) before approaching the *application software provider* (2) to obtain a quotation for the use of that particular software on the computational resource. Once an acceptable quotation has been found from the compute and application providers (and this

may be an iterative process as the cost of the software may dependent on the class of computational resource and the time the data may take to process) the quotations and reservations are confirmed and the computational resource may download and install the application software as required (3).

This process has already placed several demands on the Grid infrastructure from both an economic and general usage perspective. There is a need for a multiphase commitment to a resource reservation and a requirement for iterative negotiation to converge on an acceptable pricing for the resource reservation. Additional requirements such as authentication, authorisation, and impersonation (of the user by the Computational Resource provider in order to retrieve the application software) should be provided through the core middleware.

4.2 Service Aggregation

The process just undertaken by the user exposes them to the potential complexity of negotiating and reserving resources between different service providers. An alternative approach is for an organisation to provide this combined functionality direct to the user. This form of resource broker could be described as an application service provider as it provides a complete 'service' to the user - running their data through the application on an arbitrary resource.

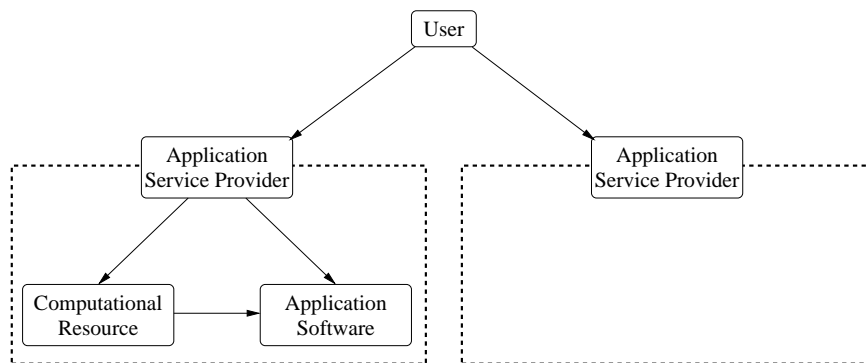


Figure 2: Service aggregation and virtualisation.

While previously the user was exposed to the full complexity of the underlying resource in this scenario the application service provider had aggregated the services to supply a 'complete pack-

age'. The application service provider has two mechanisms in providing this approach. They can provide the computational infrastructure and application software through off-line purchases of

the relevant equipment and software as would normally be expected. The service provider has full control of their costs and can offer a service directly to the user. Alternatively, they may dynamically acquire these resources in much the same way as the user did in the earlier scenario.

In this second case of service aggregation where are the economic benefits to the user? The application service provider is able to derive potential economies of scale through the bulk purchase of computer resources and software licences through the use of the economic grid infrastructure. These can be passed onto the user by reduced costs while still retaining a profit margin for the service aggregator. The service aggregator retains the flexibility to switch suppliers as long it continues to deliver any contracted service levels. From the user perspective, the service aggregator may be able to offer better pricing, faster discovery (as the a single aggregated service needs to be discovered as opposed to several compatible services having to be discovered) and faster service delivery (as software may be pre-installed).

4.3 Service Brokering

While service aggregation may offer direct benefits to the user, this is a form of service (or resource) brokering that offers a convenience function - all the required services are grouped 'under one roof'. Building on our previous examples how does a user determine which out of several application service providers should be selected for their work? The user could retain the right to select an application service provider service based on those that have been discovered from a registry service. Alternatively, this decision could be delegated to a service broker, which maintains an index of available application service providers.

The service broker is able to 'add value' to its registry of application service providers by providing extra information about their services. These differences may be as simple as cost but may include differences about the reliability, trustworthiness, the quality of the service / service level agreements and possible compensation routes to name a few. Much like a financial consultant, the broker is not expected to provide this added value service for free. Indeed, it may have a role in the financial transaction to provide an escrow account, acting as a trusted third party holding the fee until the job is complete.

5 Architectural Requirements

The previous example of application service provision does not illustrate all of the features that may be required from an economic grid services architecture. Indeed, many of the requirements from the scenario are a feature of the service oriented architecture rather than that of any economic pricing mechanism. The emergence of the Open Grid Services Architecture (OGSA) from the Grid community is providing service developers with an infrastructure that may be easily extended to support these economic services by providing a service infrastructure upon which a variety of economic models may be developed and explored.

Our purpose in this section is to sketch out the basic mechanisms required to support such an infrastructure. We will assume that economic models, dealing with issues such as price setting and Grid services market creation, will be provided by other work in this area. Our goal is to define an open infrastructure to enable the application of these pricing models to generic Grid Services.

5.1 Exploiting the Open Grid Services Architecture

The OGSA builds upon the established web services infrastructure provided through the eXtensible Markup Language (XML), the Simple Object Access Protocol (SOAP) and the Web Services Description Language (WSDL). It provides an infrastructure to securely create, manage, interact and destroy transient web service instances within distributed hosting environments that builds upon the framework used to support conventional long-lived web services [OGSA]. The Grid Service Specification defines the interface and the semantic behaviour that must be supported by the web service in order for it to be classed as a Grid Service [OGSI]. This specification is currently under development and is being standardised within the Open Grid Services Infrastructure Working Group (OGSI-WG) of the Global Grid Forum (GGF).

A Grid Service has three features that are of interest in constructing an economic framework to trade resources:

- The Grid Service Handle (GSH) provides a unique identifier to a service instance running in a service environment.
- Each Grid Service has a Service Data Elements (SDE) - an XML document - that describes the internal state of the service. The

Grid Service provides standard ports to support the update, searching etc. of the SDE by other entities.

- A Grid Service may support a 'Factory' port (or interface) that allows new service interfaces to be instantiated within the hosting environment.

A full analysis of the architectural requirements of an economic infrastructure within the context of the OGSA is being developed within the GESAWG [GESAWG].

5.1.1 Grid Service Handle (GSH)

The GSH is used by the client side code to contact the specified service or factory instance. By assuming that the economic architecture is able to embed the cost of a transient Grid service as one of the Service Data Elements of a service Factory (not an unreasonable assumption), the GSH effectively provides an identifier to a cost quotation for the use of the service. This price can of course also be advertised by other Grid advertising mechanisms; however we will assume the factory to be a reliable source of such quotations. This service price quote may vary depending on various factors, such as: time in which the service will be performed; time at which the quote is requested; identity of the requestor; the level of quality of service (QoS) factors with which the service should be performed; and the guarantee on those factors requested by the user (e.g. soft real-time versus hard real-time).

5.1.2 Service Data Elements (SDE)

The application service provider scenario has illustrated that many of the issues relating to the selection of services within an economic architecture are essentially non-functional:

- Does this service offer any bulk purchase discounts?
- Can I trust this service to deliver on its commitments?
- Is my data secure while it is residing on the remote server?
- Will I be compensated if anything goes wrong?

Such service meta-data is encapsulated within an advertising service based on the service factory data elements (SDE) structure provided by

the OGSA. This meta-data may be static (extracted from the Grid Services Description Language document that defines the service interface) or dynamic data generated by the service or inserted from other (authorised) services. Standardisation of the required and optional elements of this meta-data that is one of the challenges now facing the community.

5.1.3 An Example: Application Service Provider

We continue with our motivating example of the coordinated use of a computational resource and an application software services. The user searches a community registry for service instances that support these capabilities. The user may specify additional non-functional requirements such as a certain refund policy, or a particular architecture. The user's client contacts the factory port on each service and requests a particular level of resource use from both services (e.g. 16 processors with an interconnect greater than 100Mbs running a Solaris 2.8 operating system and a compatible version of the application software) and a minimum termination time of the reservation.

The factory generates a new service instance for each requested service use and returns these to the user. By querying the SDEs of the newly created services the user will be able to obtain the agreed price for using the service and the agreed terms and conditions. The SDE of the newly created service may differ from that of the original as the latter may support multiple approaches to setting the price of the software while the created service will only describe the agreed upon protocol. If the user is unhappy with the offered reservation, the GSH may be discarded (or retained until it expires) and the process restarted from the original service. Alternatively, the price setting protocol may allow the price to be adjusted through the newly created service, which will again generate a new GSH for further negotiation.

At some point these transient service reservations will be destroyed when their lifetime expires. If the user takes up the reservation, by invoking part of the underlying Grid Service, then the reservation will effectively be confirmed, any subsequent resource consumption will be monitored and recorded in a Resource Usage Service, and charging will then take place when the service invocation is complete through the Grid Banking Service interface.

5.2 The Grid Economic Services Architecture

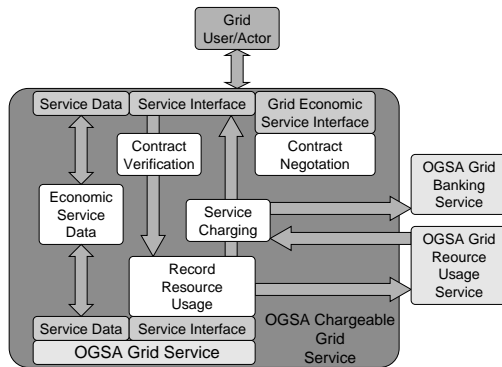


Figure 3: The current Grid Service Economic Architecture.

The constructs provided by OGSA enable a Chargeable Grid Service to be built that is able to encapsulate an existing Grid Service with the mechanisms needed to set the cost of using a service and to offer it for sale. This approach is able to exploit the basic infrastructure within OGSA for transient Grid Services while retaining a great deal of flexibility as to the eventual economic model that is used to set the cost of using the service.

Figure 3 shows a high-level representation of the internal structure within a Chargeable Grid Service. The service data elements are composed from those contained by the underlying Grid Service and the additional elements generated by the Chargeable Grid Service to describe the economic state of the service. This information is accessible through the standard Grid Service ports. Any invocation by an authorised client on the service interface is verified and passed through to the underlying service. On completion of the service invocation the resources used by the service are recorded in an external service - the Resource Usage Service. The resources consumed during the service invocation (and these might be memory, disk space, CPU time) may be charged per unit of consumed resource rather than per service invocation. The cost of using the service is passed to an external service - the Grid Banking Service - for later reconciliation.

6 Activity Within the Global Grid Fourm

We focus here on three working groups within the GGF's Scheduling and Resource Management area that are actively contributing to the definition of the economic architecture described earlier. The Grid Economic Services Architecture Working Group (GESAWG) is capturing a set of motivating use cases to inform the requirements for the underlying economic service architecture which has been defined earlier in this paper.

A key element within the overall architecture is the consumption of resources. The Resource Usage Service within GESAWG exposes the consumption of resources within an organisation by a user. Many of these resources (e.g. CPU and memory) might be used when determining the cost of having used the service. The controlled sharing of resource usage information that has been captured by the underlying service infrastructure is becoming an increasing priority with virtual organisations around the world. A service interface is being defined by the Resource Usage Service Working Group (RUS-WG) [RUS-WG] that will enable the secure uploading of consumed resource information and the extraction from the service by authorised clients.

An assumption with the RUS activity is a standard mechanism to interchange data between different Grid entities. The resource information (its values, quantities and structure) that may need to be exchanged between different centres is being defined within the Usage Records Working Group (UR-WG) [UR-WG]. Several possible interchange formats are envisaged for this information including XML.

7 Implementation

The implementation activities within the Computational Markets are ongoing. For the All Hands Meeting in September 2003 we will be able to demonstrate the simple pricing and purchase of a economically enabled Counter Grid Service, where the service invocations are recorded in a Resource Usage Service and the transfer of money is recorded in the Grid Banking Service, across distributed resources. The development and implementation of these three services have been distributed between the centres and their staff:

- Chargeable Grid Service (LeSC): Miqdad Asaria, William Lee and Anthony Mayer

- Resource Usage Service (eSNW): Jon MacLaren
- Grid Banking Service (SeSC): Kushan Namuni

These services are described in prototype form in the specification documents available from the GESA and RUS working groups within the GGF.

8 Future Plans

The last few years have seen the early adoption of Grid infrastructures within the academic and business community. While the use of Grid mechanisms is not yet widespread, the recent emergence of the Open Grid Services Architecture and its adoption of Web Services as its service infrastructure will certainly accelerate their uptake. Future Grid environments may therefore comprise of thousands of Grid Services exposing applications, software libraries, compute resources, disk storage, network links, instruments and visualisation devices for use by their communities. While this vision is appealing, a pool of Grid Services available for general use, it is not realistic, as such a service infrastructure would have to be paid for by its users.

Within such an environment we foresee the emergence of resource brokers that are capable of 'adding value' to the basic service infrastructure by finding and annotating services with information relating to their capability and trustworthiness. Users are able to obtain their required services from these brokers, who may offer a guarantee as to their capability, or alternatively seek out and discover their own services. These services need not be provided for free and for widespread acceptance of the Grid paradigm organisations must have a mechanism for defining and connecting revenue from service provision.

The key to such an infrastructure is interoperability. The work being done within the Global Grid Forum and areas of the wider Web Service community is defining standards is enabling interoperability. The 'Markets for Computational Services' project is an early adopter of the OGSA and has developed the Grid Economic Services Architecture and Resource Usage Service (RUS) working groups within the GGF. The RUS-WG will define a key lower-level service within the OGSA while activities within the GESA-WG will define key higher-level service interfaces. We are able to benefit from work that has already taken place within the Usage Record Working Group

and expect to contribute to the emerging OGSI-Agreement specification being defined within the GRAAP Working Group [GRAAP-WG].

The internet has brought us ubiquitous access to data and simple services for little or no cost. The Grid offers the possibility of ubiquitous access to more complex services but their appearance will be predicated on the service provider receiving an income for its provision. The proposed economic architecture is in its early stages of development, but will build upon OGSA to be open and extensible across many deployment scenarios and economic models and thereby providing an infrastructure that will enable utility computing. Within this architecture we can see the speculative purchase of resources by services for later resale (a futures market), customer dependent pricing policies (grid miles) and other mechanisms to encourage the maximum utilisation of resources by maximising revenue generation.

Acknowledgements

This work is being supported in part by the Computational Markets project funded under the UK e-Science Core Programme by the Department of Trade and Industry and the Engineering and Physical Sciences Research Council [MARKETS], and by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under contract W-31-109-Eng-38.

References

- [Abramson et al., 1995] Abramson, D., Sobic, R., Giddy, J., and Hall, B. (1995). Nimrod: A tool for performing parametrised simulations using distributed workstations. In *4th IEEE Intl. Symp. on High-Perf. Dist. Computing (HPDC)*.
- [Litzkow et al., 1988] Litzkow, M. J., Livny, M., and Mutka, M. W. (1988). Condor - a hunter of idle workstations. In *Proceedings of the 8th International Conference on Distributed Computing Systems*.
- [Waldspurger et al., 1992] Waldspurger, C. A., Hogg, T., Huberman, B., Kephart, J. O., and Stornetta, W. S. (1992). Spawn: a distributed computational economy. *IEEE Transactions on Software Engineering*, 18(2).

- [GESA-WG] Global Grid Forum Grid Economic Services Architecture Working Group (GESA-WG). <https://forge.gridforum.org/projects/gesa-wg/>
- [GLOBUS] The Globus Project. <http://www.globus.org>
- [GRAAP-WG] Global Grid Forum Grid Resource Allocation Agreement Protocol Working Group (GRAAP-WG). <https://forge.gridforum.org/projects/graap-wg/>
- [MARKETS] Computational markets project website. <http://www.lesc.ic.ac.uk/markets>
- [OGSA] The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Available from <https://forge.gridforum.org/projects/ogsa-wg/>
- [OGSI] Open Grid Services Infrastructure v1.0 Specification (Proposed Recommendation). Available from <https://forge.gridforum.org/projects/ogsi-wg/>
- [RUS-WG] Global Grid Forum Resource Usage Service Working Group (RUS-WG). <https://forge.gridforum.org/projects/rus-wg/>
- [UNICORE] Unicore Grid Middleware. <http://www.unicore.org/>
- [UR-WG] Global Grid Forum Usage Record Working Group (UR-WG). <https://forge.gridforum.org/projects/ur-wg/>
- [WORLDPAY] Worldpay UK website. <http://www.worldpay.co.uk/>