Economic Modeling of Grid Services

Jörn Altmann[1], Stavros Routzounis[2]
[1]*School of Information Technology, International University, Bruchsal, Germany*
Tel: +497251700130, Fax: +15016415384, Email: jorn.altmann@acm.org
[2]*Department of Informatics, Athens University of Economics and Business, Greece*
Email: sroutz@aueb.gr

Abstract: The open and global nature of the Internet provides the possibility of developing and using many new services (such as Grid services) to any user. These users can be the general public, academic institutions, SMEs, or larger organisations. Grid services could range from a simple service charged at low rate but aiming for mass usage to complex services encapsulating e-Science or HPC methods. Similarly, purchasers of such services could be from the general public, academia, or industry as well. Importantly, however, all services would be generally available; there would be no vertical segmentation of the service market according to user or service type. The ability to utilise remote Grid Utility Computing platforms frees both the service provider and the user from the need to own or acquire the necessary computational infrastructure. This paper describes the approach of GridEcon, a European Union funded project on Grid computing, that will build market-based, service-oriented Grid technology. Currently, since Grid computing suffers from not being used for commercial purpose at large scale, it reduces the benefit of Grid Computing. GridEcon offers a set of solutions to this shortcoming that are based on economic principles. This type of economics-aware Grid will create additional value for its participants by fully exploiting the economic value of the resources and providing the basis for developing creative, new business models of the next-generation Internet.

Keywords: Grid Economics, Grid Business Models, GridEcon Project, Modelling, Market-Based and Service-Oriented Architecture, Next Generation Internet

1. Introduction

Recent developments in service technologies (e.g. Web services) and Grid technologies hold the promise that the next-generation Internet can develop to be a rich economic or commercial arena where a wide variety of computational services and utilities are bought and sold. Therefore, the Grid has the potential to be more than just a large distributed system that provides virtualized services to its users and enables computing on demand on a best-effort basis at zero cost (free resources). Instead, an economics-aware Grid (and, in the future, the next-generation Internet) shall offer services that span many economically independent enterprises (in contrast to enterprise Grids or cluster Grids). In this environment, enterprises can make independent decisions, forming a real economy. The Grid is expected to evolve into an infrastructure that allows resolving differences in service preferences and utility of participants through interaction in markets. This infrastructure will create additional value for its participants since they will be able to better express their preferences, share resources more effectively, and generate revenue. In order to leverage the opportunities of this next-generation Internet, additional services and components that enhance the Grid infrastructure have to be identified and realised. These additional services will realise these market opportunities as fully and as efficiently as possible. These services are located within the services framework and within existing Grid technologies.

The development of Grid middleware and cluster management systems (Globus [1], Sun Grid Engine [7] and Condor [8]) provides the means whereby physical computational and storage resources can be shared and made available for use by communities beyond their immediate owners. Computational resources can be realised through the development of Grid Utility Computing platforms. Utility Computing is the provision, by third parties, of remote execution or storage environments, often realised by large localised installations of servers or clusters, to support the computational requirements of customers on a use-on-demand, pay-per-use basis. Utility Computing dramatically lowers the cost of ownership (e.g. for HPC users) and has many other benefits accruing from specialisation and economies of scale. One important recent development in Utility Computing is Sun's N1 Grid Engine Pay-per-Cycles facility. Here, Grid computing technology is used to provide Internet-available execution and storage platforms that can be acquired and used on a pay-per-use basis.

The development of transparent application mapping and job submission systems (e.g. GridSAM [11]) means that such execution platforms become interchangeable or substitutable as applications can be mapped or adapted to alternative platforms. This in turn makes execution itself a tradable good (or commodity), opening up the potential for markets in processor cycles or storage. However, Grid middleware systems, to date, neither provide support for concurrent execution of applications nor support for applications to run economically efficient [16][17].

A further, important development has been the convergence of the Grid computing with the WWW and business communities, and the adoption of Web Services as the application construction and deployment framework. Web Services provide a framework where potentially complex procedures can be deployed as easy-to-use (and easy-to-combine) services on the web. Importantly, this means that the Grid world can coincide with the public Web and Business Enterprise worlds and any results emanating from Grid research can now find ready application and adoption on the public Internet. The prospect of an open public Internet market is exciting. In such an environment, each member of the general public, academic institutions, SMEs or large organisations can purchase, develop, and consume services. If the same level of ingenuity and public enterprise that has made the current (information) Web such a phenomenon could be harnessed to the production of innovative services on the public Web, the potential for exciting and enterprising activity would be considerable. Such services could range from the simple (e.g. e-mail filters, photo processing and archiving, or audio mixing), charged at very low rates but aiming for mass usage, to the complex, perhaps, encapsulating e-Science or HPC methods. Importantly, however, all services would be generally available; there would be no vertical segmentation of the service market according to user or service type. The ability to utilise remote Grid Utility computing platforms frees both the service provider and the user from the need to own or acquire the necessary computational infrastructure. This environment, which is called the service-oriented, market-based Grid, will be developed within the GridEcon project.

The remainder of the paper is organized as follows. The next section describes the state-of-the-art in Grid research. Section 3 points out briefly the objectives of the GridEcon project. The architecture of the service-oriented, market-based Grid as envisioned within the GridEcon project is illustrated in Section 4. This section also determines the Grid technology components that have to be designed or adapted in order to successfully implement the next-generation Grid.

2. State-of-the-Art in Grid Research

So far, Grid economics research has focused mainly on simple resource definitions (CPU cycles) and price mechanisms (mainly auctions). These ideas, which fall into the general category of market mechanisms for contention resolution in IT systems, fall short in capturing the issues such as risk, uncertainty, principle shift in business models, effect of disruptive technology, and trust. Those issues distinguish Grid from traditional distributed systems.

The Globus Toolkit project [1], the NextGrid project [2], the SweGrid project [19], the EGEE project [18], the GridBus [20], and the Akogrimo project [3] have all developed basic Grid components and architectural concepts that allow building economic-aware components. Projects as GRASP [9], Akogrimo, GRIA [4], TrustCom [10], and NextGrid started to consider specific aspects of risk, trust, and service-level agreements, such as SLA aggregation and risk assessment in Grid computing. However, all approaches differ from the GridEcon approach in that they do not provide a generalize framework for all issues on Grid economics.

The NextGrid Project develops a next-generation Grid architecture, which addresses the issues that exist in today's Grid, namely missing capabilities to express the needs of business users, to ensure security, to consider business models, legal requirements, and privacy.

Akogrimo aims at an advancement of pervasiveness of Grid computing [21]. Besides considering layers and technologies of the next-generation Grids, which focus on knowledge-related and semantics-driven web services, Akogrimo architects a next-generation Grid, which closely co-operates with evolving mobile Internet infrastructures based on IPv6. Because of the mobility, the components of the architecture of Akogrimo have to consider a continuously changing demand for resources. If the resources are not allocated in an economically efficient way, service quality cannot be guaranteed for business-critical applications.

GRIA is a Grid middleware that allows users to trade computational resources on a commercial basis [4]. It allows resource owners and users to negotiate terms for access to resources. This work, which is based on existing business models and business processes, provides components for trust management, security, and authentication in a service-oriented architecture. The Grid middleware allows modelling and implementation of new Grid market services, new components, and extensions have to be designed on top of this existing middleware.

BEinGrid project aims to exploit European Grid middleware by creating a toolset repository of Grid services from across the Grid research domain and to use these services to deliver a set of successful business experiments that stimulate the early adoption of Grid technologies [5]. It is focusing on applying existing theory and developing a large number of business models for Grid applications.

The Sorma project [6] is also addressing issues related to Grid economics and markets. The difference is that Sorma is going bottom up, starting by defining market mechanisms for trading the basic CPU and storage resources and then by designing the appropriate middleware components for brokering, accounting, and charging.

GRASP analyses the issues of composition of Grid Services and licensing. It investigated innovative business models for Application Service Providers (ASP), based on collaborative Virtual Organizations.

3. Objectives

GridEcon analysis is top down: it starts from scenarios, defines the appropriate business models, and as a result will first identify the kind of higher-level goods that applications would like to obtain in a commercial Grid environment. These goods are not expected to be direct CPU and storage mappings but contracts for more complex services expressing quality, reliability, time constraints, and risk aversion. Those services are expressed in the methods of payment (flat fee, usage based, pay as you go, etc.). Many of these higher-level services may be offered by new stakeholders. These stakeholders will be the new Grid intermediaries and will act as aggregators and insurance providers (risk brokers). We place strong emphasis on the higher-level market concept and its relation to the lower level market, which we consider very important for the commercial adoption of Grid.

We believe that the understanding of Grid economics will allow users to fully exploit the potential value of the Grid and service technologies and the resulting externalities of large user participation [12][13]. For example, the new Grid Utility Computing paradigm greatly impacts the management of resource needs within a company and the organization of its IT infrastructure due to the business model shift from “buy a computer” to “computing on demand”. Moreover, future “Grid-friendly” applications should be capable of dealing with the special features of the Grid environment in terms of uncertainty and dynamic resource provisioning. This may imply a move away from the traditional business model, where applications assume hard service level agreements (SLAs) for servers (which are easy to implement due to ownership) and networks, and towards a new dynamic resource environment, in which applications have to be price-aware and able to deal with a certain degree of uncertainty [14][15].

4. The Model of an Economic-Aware Grid

4.1 Architecture

In such a service-oriented, market-based Grid, services can be highly customized and composed from other services. As depicted in Figure 1, services can range from buying software execution time (i.e. software will not be ‘sold’ anymore) to enabling the composition of services, to mapping services onto physical resources, or to brokering physical resources (i.e. providing virtual computing services). The development of a properly functioning Internet market for services will be facilitated by and will make heavy demands on Utility Computing platforms, for the transparent and flexible delivery of computational and storage facilities. The development of higher level open markets for services on the Internet is an exciting possibility that could make the developments so far in Grid middleware much more applicable in public and commercial arenas, not just in scientific computing. However, in order to make these Grid-enabled platforms capable of supporting these higher-level, service-based commercial applications, the Grid middleware systems themselves have to be endowed with sufficient market awareness to allow them to support these higher level markets efficiently and effectively.

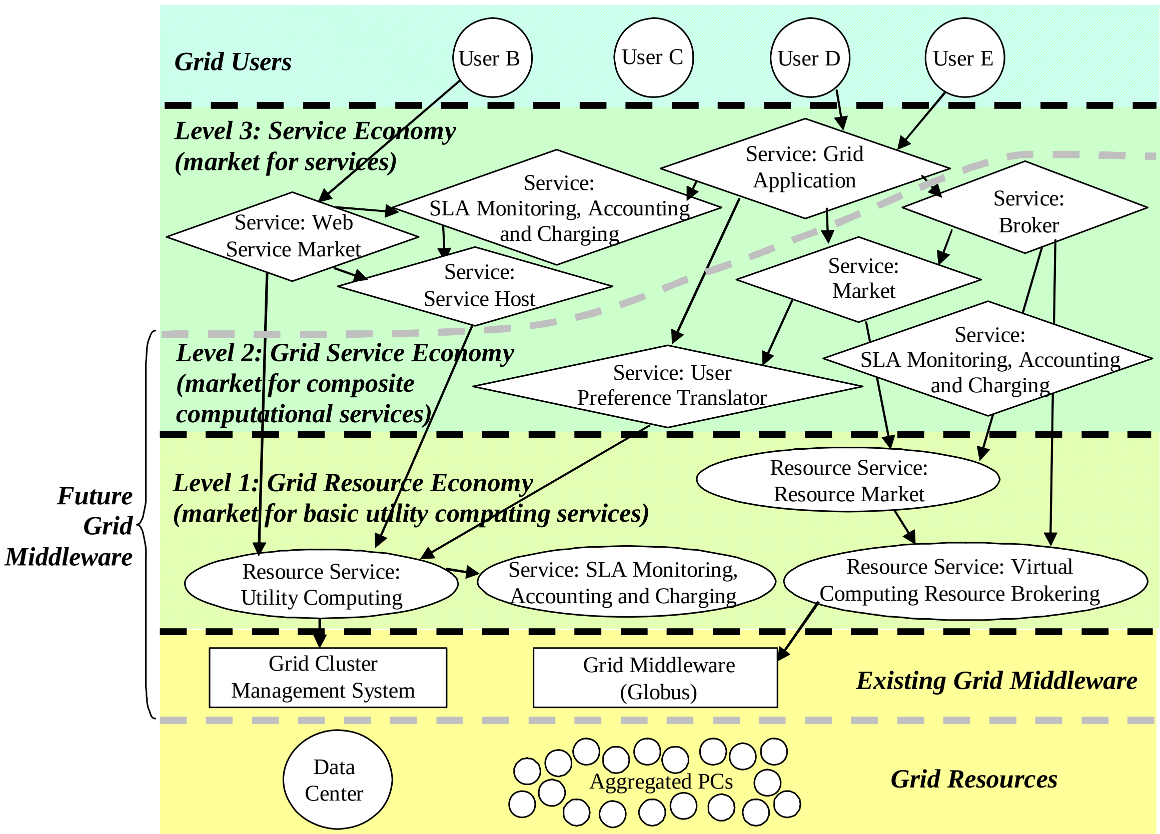

Figure 1: The service-oriented, market-based Grid

The available computing resources will be huge data centres (even possibly spanning many organizations) or co-ordinated collection of PCs owned by different individuals. The role of the Grid middleware shifts to providing dynamically configurable on-demand utility services where the highly dynamic aspect of the underlying resources poses new requirements on pricing models, resource scheduling, control, and quality assurance. In the envisioned service-oriented, market-based Grid architecture, the users, via a service (i.e. application), can express their preferences to the Grid by submitting a request for resources to resource service providers directly (*Resource Service: Utility Computing* in Figure 1) or to a high-level market (*Service: Market* in Figure 1), where Grid resources are traded. If the trade is executed, the different Market-Managed Resource Allocation Systems (Level 1 in Figure 1) of the Grid middleware allocate the resources to the user, which might belong to independent organizations. These mechanisms are responsible to determine the priority among the different service requests and schedule the resource requests accordingly. They are based on the market value of the request. The *SLA Monitoring, Accounting and Charging System* (Figure 1) records the service usage by a user, checks it against the SLA, and calculates the total charge.

Figure 1 also shows the layering of these service markets, which allow users to buy and sell services. Each market may have similar components (for accounting, SLA management, brokerage, directories, etc.) but its specific type of goods sold may require different types of market mechanisms and interaction with different intermediaries. For instance, the Grid service layer (offering composite computation services) offers the capability of buying reliability, quality for storage, and computation bundles. Those services are offered by intermediaries, which may aggregate simple utility contracts from the basic computational services market.

The fixtures in the computational world are manufacturers who construct hardware systems and users who want to use these resources to accomplish some task. In the reconstructed model, hardware manufacturers sell their products to Grid Utility Computing operatives (or they may operate them by themselves). These Utility Computing service providers sell computing cycles, bandwidth, or storage on an open market (see Level 1 in Figure 1).

These facilities can then be utilised, for instance, by service hosts to provide an environment where new services can be both authored and executed (see Level 3 in Figure 1). This in turn simplifies the production of new services (i.e. Grid applications) by third parties (individuals, academia, SMEs or large organisations). All services can be used by end-users.

In such open markets, there is a clear need, and hence a market opportunity, for new kinds of services (e.g. risk brokers, SLA creators, resource aggregators) to act as intermediaries between users and services (which is advancement to the state-of-the-art). Such brokers could combine their user-facing service discovery or recommendation operations with back-end services (e.g. deploying such selected services on appropriate utility platforms).

In such a flexible and dynamic environment, a significant number of economic-related problems will occur, ranging from incentives for participation and sharing of information, economic efficiency in resource allocation, information structures for making decisions and evaluating risk, to the capability to express preferences through charging ("pay more and get more"). Therefore, unless the new generation of services and Grid technology provides the means for users to operate in this economics-aware environment and to resolve those economics problems, its full potential and its commercial adoption will be in jeopardy. However, if such an architecture could be realised, it would provide exciting commercial opportunities. Each service is self-contained and the interfaces between them correspond to economic or commercial transactions. Thus, each layer provides market opportunities. As this market would operate over the global Internet, there is no restriction as to who could be a producer or a user of such services. That means that the current divisions of applications (services) between the general public, academia, or business would be eroded. Anyone could be a producer or user (or both) of such services, consuming Grid services on demand.

4.2 Required New Grid Market Services

In order to discover the most challenging economic-related problems in such an environment, an understanding of potential business models is necessary. The analysis of business models reveals roles that companies can take on and helps to discover the missing components in this market-based, service-oriented Grid. We have studied several scenarios in all sectors (the general public area, the industrial sector, and the academia area), since we are targeted at the service-oriented, market-based Grid. In the following, in order to illustrate this vision, we are laying out a scenario from the industry sector.

The scenario is set around a large, multinational financial services group with a range of activities covering banking, brokerage, and insurance. This enterprise believes that it can best deliver value to shareholders by operating as a series of autonomous business units (BUs), which can meet the specific needs of individual markets and geographies. It has identified 4000 individual BUs. Given that many of these units are very small, the organization believes that they will have the required element of entrepreneurship to drive growth.

The two key challenges for the organization are to give each BU the freedom to follow their entrepreneurial ideas and, at the same time, to provide them with the resources and the benefits of being part of a larger organization. A larger organisation can acquire resources at low costs. IT is considered such a resource and central to the performance of each BU, fulfilling its business needs. Therefore, the enterprise wants to fully understand the IT needs of each BU, wants to know how it can best provide IT so as to meet these individual business needs in the most cost-effective way, and how the IT costs should be allocated to each BU, so that it can realistically assess profit and loss at the BU level.

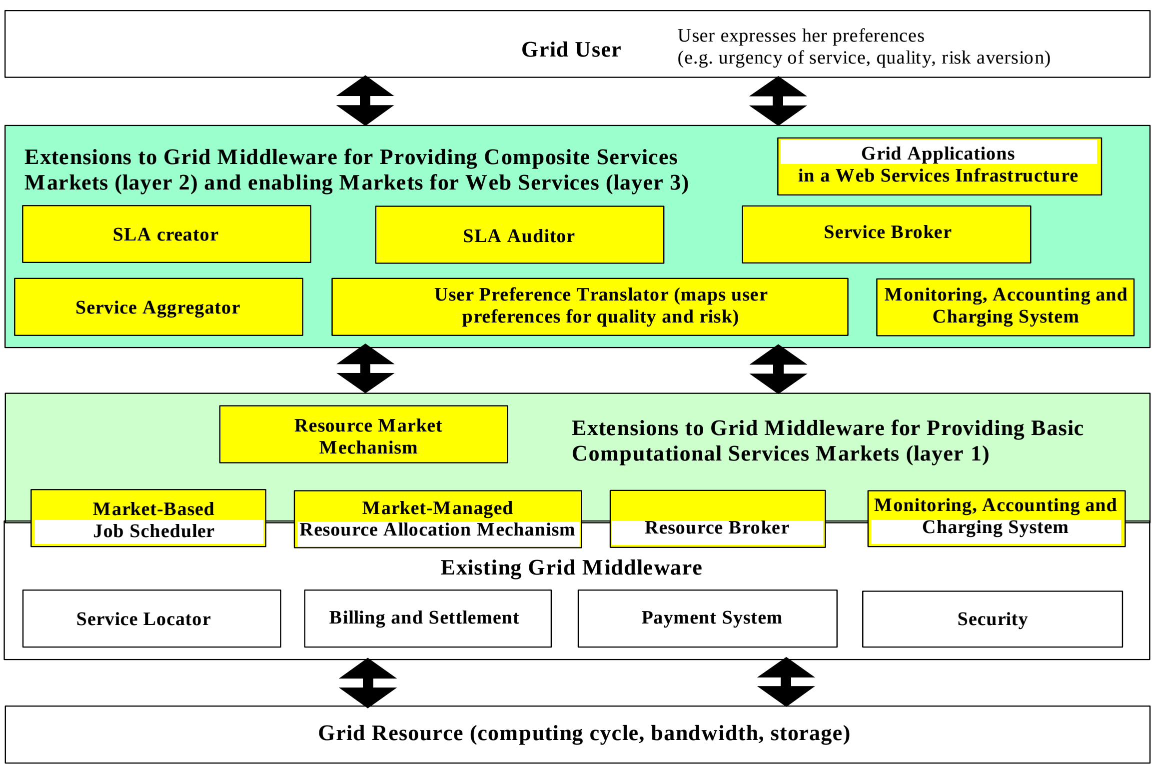

Figure 2: Grid services and components that are available (white), that have to be provided (grey), or that have to be modified (white/grey)

In order to drive business (e.g. lower the IT cost, harmonize IT usage and requirements across different BUs) and deal with business challenges (e.g. rivalries between departments, concerns from BUs that the cost will rise), the enterprise's aim is to develop an enterprise Grid with an accounting and charging service such that each BU pays for its IT usage (processing power, storage, bandwidth and software licenses) based on different economic parameter. For the enterprise Grid, it will use a mixture of new commodity hardware and re-provisioned existing hardware. The organization will choose a Grid middleware with a strong regard for standards. Given the challenges of migrating existing software, the focus will be to run applications on the Grid.

This scenario is a nice example for what the state-of-the-art in Grid computing cannot provide right now. Therefore, the successful implementation of this scenario requires additional services on top of the existing Grid middleware. Those services, which are parts of the three market layers of the service-oriented Grid (see Figure 1), are depicted in more detail in Figure 2.

These additional, economics-related services, which can be taken on by different stakeholders within the scenario and are illustrated in Figure 2, are:

- Broker capabilities for purchasing application, computation, and storage services that take into account the utility of quality of service from the side of the buyer BU (i.e. the willingness to pay for reliability, availability, and security).
- Service aggregator capabilities for offering large utility services by aggregating simple utility services. The simple utility services are bought internally and are possibly combined with external ones. The available Grid middleware offers virtualized resources with low availability or reliability guarantees, but a BU may be reluctant to rely on such services. A service aggregator may also be able to sell advanced reservations by predicting the future state of the system.
- Mapping services that translate preferences for information services into combinations of resources (server, storage, bandwidth) by also taking into account requirements of the resources.
- Resource reservation services, which allow BUs to advertise resource availability, also allow securing computing resources in advance. The types and attributes of the reservations should be inline with the actual requirements of the BUs.
- SLA creators, negotiators, and monitors deal with SLAs that include the above mentioned quality aspects of services. BUs may need such SLAs internally since they cannot trust other BUs making their resources available as promised. There should be incentives for BUs to honour their SLAs.
- Service hosts so that any BU can easily offer new services (e.g. new software application or new hardware).
- Monitoring services, accounting services, and charging services for settling any service transactions.
- Business intelligence services for evaluating the efficient operation of the system and for updating its infrastructure. For example, a service that provides each BU with an overview of prices for IT services over the course of a month. Based on this information, each BU could adapt its IT service purchasing behaviour to lower the cost.

4.3 Research Questions

This analysis of services offered by basic stakeholders or possible combination of stakeholders, and particularly the interaction of stakeholders, is of great interest. Since these stakeholders are different business entities, which might be in competition with each other, each composition of these services opens the possibility for market games between the stakeholders. These games have not been investigated yet. Within the scenario above, those games could be played between all business units. One business unit might be capable to manipulate the market such as it will provide the lowest possible cost for itself but increasing the cost for all the other business units (i.e. decrease in social welfare).

Another challenging research question is how to enable end-users to express their preferences and needs for (Grid application) services. The kind of preferences that end-users (or BUs) might have can be based on:

- The quality of service (e.g. reliability of the service),
- The end-user's risk aversion related to fluctuations (unpredictability) of prices and quality (availability and reliability),
- The security of the service,
- The SLA liability and enforcement,
- The time of service consumption and the ability of resource reservation.

Mechanisms must be capable to let users express their preference and to translate these preferences into actual demand for Grid resource, considering the preferences of all the other market participants. In particular, the research questions to be addressed are how we can design a mechanism that allows users to express their need for reliable services or a mechanism that allows users to express their risk aversion of not receiving any service.

After designing such mechanisms, they will have to be implemented and integrated with existing Grid middleware, in a way that the access to these services shall be feasible by existing Grid middleware (e.g. EGEE, SweGrid SGAS, Globus Toolkit OSGA GT4). This way, this new technology shall be used by a large number of Grid developers. In addition to this, the implementation of new market mechanisms would most probably require the modification of existing Grid technology. In particular, the job scheduler and resource allocation algorithm might have to be adapted.

5. Conclusions

We have presented the model of a service-oriented, market-based Grid. In this environment, we pointed out the most challenging economic-related problems, which seem to prevent users from fully exploiting the potential value of the Grid. We also suggested that the analysis of potential business models can help discovering missing Grid services and components (on top of the existing Grid middleware), that would resolve the major economic-related problems and would support the implementation of new Grid business models.

References

[1] Globus Toolkit OSGA, http://www.globus.org/ .
[2] NextGrid project, http://www.nextgrid.org/ .
[3] Akogrimo project, http://www.mobilegrids.org/ .
[4] GRIA project, http://www.gria.org/ .
[5] BEinGrid project, http://www.beingrid.com/ .
[6] Sorma project, http://www.sorma-project.org/ .
[7] Sun Grid Engine project, http://gridengine.sunsource.net/ .
[8] Condor project, http://www.cs.wisc.edu/condor/ .

[9] GRASP Project, http://eu-grasp.net/english/default.htm, January 2005.
[10] TrustCom, http://www.eu-trustcom.com/, July 2005.
[11] W. Lee, A.S. McGough, and J. Darlington, "Performance Evaluation of the GridSAM Job Submission and Monitoring System," UK e-Science All Hands Meeting, Nottingham, UK, Sept 2005.
[12] Junseok Hwang, "Grid and P2P Economics and Market Models," GECON 2005, 1st Workshop on Grid Economics and Business Models, Seoul, South Korea, April 2004.
[13] Steven Miller, "Analyzing Demand Patterns, Business Models and Governance to facilitate the growth of Grid Services in Singapore," GECON 2005, 2nd Workshop on Grid Economics and Business Models, Seoul, South Korea, March 2005.
[14] Kenyon Chris, "Creating Services with Hard Guarantees from Cycle-Harvesting Systems," IBM Research Report RZ 3461 11/11/2002. CCGrid 2003, Tokyo, May 2003.
[15] Kenyon Chris, "Grid Resource Commercialization: Economic Engineering and Delivery Scenarios," in: Grid Resource Management: State of the Art and Research Issues. Editors: J. Nabrzyski, J. Schopf and J. Weglarz, Kluwer, 2004.
[16] Klaus Krauter, Rajkumar Buyya, and Muthucumaru Maheswaran, "A Taxonomy and Survey of Grid Resource Management Systems for Distributed Computing," International Journal of Software: Practice and Experience (SPE), Wiley Press, New York, USA, May 2002.
[17] Rajkumar Buyya, David Abramson, Jonathan Giddy, "Nimrod/G: An Architecture for a Resource Management and Scheduling System in a Global Computational Grid," The 4th International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia 2000), Beijing, China. IEEE Computer Society Press, USA, May 2000.
[18] EGEE, "EGEE Middleware Architecture," available at: https://edms.cern.ch/file/594698/1.0/architecture.pdf , 2005.
[19] SweGrid, "SGAS: A Coordinated Accounting Solution for SweGrid," available at http://www.sgas.se/docs/SGAS-0.1.3.pdf , 2003.
[20] Alexander Barmouta and Rajkumar Buyya, "GridBank: A Grid Accounting Services Architecture (GASA) for Distributed Systems Sharing and Integration," Workshop on Internet Computing and E-Commerce, Proceedings of the 17th Annual International Parallel and Distributed Processing Symposium (IPDPS 2003), IEEE Computer Society Press, USA, Nice, France, April 2003.
[21] Wesner, S., Jähnert, J. M., Aránzazu, M., Escudero, T., "Mobile Collaborative Business Grids: A short overview of the Akogrimo Project," Available at http://www.akogrimo.org/download/White_Papers_and_Publications/Akogrimo_WhitePaper_Overview.pdf , 2005.