

Capacity Planning in Economic Grid Markets

Marcel Risch¹, Jörn Altmann²

¹International University in Germany, School of Information Technology
Campus 3, 76646 Bruchsal, Germany
marcel.risch@i-u.de

²TEMEP, School of Industrial and Management Engineering
College of Engineering, Seoul National University,
San 56-1, Sillim-Dong, Gwanak-Gu, Seoul, 151-742, South-Korea
jorn.altmann@acm.org

Abstract. Due to the few computing resource planning options currently available in Grid computing, capacity planning, an old discipline for analyzing resource purchases, is simple to perform. However, once a commercial computing Grid is established, which provides many different resource types at variable prices, capacity planning will become more complex and the user will require support for handling this difficult process. The support could come from an online Grid Capacity Planning Service, which helps users with little IT expertise to make use of the Grid in a cost-effective manner. This Grid Capacity Planning Service is a stand-alone service, enabling companies to outsource their capacity planning task. This paper describes the Grid Capacity Planning Service and demonstrates the workings of the service through simulations.

Keywords: Grid Economics, Grid Capacity Planning, Service-Oriented Computing, Grid Computing, Resource Allocation, Utility Computing.

1 Introduction

Capacity planning is being applied in many variations in companies. The more the company depends on capacity planning decisions, the more effort is allocated to it. For example, in data centers, capacity planning is extensively used to determine the computing resource needs. To ensure that all applications run with the required QoS and none of the computing resources becomes overloaded, the IT staff continuously monitors system data and resource usage, forecasts the future demand of applications and, thus, predicts their resource requirements. This requirements list can then be turned into an allocation plan and, if the existing resources are insufficient, into a list of required resources which need to be purchased.

At present, the computing capacity planning process of companies is fairly simple, since required computing resources can only be purchased or leased. With the advent of commercial Grids, however, capacity planning becomes more involved. Any company that requires additional resources now is offered a new option for satisfying its computing resource needs: purchasing Grid resources from the commercial Grid. This additional option adds additional complexity to the resource purchase decision mak-

ing process, since three issues have to be addressed: Firstly, it needs to be decided which applications are suitable to run on the Grid. Secondly, since a Grid market is expected to be competitive, prices will fluctuate with changes in supply and demand. Thus, if the overall cost of Grid usage has to be determined, the price for Grid resources has to be predicted accurately. Finally, the demand fluctuations have to be predicted accurately, since the benefit from using the Grid comes from selling spare capacity on the Grid and buying additional resources at times of peak demand times. From these three issues, it can be seen that, while capacity planning is vital to using computing resources in an economically efficient manner, it is extremely difficult to perform it properly.

Because of this difficulty, we propose a new service in the commercial Grid environment: the Grid Capacity Planning Service (GCPS). The remainder of the paper is organized as follows. Section 2 gives an introduction to capacity planning, while section 3 elaborates on the difference between traditional capacity planning and Grid capacity planning. Our capacity planning model is introduced in section 4 and expanded in section 5. The workings of the model are then demonstrated with the use of simulations in section 6.

2 An Introduction to Capacity Planning

The term “capacity planning” is often used but rarely defined. To avoid any ambiguity, we follow the definition given by IBM [1]: “*Capacity Planning encompasses the process of planning for adequate IT resources required to fulfill current and future resource requirements so that the customer's workload requirements are met and the service provider's costs are recovered.*”

This definition allows us to categorize the users of capacity planning into two groups: customers and providers. Current research has largely taken the stance that the provider’s problem in Grids is a resource allocation problem to which economic mechanisms can be applied [2, 3, 4, 5, 6, 7]. Other researchers have taken a more long-term view of capacity planning which works with reservations [8] while still others have applied the problem to specialized fields, such as phased workloads [9].

However, there is, as of yet, no research being done on the customer’s need for capacity planning in a utility computing environment. We will remedy this situation by focusing solely on the customer’s capacity planning problem, which is at least as challenging as the provider’s.

The following three tasks are at the heart of the capacity planning process, according to the definition given above: (1) monitoring of the current resource utilization, (2) estimation of future resource requirements of applications and, finally, (3) cost estimation to ensure that a company does not overspend.

2.1 Capacity Planning Tasks

Before the introduction of commercial Grids, capacity planning had only been a long-term approach. Data center staff had to analyze the current application-to-resource mapping, the monitoring data, and some economic data, such as the income

generated by certain applications. Using this input, the data center staff then had to determine whether the resource pool is able to run all applications at the required QoS. If this had not been the case, the data center staff had to determine which additional resources have to be purchased or leased and then create a migration plan for the applications that have to be migrated.

With the advent of commercial Grid offers, capacity planning can also be used to solve short-term capacity problems. In this case, the data center staff can purchase additional Grid resources if the applications no longer run at the required QoS. Since this decision can be implemented within minutes, the capacity planning process now takes on a short-term aspect as well. However, as has been shown in [10], using the Grid excessively is also not to be encouraged, as the Grid becomes more expensive than in-house resources in the long run.

We can therefore say that the capacity planning process for utility computing consists of two parts: The short-term capacity planning process and the long-term capacity planning process which has been used in datacenters before the introduction of utility computing environments. This idea is illustrated in the figure below.

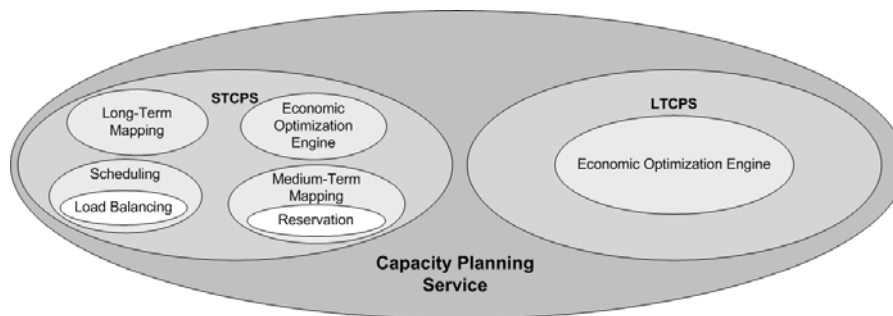


Fig. 1. Capacity Planning Structure for Utility Computing.

Each of the two capacity planning sub-services, the Short-Term Capacity Planning Service (STCPS) and the Long-Term Capacity Planning Service (LTCPS), has a number of tasks to perform. For the STCPS, the main task is the scheduling of applications to resources that are available. A further subtask of scheduling is load balancing which ensures that all resources are used evenly. Furthermore, the STCPS has the capability to perform medium-term mapping of applications to resources (i.e. resource reservation). This is useful in the case of daily demand peaks which can then be planned for. Such a module reserves utility computing resources (e.g. on a day-ahead basis) to ensure that the scheduler has sufficient resources to schedule all applications. Lastly, the STCPS also has to take the economics of Grid usage into account. Not only can the Grid usage become expensive over time, it can also be more expensive than letting an application run slower. To determine whether using the Grid is economically efficient, the STCPS has to have an economic optimization module.

The LTCPS, on the other hand, is mostly concerned with the economics of resource purchases. In other words, its focus lies on the question of which resource purchase is the most economically efficient one. This procedure has to take into account the current mapping of application to resources, the performance of each application and the costs incurred by using Grid resources. Furthermore, the user's budget con-

straints and economic requirements (e.g. importance of applications to the user's business, the expected long-term benefits of providing good QoS) have to be considered when developing a new application mapping.

Especially, the LTCPS also has to consider risks. Some risks are inherent to the system, such as resource failures or provisioning issues. To avoid these problems, the Economic Optimization Engine has to take into account the risks inherent to using in-house and Grid resources and has to determine which course of action (e.g. fault tolerance mechanisms) can minimize these risks.

2.2 Capacity Planning Inputs and Outputs

To perform their tasks, both capacity planning services require a number of inputs which are shown together with the outputs of the capacity planners in the following figure.

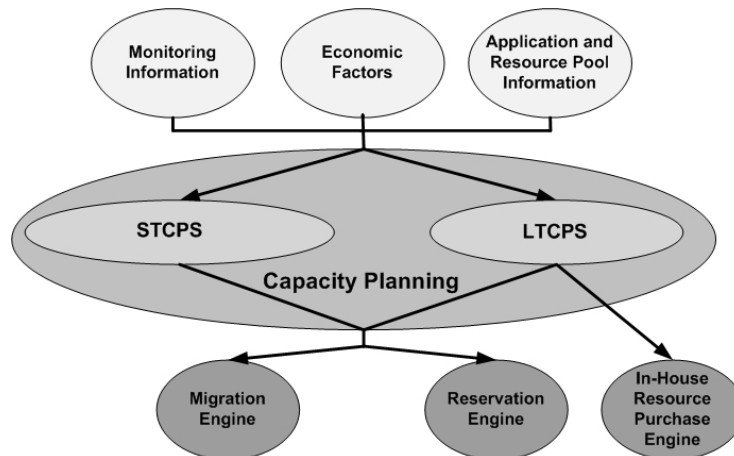


Fig. 2. Capacity Planning Inputs and Outputs.

The first input parameter is the monitoring information. In this case, monitoring refers to three types of actions: measuring the utilization rates of resources, response time analysis of applications, and traffic analysis. In the utilization rate measuring process, the IT staff analyzes to which percentage any given resource is used. Once the utilization approaches a critical level, the resource is classified as overloaded.

The response time analysis determines the response time of applications. The rise of the response time over a certain threshold level indicates that the application has insufficient resources available.

The traffic analysis is used to determine traffic flows within the data center and the traffic flows into and out of the data center. This data can then be used to determine whether individual resources need to be connected differently. The traffic flow information can also be used to determine whether the infrastructure is able to handle all data transmissions.

The second type of input is the economic factors which are stated by the user. These can include requirements (e.g. certain applications have to run in-house) and restrictions. Restrictions can be categorized into financial restrictions or into purchasing restrictions. Furthermore, the user may have a certain budget which has to be considered when creating a new capacity plan.

The third type of input is the information about the resource and application pool. In particular, the capacity planning service has to know which applications are running and what they are being used for, since the use can have a big impact on the resource requirements. For example, a Web server for text-based Web pages has a different load pattern than a Web server which is used for streaming videos.

Furthermore, the capacity planning service has to know which resources are available. This includes not only in-house resources but also resources that have been purchased on the Grid and resources that are available on the Grid.

Based on these inputs, the capacity planning service creates a number of outputs which can either be used by automated programs or by the data center staff. The former is the Migration Engine which is responsible for either migrating applications according to the resource allocation plan generated by the capacity planning service.

The second output consists of a recommendation list for making reservations of computing capacity on the Grid. These can be either short-term reservation recommendations which come from the STCPS or long-term reservation recommendations from the LTCPS. The actual reservations, based on the recommendation list, can be made on behalf of the user by an automated Reservation Engine. Alternatively, the reservations can also be made by the user.

Lastly, the Long-Term Capacity Planning Service can also create a plan for purchasing in-house resources. Since this task cannot be performed automatically, the LTCPS only gives out human-readable list of resources that have to be purchased and the store at which to purchase them.

3 Grid Capacity Planning and Traditional Capacity Planning

There are a number of differences between traditional capacity planning as it is performed today and Grid capacity planning. This section will illustrate these differences and thereby demonstrate the need for a Grid Capacity Planning Service.

3.1 Resource Selection

The outcomes of traditional capacity planning are fairly limited, since there are only three courses of action: purchasing in-house resources, renting or leasing in-house resources, or doing nothing. This lack of fine-grained options does not require a long-winded capacity planning process for small and medium-sized companies. Therefore, the decisions of those companies that can be made can be made quickly, optimizing the costs for the capacity planning procedure [11].

While capacity planning is not an attractive tool in non-utility computing environments, it becomes more important in commercial Grids due to the wider range of options: purchase Grid resources, purchase in-house resources, lease resources, any

combination of the previous, sell spare computing capacity, or do nothing. This increased number of options leads to the problem that an optimal capacity planning solution is not obvious anymore. For example, users willing to sell computing resources must consider the expected income during the capacity planning process.

Overall, due to the increased complexity, the capacity planning staff requires more time, which makes the capacity planning process more expensive and, thus, a utility computing environment less attractive.

3.2 Price Volatility and Demand Fluctuation

The prices of the current computing resource market are static, i.e. resource prices do not change frequently. Differences only occur because of special offers or economies of scale. This means that the capacity planning team does not need to rush the process to avoid rising prices, since even the currently available utility computing resource prices remain constant (e.g. Amazon [12], Tsunamic Technologies [13]).

With the advent of commercial Grids in which companies can purchase and sell resources according to their needs, the changes in supply and demand will lead to fluctuating prices [14, 15, 16, 17]. These varying prices must be taken into account in the capacity planning process.

Furthermore, taking fluctuations of demand into account, it becomes necessary to predict how prices will develop in the future, and thus, the timing of purchases may become a relevant parameter in the capacity planning process. To achieve a precise prediction, the capacity planner must consider the past behavior of the market with respect to the available resources. This means that the capacity planner should not only look at the average demand but also at peak demand times on the utility computing market, which can occur when many Grid users require additional resources. Furthermore, the own demand must be seen in comparison to the peak demand. If there are regular demand peaks on the utility computing market and the own demand peaks occur at the same time, the required Grid resources might only be available at very high prices, which might cause budget problems. On the other hand, if the own demand is anti-cyclical to the market demand, Grid resource prices should not be an issue.

3.3 Application Mapping

Optimizing the mapping of applications to resources also becomes more convoluted in a Grid market environment. In traditional capacity planning scenarios, companies only have to find a mapping of applications to their in-house resources and, if necessary, purchase additional resources for in-house installation. This approach, while not trivial, is manageable, since the number of possible mappings and the resource diversity are fairly small. In fact, once a company knows which resources have to be purchased, the suitable products can be ranked according to their cost.

On the other hand, optimizing the application mapping in a utility computing environment is also more involved. The application-to-resource mapping depends on the resources that could potentially be purchased on the Grid. Therefore, for each applica-

tion, two groups of options have to be considered: running the application on one of the suitable in-house resources, or running the application on one of the suitable Grid resources. Each of the Grid options has its own price, since the pricing structures differ between resource types and resource providers.

Furthermore, applications have to be sorted according to whether they are suitable to run on the Grid or not. Some applications may not run on the Grid because of several reasons, such as applications that require sensitive information for their calculation which is not allowed to be transmitted to external resources.

3.4 Comparison

Grid capacity planning is more elaborate than traditional capacity planning due to the additional options available in computing resource markets. These differences are summarized in Table 1. This increased complexity will mean that companies with little or no IT expertise that are new to the Grid will either not use it or overspend.

Table 1. Comparison of Traditional and Grid Capacity Planning.

	Traditional Capacity Planning	Grid Capacity Planning
Resource Selection	Few courses of action	Many courses of action
Price Volatility	Small	Large
Application Mapping	Small	Large

However, all companies participating in the Grid need to perform the same capacity planning steps and many run similar applications with similar loads. Therefore, it would be useful if the capacity planning process could be outsourced to an external entity which specializes in providing a capacity planning service. This service, the Grid Capacity Planning Service (GCPS), would allow companies to benefit from utility computing by optimizing companies' Grid resource purchases at low costs. Therefore, this service would be a Grid market enabler.

4 A Capacity Planner Model

Following the general model introduced in section 2, the GCPS consists of two parts working in concert. Their workings and interaction is illustrated in more detail in this section.

4.1 The Long-Term Capacity Planning Service

The Long-Term Capacity Planning Service (LTCPS) performs the long-term data analysis as described in section 2. Since its main task is to analyze the current data center computing resource pool and the current application mapping, it must be given this information, in addition to economic information, such as the budget of customer

(both for Grid and in-house resource purchases), the relative importance of each application and whether the customer would be willing to sell resources on the Grid market.

The next step of the LTCPS is to analyze how the applications which have to run in-house (so-called in-house applications) can be mapped to existing resources. The outcome of this analysis can fall into the following categories: (1) the user has to purchase additional in-house resources, (2) the user has idle in-house resources, (3) the user has idle in-house resources but also has to purchase additional resources to satisfy the demand, or (4) the user has no idle in-house resources and all applications have been mapped. In cases 1 and 3, the user has to purchase additional in-house resources. In the remaining cases, the LTCPS can continue the capacity planning process. This is illustrated in the following figure.

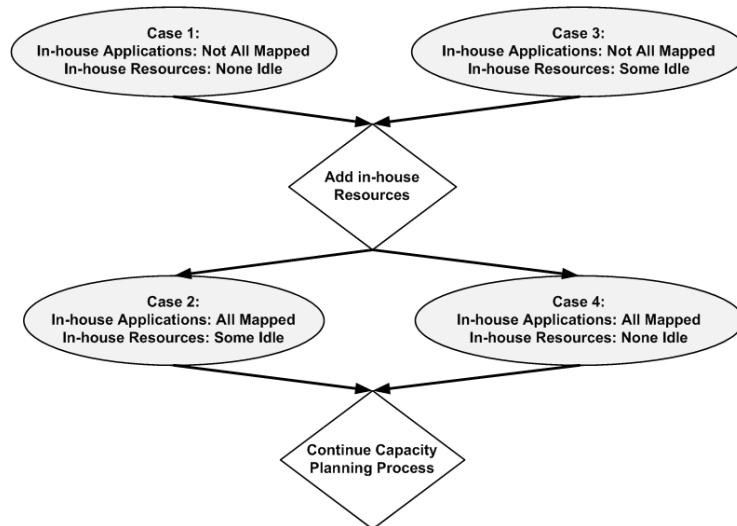


Fig. 3. Initial Steps of the LTCPS.

In the next step, the LTCPS will have to consider the costs and benefits of using Grid resources. In general, it has to weigh (1) using in-house resources versus using Grid resources, (2) purchasing new in-house resources versus using Grid resources and (3) whether the user is willing to sell resources on the Grid and if so, which resource configuration would be the best selling option. All these factors then have to be analyzed with regards to the market issues that have been described previously, namely demand fluctuation and price fluctuations.

The result of this step will be an optimal or near-optimal mapping of applications to resources such that all user requirements are met and that the resources are used as efficiently as possible. The result can fall into one of the following categories: Purchase Grid resources, purchase in-house resource, and purchase both in-house and Grid resources. The purchase of Grid resources could be implemented in the form of a purchasing plan. To avoid high prices, such Grid purchases could be done far in advance by using the Reservation Engine. For regularly occurring demand peaks, peaks could be covered by using Grid resources in addition to in-house resources.

5.2 The Short-Term Capacity Planning Service

The Short-Term Capacity Planning Service (STCPS) performs measurements on the in-house resources to determine their load and the response time of applications. To ensure that these tests do not affect the system adversely, STCPS will only do so periodically. Should it notice that either a resource is being used to maximum capacity or that an application response time is decreasing, it will determine which Grid resource can take up the additional demand.

Furthermore, the STCPS will consider the number of times a similar Grid resource has been purchased in the past. This will allow the STCPS to monitor two important issues: On one hand, it can determine whether these Grid purchases are necessary at regular intervals. If so, it can determine when the Grid resources will be required again and can then suggest reserving Grid resources.

The STCPS will inform the LTCPS of the Grid resource purchase. This allows the LTCPS to determine whether the total cost for these Grid resources approaches the costs of an in-house resource. If this is the case, the LTCPS can warn the user, since this may be a sign that the capacity plan is outdated.

The purchasing information also allows the LTCPS to determine whether Grid resource purchases are occurring at regular intervals. If this is the case, the next purchasing date can be predicted without difficulty and Grid resources can be purchased in advance.

6 Implementation and Validation

An initial test of the performance of both components has been implemented. The services are expected to function within a continuous double auction (CDA) setting, which was implemented using Repast [18]. The simulation environment consisted of 500 agents, which traded resources within this market for 500 days. At the beginning of the day, each agent determines its demand. If the demand is larger than the number of in-house resources, the agent will bid for resources on the Grid market. Should the number of required resources be lower than the number of in-house resources, the agent would attempt to sell the excess resources. The traded resources were made available the following day.

Table 2. Simulation Parameters Overview.

Parameters	Value
Number of agents	500
Number of in-house resources (per agent)	20-40
Market mechanism	CDA
Number of simulated trading days	500
Offer expiration time	1 day
Demand distribution	Normal
Distribution Mean	30
Distribution Variance	30

Using this setup, we developed two scenarios: In the first scenario, the agents used their current demand level to purchase or sell resources. The result of this simulation can be seen in Fig. 4, which shows the number of available resources. A negative value shows that the agent has fewer resources than required, while a positive value shows that the agent has more resources than it requires.

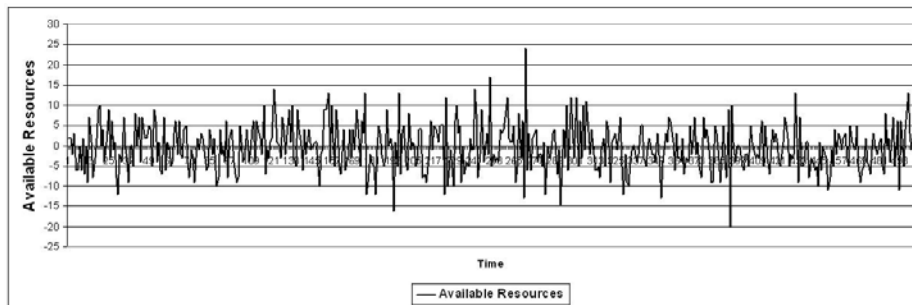


Fig. 4. Resource Availability with Basic Capacity Planning.

The spikes in the graph show that the agent rarely has the correct number of resources available. This fact shows that this very basic capacity planning approach is far from optimal when it comes to predicting the resource.

The second scenario worked with a more complex capacity planning approach: The agents' capabilities were expanded to allow predicting their demand based on past resource requirements. The requirements prediction was implemented using the linear regression tool of the Apache Commons Math Toolbox [19]. The linear regression used the demand from the past 30 days to predict the demand for the next day. This information was then used to buy or sell resources. The result of this simulation is shown in Fig. 5 below.

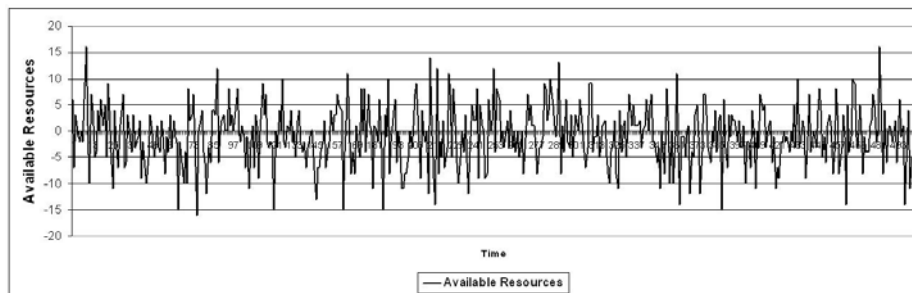


Fig. 5. Resource Availability with More Advanced Capacity Planning.

Fig. 5 shows that the peaks are no longer as large as before and that the extreme peaks no longer occur with the agents. While this is not a marked improvement, it should be noted that the prediction algorithm is still fairly basic.

The simulations demonstrated that the GCPS is indeed a valuable tool in a Grid market environment in which price volatility and demand fluctuation have to be con-

sidered. The GCPS can ensure that a company will have sufficient resources at its disposal in such an environment. Since these comparisons are also computationally fast, the entire capacity planning process in this environment took only a few milliseconds per agent. However, the simulations also showed that much remains to be done to improve the predictive capabilities of this service.

7 Conclusion

In this paper we defined capacity planning for utility computing and placed it in context with load balancing, scheduling, and reservations. Furthermore, we have shown that capacity planning is more complex in a Grid environment than traditional capacity planning. Due to the complexity, we believe that a Grid Capacity Planning Service is required for a successful Grid usage, since performing capacity planning using in-house staff is costly and would negate the benefits of utility computing.

The GCPS described in this paper consists of two distinct parts: the Short-Term Capacity Planning Service and the Long-Term Capacity Planning Service. This structure reflects the fact that capacity planning in a commercial Grid environment has to be used to solve short-term and long-term problems. The first is responsible for ensuring that all applications and resources are running as required by the user and will give advice regarding additional resources if necessary. The latter is responsible for long-term planning of data centers and takes into account the resource requirements of all applications, the available in-house resources, prices, demand fluctuations, and the user requirements. Using this information, a mapping of all applications is found and (if necessary) recommendations for resource purchases are made.

Furthermore, the GCPS has been implemented and initial tests have shown that the performance overhead is low. Future work will center on refining the capacity planning algorithms of the two components, since the simulations have also shown that the demand prediction has to be improved.

8 References

1. IBM: A Statistical Approach to Capacity Planning for On-Demand Computing Services,
<http://domino.watson.ibm.com/comm/research.nsf/pages/r.statistics.innovation2.html>
2. Li, C., Li, L.: Competitive proportional resource allocation policy for computational grid. *Future Generation Computer Systems*, vol. 20, no. 6, 1041-1054 (2004)
3. Yu, J., Li, M., Ying, L., Hong, F., Gao, M.: A Framework for Price-Based Resource Allocation on the Grid. In: Liew, K.M., Shen, H., See, S., Cai, W., Fan, P., Horiguchi, S. (eds.) *PDCAT 2004. LNCS*, vol. 3320, pp. 341-344. Springer, Heidelberg (2004)
4. Li, C., Li, L.: Dynamic resource allocation for joint grid user and provider optimisation in computational grid. *International Journal of Computer Applications in Technology*, vol. 26, no.4, 242 - 250 (2006)

5. Wolski, R., Brevik, J., Plank, J. S., Bryan, T.: Grid Resource Allocation and Control Using Computational Economies. In: Berman, F., Fox, G., Hey, T. (eds.), pp. 747-771. John Wiley & Sons, Hoboken (2003)
6. Pourebrahimi, B., Bertels, K., Kandru, G.M., Vassiliadis, S.: Market-Based Resource Allocation in Grids. In: Second IEEE International Conference on e-Science and Grid Computing, pp. 80-88. IEEE Press, New York (2006)
7. Afzal, A., McGough, A.S., Darlington, J.: Capacity planning and scheduling in Grid computing environments. *Future Generation Computer Systems*, vol. 24, no. 5, 404-414 (2008)
8. Siddiqui, M., Villazon, A., Fahringer, T.: Grid Capacity Planning with Negotiation-based Advance Reservation for Optimized QoS. In: SC'06, pp. 21-37. IEEE Press, New York (2006)
9. Borowsky, E., Golding, R., Jacobson, P., Merchant, A., Schreier, L., Spasojevic, M., Wilkes, J.: Capacity planning with phased workloads. In: Proceedings of the 1st international Workshop on Software and Performance, pp. 199-207. ACM, New York (1998)
10. Risch, M., Altmann, J.: Cost Analysis of Current Grids and its Implications for Future Grid Markets. In: Altmann, J., Neumann, D., Fahringer, T. (eds.) Gecon 2008. LNCS, vol. 5206, pp.13-27. Springer, Heidelberg (2008)
11. Risch, M., Altmann, J., Makrypoulas, Y., Soursos, S.: Economics-Aware Capacity Planning for Commercial Grids. In: Collaborations and the Knowledge Economy, vol. 5, pp. 1197-1205. IOS Press, Amsterdam (2008)
12. Amazon Elastic Compute Cloud (Amazon EC2), <http://www.amazon.com/gp/browse.html?node=201590011>
13. Tsunami Technologies Inc., <http://www.clusterondemand.com/>
14. Regev, O. and Nisan, N. 1998. The POPCORN market—an online market for computational resources. In: Proceedings of the First international Conference on information and Computation, pp. 148-157. ACM, New York (1998)
15. Waldspurger, C.A., Hogg, T., Huberman, B.A., Kephart, J.O., Stornetta, W.S.: Spawn: A Distributed Computational Economy. In: *IEEE Transactions on Software Engineering*, vol. 18, no. 2, 103-117 (1992)
16. Buyya R, Abramson D, Giddy J.: An economy grid architecture for service-oriented grid computing. In: 10th IEEE International Heterogeneous Computing Workshop. IEEE Computer Society Press, Los Alamitos (2001)
17. Lai, K., Rasmusson, L., Adar, E., Zhang, L., Huberman, B.A.: Tycoon: An implementation of a distributed, market-based resource allocation system. *Multiagent and Grid Systems*, vol. 1, no. 3, 169-182 (2005)
18. Repast Simulation Environment, <http://repast.sourceforge.net/>
19. Apache Commons Math Libraries, <http://commons.apache.org/math/>