Economics-Aware Capacity Planning for Commercial Grids

Marcel RISCH[1], Jörn ALTMANN[1,2], Yannis MAKRYPOULIAS[3], Sergios SOURSOS[3]
[1]*International University in Germany, School of Information Technology,*
Campus 3 Bruchsal, 76646, Germany,
Tel: +49 (0)7251 7000, Fax: + 49 (0)7251 700150, Email: marcel.risch@i-u.de
[2]*TEMEP, School of Engineering, San 56-1, Sillim-Dong,*
Gwanak-Gu, Seoul, 151-742, South-Korea
Tel: +1 510 972 3062, Fax: +1501 641 5384, Email: jorn.altmann@acm.org
[3]*Athens University of Economics and Business, Dept. of Informatics,*
76, Patission str, Athens, 10434, Greece
Tel: +30 210 8203693, Fax: + 30 210 8203693, Email: {makrupoul06, sns}@aueb.gr.

Abstract: Currently, capacity planning is fairly simple, due to the few options that are available. With the advent of Grid markets, this discipline for analyzing resource purchases must be adapted. A commercial Grid provides many different resource types at variable prices, making capacity planning more complicated than it currently is. In this paper, we describe the functionality of an online Grid Capacity Planning Service, which helps companies with little IT expertise to make use of the Grid in a cost-effective manner. The requirements of the capacity planning service are derived in part from a survey carried out among SMEs in the region of the German city of Bruchsal. Using the requirements, we identified all the necessary information from the Grid and we designed and implemented parts of the capacity planning service.

1. Introduction

The main foci of Grid research fall into two categories: In the first category are the research and development projects, which aim at developing Grid architectures. These include Globus [1], GRIA [2], Gridbus [3], glite [4] and NextGRID [5]. In the second category are the business-related projects, which analyze and develop business models for Grid systems. These include, amongst others, AssessGrid [6], Biz2Grid [7] and BeInGrid [8].

However, the commercial Grids, which are slowly developing, do not seem to rely on the principles developed by these projects. Instead, existing commercial Grids have been developed by companies, which have created their own approaches to Grid computing, such as the Amazon.com EC2 [9], the Sun Grid [10], and the Tsunamic Technologies Grid [11]. Some of these services have started to become very popular and, thus, prove that the idea of commercial Grid computing is indeed acceptable to resource providers and buyers [12]. This is illustrated in SecondLife Blog [13], which illustrated that for short demand peaks, the Amazon service was cheaper than installing additional in-house bandwidth capacity.

Existing research in commercial Grids has largely failed to consider the economic aspects of Grid computing. It has been tacitly assumed that, by providing the proper technical environment, customers would be convinced of the advantages of Grid technologies. This shortcoming of existing Grid research has been identified by the GridEcon project [14], which started out with the premise that economic considerations complement technical solutions. A number of economic-enhanced support services have been identified which are needed for a commercial Grid to function. These services include

various brokers but also a capacity planning service, which helps customers to determine their resource needs.

This paper shows the steps towards the development of a Grid Capacity Planning Service, which will help Grid users to plan their capacity requirements. First, we will define the term 'capacity planning', since capacity planning has lost some of its importance in data center planning and since there are many different definitions for this task. Second, we demonstrate the need for a capacity planning service in a commercial Grid environment through the analysis of the differences between traditional capacity planning and Grid capacity planning as well as through the analysis of a qualitative survey. The survey addressed SMEs from diverse fields such as banking, software development, and car sales. From the survey results, we determined the requirements that SMEs have towards Grid markets and towards capacity planning in particular. Based on these results, the requirements of the capacity planning service are determined and an architecture of a Grid Capacity Planning Service is designed and developed.

The structure of the paper is devised as follows: In Section 2, capacity planning is put into context with other data center tasks. In Section 3, the differences between traditional and Grid capacity planning are explained and, from these differences, the requirements for a Grid capacity planning service as well as the design of the service are derived. The Short-Term Capacity Planning Service is then explained in more detail in Section 4, before the paper is concluded in Section 5.

2. Capacity Planning Definition

2.1 Capacity Planning Definition

In [15], it has been remarked that the term "capacity planning" is frequently used but rarely defined. To correct this deficiency, we will start by defining the term, using the definition given by IBM in [16]:

> "*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. While some research has been done on the provider's capacity planning problems [17][18], no research has been done, as of yet, on the customer's need for capacity planning in a Grid environment.

According to the definition, the following three tasks are at the heart of the capacity planning process: (1) Monitoring the current resource utilization rate and the application response times; (2) Estimating the future resource requirements of applications; (3) Cost monitoring to ensure that a company does not overspend. Based on these tasks, four courses of action are open to companies: (1) purchasing in-house resources, (2) renting or leasing in-house resources, (3) doing nothing, or (4), in the case of the existence of a commercial Grid, purchasing Grid resources.

2.2 Conceptual Classification of Capacity Planning

The capacity planning process should guarantee that a basic mapping of applications to resources (i.e. resource allocation) is always possible. This implies that resource allocation is a sub-task of the capacity planning process. At the same time, the capacity planning process should ensure that resources have similar load levels, which implies that load balancing is also a subtask of the capacity planning process. Load balancing and resource allocation already have one task in common, namely monitoring.

Economic aspects have always been a part of capacity planning. Since data centers have budget constraints, the IT personnel needs to determine which resource has the best value

for the price. These issues are also faced in a commercial Grid, where the IT personnel has to compare different Grid alternatives. As we will outline in the next section, the task of choosing the optimal resource is even more complex because of the usage-based pricing structure of commercial Grids. It forces the data center personnel to predict the resource usage very precisely, in order to achieve low costs for Grid usage.

3. Towards a Grid Capacity Planning Service

3.1 Complexity of Decision Making in Capacity Planning

This section demonstrates why capacity planning is very important in a Grid environment. It also discusses the differences between traditional capacity planning and Grid capacity planning. For this analysis, we assume that a functioning Grid economy exists, that is, prices are determined according to the supply and demand situation. The currently existing Grid computing offers (such as Amazon EC2, Sun and Tsunamic) have fixed prices.

3.1.1 Effort of Capacity Planning

Capacity planning is not popular with companies, since the effort does not bear any relation to the expected benefits, as can be demonstrated with the following example:

We assume that an employee incurs costs of $4500 per month (salary and human resource management) to a company. Thus, one week of work on planning the computing resources costs the company about $1000. This cost must be added to the cost of the computing resources. If this company is a small company, the actual savings through an accurate capacity planning procedure is very low or even non-existent. Consequently, capacity planning is not widely used. This fact could be verified during a qualitative survey of SMEs in the Bruchsal region (Germany): When asked to describe their capacity planning process, companies stated that they used their “gut feeling” to decide which resource to purchase. Other companies simply bought one of the most powerful computing resources available to ensure that it would be able to run future applications, which are expected to require more powerful resources.

In a commercial Grid environment, provider companies have even more capacity planning inputs to consider to determine the best resource allocation. Thus, the capacity planning process becomes even more expensive. This process is further complicated by the fact that Grid users may be willing to sell excess resources on the Grid. In this case, the expected income must be taken into account when calculating the Grid usage costs. This increased complexity will require the data center staff to spend more time on the capacity planning process, which in turn reduces the benefit of capacity planning further.

3.1.2 Resource Diversity

In traditional capacity planning, the IT personnel only has to select new hardware from the resources currently available in the market. Although not standardized, these resources usually have similar features and thus pose few difficulties for professional staff. Therefore, should a resource be added or replaced, the data center staff will be able to find a similar resource that can perform the job adequately well, without having to perform any performance testing. This has also been reported in the survey: Companies stated that new resources are better in every aspect than existing resources, so that all existing applications and future applications (with increased resource requirements) can still be executed.

In a Grid environment, however, the diversity of resources is much greater, since providers may offer any kind of resource (e.g. virtual machines of old computers).

3.1.3 Price Volatility

The current computing resource market is fairly static in that resource prices do not change frequently. Since current resources are usually bought for in-house installation, price variations only occur because of special offers or economies of scale. Therefore, the capacity planning team does not need to rush the capacity planning process to avoid changing prices. Even if the capacity planning team decides to make use of current commercial Grid resources (e.g. Amazon [9], Sun [10], Tsunamic [11]), those prices, although high, remain unchanged [19].

With the advent of commercial Grids, which sell resources at competitive prices, prices will change according to the variation in supply and demand. The capacity planning team has to consider these price fluctuations and has to predict how the prices will develop. Furthermore, with changing prices, the timing of purchases may become a relevant parameter in the capacity planning process: The demand peaks have to be analyzed with respect to the market prices, in order to determine whether the demand peaks coincide with times of high Grid prices.

3.2 Requirements of a Grid Capacity Planning Service

In order to be accepted by Grid users, a Grid Capacity Planning Service (GCPS) should fulfill a number of requirements and provide a set of basic functionality.

The functionality that should be offered by the GCPS is the monitoring service. It monitors Grid and in-house resources. If a performance requirement is no longer met (e.g. an exceeded response time limit or exceeded load level), the GCPS triggers the resource analysis process in order to determine a course of action to satisfy all user demands.

During the resource analysis process, the Grid Capacity Planning Service should take the user's in-house resources into account and should ensure that these are used. Since these resources have been purchased and have been installed, they should be used as much as possible to reduce costs. However, if the user is willing to sell certain resources on the Grid, the GCPS has to take this into account as well. There may be cases in which it is advantageous to sell some of the in-house resources on the Grid.

When the GCPS has determined which courses of action are viable, it should give a ranked list of options to the user, who can then decide which of these options should be executed. This final selection step allows the user to remain in control of his budget and to ensure that tacit requirements are also met.

Furthermore, the GCPS should perform application-resource-mappings. Since the GCPS needs to determine the requirements of applications under different circumstances, the GCPS has to run performance tests for applications on various resource types. At the same time, this procedure will have to be performed in a cost-effective manner. In a commercial Grid environment in which prices fluctuate, the GCPS should perform these actions during times of low resource prices. The results of the performance tests and the thereby derived requirements for each application should be stored locally within the GCPS to ensure that all information is readily available.

Finally, the Grid Capacity Planning Service should perform as many of these actions as possible automatically, i.e. without the help of a human operator. There are two arguments that count against human involvement: the first is that people are slow and error prone and thus the result generated by the GCPS would not be as accurate. The second reason is that experienced and specialized employees are expensive and, thus, would cause higher costs.

3.3 A Grid Capacity Planning Service Model

To fulfill all requirements, we envision a Grid Capacity Planning Service (GCPS), which consists of two distinct parts that work in concert:

- Part 1: Long-Term Capacity Planning Service (LTCPS): This is an online service, which performs the long-term analysis of the user's resource situation. It takes input parameters (e.g. information about in-house resources, application information, and user requirements, and information from the Short-Term Capacity Planning Service) before determining a ranked list of possible courses of action for in-house resource purchases and rentals, and Grid resource purchases and sales.
- Part 2: Short-Term Capacity Planning Service (STCPS): This service suggests to the user the number of machines required to meet his short-term performance or economic goals, over a specific time period. To do so, it uses information on past usage of Grid resources, considers characteristics of the application the user wants to run on the Grid and estimates the application's load that is expected for the requested time period.

These two capacity planning services complement each other, since one addresses the short-term problems while the other addresses the medium- to long-term problems. The two services will be described in more detail in the following two sections.

3.3.1 The Long-Term Capacity Planning Service

Since the LTCPS should plan as far into the future as possible, it requires a lot of information, which can be grouped into three categories:

The first category includes information about user-owned resources, user-owned applications, as well as user requirements. Information about user-owned resources and applications describe which resources are installed in-house and which applications are to be run. The information about user requirements is more complex. For example, many companies have sensitive data, which should not be processed by applications beyond company boundaries. These applications are not permitted to run on the Grid, since this would violate company policy (i.e. a user requirement). Another example for a user requirement is the maximum expenditures for Grid resources, runtime limits, the response time limits, or the fact that the in-house resources must be used to their full extend.

The second category consists of application requirements. These can be divided into minimum and optimal requirements and include items such as hardware requirements and software requirements. The software requirements provide information about additional software that must be available on a machine. The software requirements also include requirements for communication with other software components (e.g. communication with other sub-jobs within a workflow). Furthermore, the application requirements should also include the usage frequency of the application and the average usage duration.

The third category contains information about the Grid market and is collected by the LTCPS. It comprises the current and past prices of Grid resources, the availability of suitable Grid resources, as well as the prices for in-house resources. This last point is still a challenge, since many resource vendors do not use any standard for composing computers that would allow a potential buyer to determine computer prices automatically.

Next, these three sets of information are used to determine a mapping of applications to resources in such a way that all user requirements are met. The result can fall into one of the following categories: in-house resources are sufficient, purchase Grid resources, purchase in-house resources, sell in-house resources on the Grid, or purchase both in-house and Grid resources. The operation of the LTCPS can be seen in Figure 1 below.

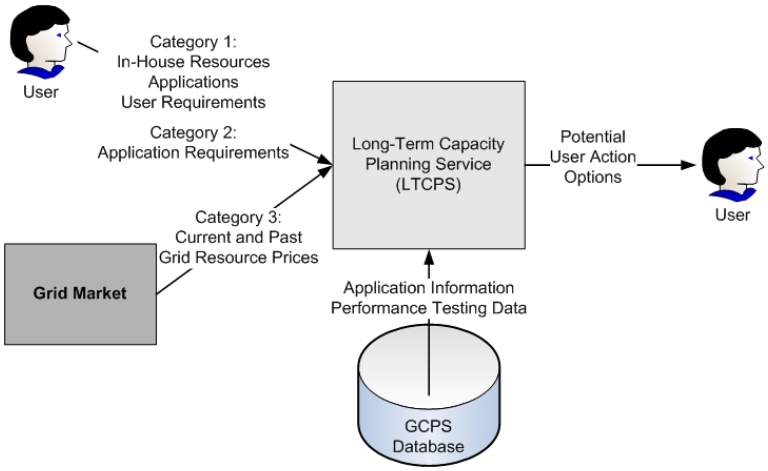

Figure 1: The Long-Term Capacity Planning Service

3.3.2 The Short-Term Capacity Planning Service

The STCPS takes input from the user with respect to the performance and cost-minimization objectives. Then, it proposes the number of machines to be purchased on the Grid, in order to fulfill the user requirements.

We assume that performance objectives are expressed in terms of time delay (i.e. the response time of an application for serving a request). Users (e.g. application providers) find it easy to express their requirements in terms of delays, since the system response time can be experienced by users.

For cost-minimization purposes, further information is required. First, the information about the cost of purchasing a single machine on the Grid market is necessary. Second, it is required that the user provides a cost function, i.e. a function that relates the experienced delay per application request with a monetary value, representing the cost that is incurred to the provider per millisecond of delay per request.

In addition to the information about the objectives of the user, the STCPS needs to have access to monitoring data about the resource load. The monitoring data is needed to make the necessary estimations and predictions of the load (number and type of requests) that the application will face in the future time period for which the plan is required. All these input are summarized in Figure 2 below.

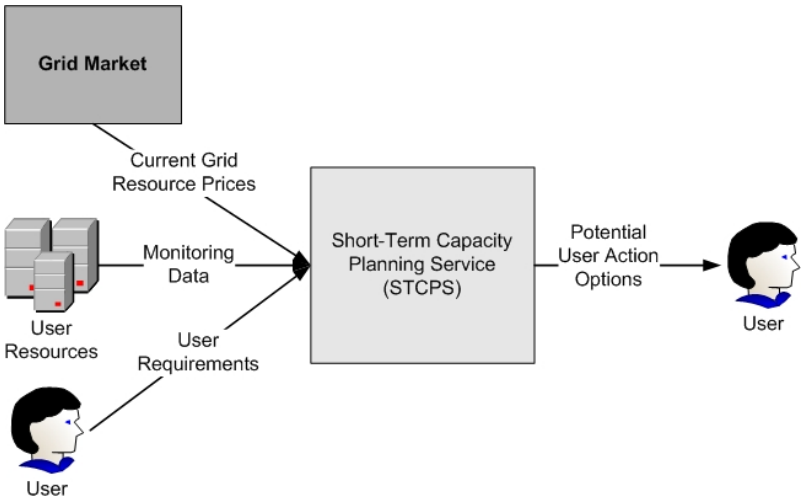

Figure 2: The Short-Term Capacity Planning Service

4. A Closer Look at the Short-Term Capacity Planning Service

Within the GridEcon project [14], we have designed and implemented the Short-Term Capacity Planning Service. In the following sections, we will present the architecture of this system and describe the functionality of the components.

For the STCPS to function, we have to collect monitoring data about every application that runs on the Grid and to be able to categorize applications according to some common characteristics. We assume that every Grid system provides some public interfaces for accessing monitoring data. Using these interfaces, we can obtain all the needed information and store it to a local database, namely the History Component.

After collecting the necessary data, we run a classification algorithm, executed by the Clustering Component. Clustering is necessary for the STCPS, since it allows working with an aggregated set of application-related data. The criterion according to which the application is classified is the execution time of a single request to a virtual machine on the Grid. Our assumption is that applications with similar execution times per request are similar enough to be considered as identical applications. For sake of simplicity, we use clustering with only two classes as an output. This parameter of the system can be changed to allow clustering results with a higher number of classes.

After having aggregated the historical data about the load, we can estimate the load expected for the time period defined by the user. This task is assigned to another module, the Workload Predictor Component. It takes into account the load of the previous hours, along with the load encountered at the same day and time over the past few weeks. Thus, we capture the current trend of the load as well as the behavior that appears periodically.

After the workload prediction is complete, the Decision Support Component has all information necessary to propose to the user the minimum number of virtual machines needed to meet the requirements. To give the user a better recommendation, we have implemented two models: an M/M/k queuing model and a Support Vector Machine model. Both models give answers to the same questions, e.g. how many machines are required such that the application A provides an average service time of Y milliseconds for the next 2 hours? Or, how many machines are required to minimize the provisioning cost, provided that each virtual machine costs x Euros per hour and the cost function is of a certain shape?

All the aforementioned components, along with their internal and external interactions, are depicted in Figure 3.

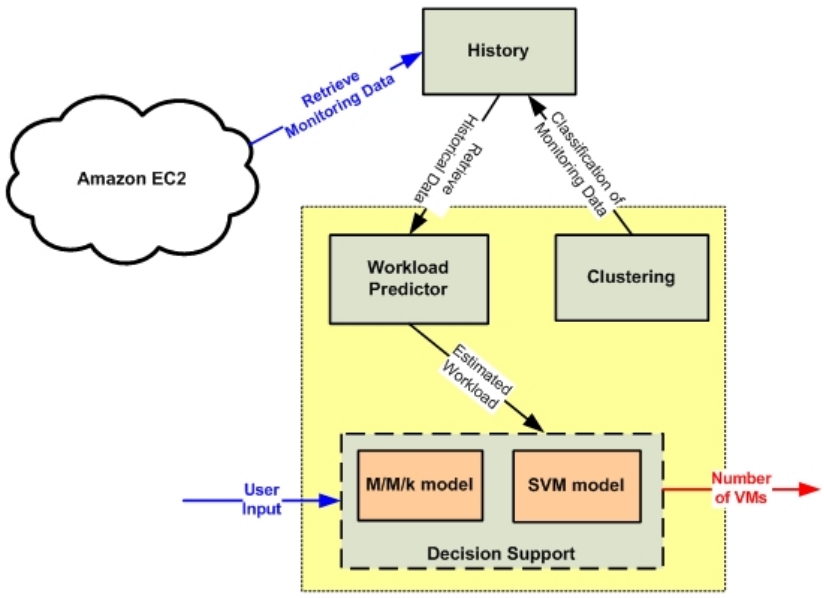

Figure 3: The Architecture of the STCPS

The system is already implemented using the .NET platform and relying on the Amazon EC2 environment. For our tests, we also designed and implemented an application-level monitoring module. The test application that is executed on EC2 is a Web Server with different JSP web pages that emulate the different Web applications. We are currently testing the application and the various approaches used in the design, in order to evaluate the contribution of such a system.

5. Conclusions

In this paper we have defined the term "capacity planning" and have placed it in a conceptual context for Grid computing. Furthermore, we have shown that capacity planning is rarely used in companies today and that it is more complex in a Grid environment. Due to the complexity, we believe that a Grid Capacity Planning Service (GCPS) is necessary for not only a successful Grid usage but also for making Grid computing widely used.

The GCPS described in this paper has two parts: the Short-Term Capacity Planning Service and the Long-Term Capacity Planning Service. The first is responsible for ensuring that all applications are running as required and will give advice regarding additional resources if they become necessary for a short time period. 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, and the user requirements.

Since only an outline of the LTCPS and a basic testing implementation of the STCPS have been developed, several issues still need to be addressed in implementing the GCPS: Firstly, the efficiency of this service needs to be investigated. Since performance testing is expensive, the GCPS needs to perform as little performance testing as possible without ignoring available resource types. Secondly, the GCPS needs to take into account that an application can have different resource requirements depending on how it is used. Finally, the GCPS has to ensure that it performs all capacity planning tasks quickly, even if the user has many courses of action open. This may require storing common resource allocation solutions but, at the same time, avoiding the storing of excessive amounts of data.

References

[1] The globus alliance, http://www.globus.org/, 2008.
[2] Gria, http://www.gria.org/, 2008.
[3] Gridbus, http://www.gridbus.org/, 2008.
[4] gLite, http://glite.web.cern.ch/glite/, 2008.
[5] NextGRID: Architecture for Next Generation Grids, http://www.nextgrid.org/, 2008.
[6] AssessGrid, http://www.assessgrid.eu/, 2008.
[7] Biz2Grid, http://www.d-grid.de/index.php?id=407&L=1, 2008.
[8] BeInGrid, http://www.beingrid.eu/, 2008.
[9] Amazon Elastic Compute Cloud (Amazon EC2), http://www.amazon.com/gp/browse.html?node=201590011, 2008.
[10] Sun Grid, http://www.sun.com/service/sungrid/index.jsp, 2008.
[11] Tsunamic Technologies Inc., http://www.clusterondemand.com/, 2008.
[12] Techcrunch: http://www.techcrunch.com/2008/04/21/who-are-the-biggest-users-of-amazon-web-services-its-not-startups/, 2008.
[13] SecondLife Blog, http://blog.secondlife.com/2006/10/26/amazon-s3-for-the-win/, 2008.
[14] GridEcon, http://www.gridecon.eu, 2008.
[15] Cortada, J.W.: Managing DP Hardware. Capacity Planning, Cost Justification, Availability and Energy Management. Prentice-Hall, Inc., Englewood Cliffs (1983).
[16] 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, 2008.
[17] Siddiqui, M., Villazon, A., Fahringer, T.: Grid Capacity Planning with Negotiation-based Advance Reservation for Optimized QoS. In: Supercomputing, 2006. SC '06. Proceedings of the ACM/IEEE SC 2006 Conference, IEEE, 2006.

[18] Borowsky, E., Golding, R., Jacobson, P., Merchant, A., Schreier, L., Spasojevic, M., and Wilkes, J.: Capacity planning with phased workloads. In: Proceedings of the 1st international Workshop on Software and Performance. WOSP '98, pp. 199-207 ACM, New York, NY, 1998.
[19] Risch, M., Altmann, J.: Cost Analysis of Current Grids and its Implications for Future Grid Markets. In: Proceedings of the Grid Economics and Business Model Workshop. Gecon 2008, pp.13-27. Springer LNCS. Heidelberg, 2008.