Services Pricing: A Shared Grid Case Study

Nabil Abdennadher, Aarti Agrawal, Emmanuel Fragniére and Francesco Moresino

Abstract—In recent years much research has gone into grid computing and building shared resource networks. The grid computing has given access to the research community to high computational resources. The pricing of grid resources has been evolved along with resource management however there is a need for a better pricing model which is well suited for services. In this paper, we first explain the notion of service pricing in the context of shared grid computing. Then, we propose an original development of a mathematical programming model based on Markov Decision Processes (MDP) in order to maximize grid service level. An example written in AMPL modeling language is provided.

I. INTRODUCTION

Grid computing has become increasingly important in recent years with ever increasing demand of computing and storage resources [6]. Sharing, selection and collection of geographically distributed resources such as super computers, storage systems, data resources and specialized devices is made possible by grid network for solving large-scale resource intensive problems in science, engineering and commerce. Organization that are looking for extremely high computing power for short periods of time, but which do not necessarily want to invest further in their own computing resources are finding grid computing as an attractive alternative. Several large cooperation’s such as IBM, Unilever, Ericsson and Hitachi are investing large sums of money in developing grid computing initiatives [7].

The idea of creating a grid to share resources which seems to be very intuitive now was originally borrowed from electrical power grid. In the mid-1990 scientists began to explore the design and development of an infrastructure which was analogous to electric grid due to its pervasiveness, ease of use and reliability [8]. The motivation for computational grids was initially driven by large scale, resource (computational and data) intensive scientific applications that require more resource than a single computer (PC, workstation, supercomputer, or cluster) [6].

A. The EZ-Grid and Interreg grid projects

The Virtual EZ-Grid project involves several academic Swiss institutions: Geneva University, Swiss Italian University, Neuchatel University and Western Swiss University of Applied Science. The Virtual EZ-Grid project has been recently connected to another project call Interreg involving the EZ-Grid team plus academic partners from France (in particular the University of Franche Comté). The platform that will be developed in the context of the project can be used to deploy a large range of high performance computing applications. Direct users of the grid will be the researchers griddifying existing applications, whereas for the end users of the systems, the use of a grid for the computation will be invisible. Practically, we are building a grid that will comprise the computing resources at around ten different institutes connected via internet. This grid can later be extended to more universities or public users. This platform will essentially be used to deploy two medical applications, each having a large user community: medGIFT and NeuroWeb. The overall infrastructure will involve more than 1500 nodes spread around the institutions taking part to the project. The expected number of users is 20 in the first years. The authors of this paper are in charge of setting up a new pricing scheme for the optimum resource usage of the EZ-Grid.

B. Organization of the text

This paper is organized as follows. In Section II, we present the particular nature of grid pricing services. In Section III, we explain notions of market based resources in the context of shared grid computing. In Section IV, we present the pricing model we have developed whose goal is to optimize an intangible factor of perceived value, namely the respect of priorities in the execution of jobs. In Section V, we illustrate our model with a simple instance and present its main results. In Conclusion, we indicate further directions.

II. SERVICE PROPOSAL AND PRICING FOR GRID USERS

In most grid projects, the connected PCs are generally made available at no cost. Indeed it serves the needs of research projects requiring huge computing power. However, economic theory tells us that if prices are set in a fair manner between the provider and the user of the service, this should lead to an optimal allocation of resources. For users, the value of a grid is higher than its accounting value. Indeed, the grid offers more than software and hardware: it offers a service. In practice, the pricing of services is mainly based on the cost structure, which does not take into account the real value provided to the client.

What could be done for pricing grid services? Most of the pricing techniques applied to the service sector are devoted to service commodities like airplane seats or hotel rooms. Service activities are traditionally described with the help of the IHIP paradigm (intangibility, heterogeneity, instantaneity and perishability). Intangibility and heterogeneity make the pricing scheme not easy to model services. The pricing of a service relies upon three pillars: internal organizational
costs, the competitors prices and the perceived value by the customers. In this research, we focus our attention on the latter, i.e. the value perceived by the customers in the service experience.

As a matter of fact the question of pricing is essential to analyzing the system of service production. Pricing directly impacts how the service is positioned, and influences equally how clients and coproducers will interact. The paradigm of the price/quantity economic model cannot be fully realized in the service world, however. Quantity is replaced by the idea of value perception, which naturally impacts the production model.

Grid services fall in the category of knowledge based services. Consequently, a service plan based on know-how and expertise depends on the following: a thorough and clear understanding of the needs and expectations of clients, the ability to elaborate a diagnosis of client needs from limited information, the outline of a specific service proposal, the efficient use of delivery processes and of existing products (or product modules), and finally a custom-made solution that incorporates perceived value-added (often referred to as problem resolution).

Very often, delivering more value in the knowledge-based services means delivering more tangibility and more customization. Customization has been defined as the ability of the service delivery system and its employees to attend flexibly to customer needs [3]. Customization may also induce more risky operations as services would become inherently variable in how they are conducted, and according to [4], it is to be expected that problems will occur. Therefore, how to make attractive and price service operations that increase tangibility and customization for the customer while reducing (or maintaining) the complexity of managing real heterogeneity?

In recent research, we are exploring the potential pricing schemes that could be established and the relevant criteria that should be used to determine a fair price in the context of grid services taking into account the point of views of providers and consumers of knowledge-based services. To answer to this question, we must develop a grid service model based on the main attributes (also called salient attributes of the service in the share of choice model terminology [5]) as perceived by the user. In our case, we could retain several attributes such as the capacity made available to the user (three formats - small, medium and large), the proper priority management of jobs executed (the objective that is to be retained in the model presented hereafter), or reporting information for the users. In this paper we will assume that priority management is a salient attribute of grid services, which is perceived as important in terms of value. However it is clear that in a subsequent research, we need to investigate through survey techniques what represent the most important salient attributes for our grid users (mainly researchers in medical studies).

In this project, we intend to develop a pricing scheme based on the perceived value by the user of an actual grid application (EZ-Grid and Interreg projects). This contribution is based on interdisciplinary research. The pricing model essentially borrows academic findings arising from Marketing and Service Science. The solution relies on a number of different fields: Operations Research, Game Theory and Negotiation Theory. We consider three cases that are relevant for different situations: 1. the grid is managed by a central agency, 2. the actors in the grid are competing without cooperation, 3. the actors in the grid are competing but cooperation is possible.

In this paper, we will solely tackle the case where the grid is managed by a central agency. For instance, we have already explored the shadow prices approach through an optimization problem called the share of choice model [5]. This latter model attempted to optimize the design of a service by selecting its attributes according to the perceived value of a sample of clients. Perceived value were expressed as utility functions obtained from conjoint analysis. Conjoint analysis techniques enable constructions of path-worth or utility functions for each respondent regarding the different attributes of the service.

Regarding what has been presented, our goal is compute a price for the utilization of the resources such as CPU, RAM, bandwidth, libraries, etc. However, the price will be differentiated as a function of the quality of the service (in particular, privileged users that can have priority for their jobs) and also as a function of the level of demand (different prices for rush hours and off-peak hours).

III. MARKET BASED RESOURCE MANAGEMENT

A. Overview

There has been extensive research on how to manage the resources in efficient way, optimize the usage, balance the load and reach maximum user satisfaction. Sharing resources fairly with economic efficiency, where efficiency can be defined as ratio of the actual total benefit for all users to optimal total benefit, at a low cost still remains a challenge [9]. The fact that consumers have different goals, preferences and policy further adds to the complexity of resources management [10]. Software agents such as automatic scheduling programs and negotiation agents can play an essential role in realizing this vision of the grid. Numerous economic models for grid resource management such as commodity market models, auction, contract-net/tendering models, bargaining models, posted price models, bid-based proportional resource sharing models, cooperative bartering models, and monopoly and oligopoly had been proposed in the literature [11]. While some of the more commonly-referenced work focused on commodity markets, auctions and fixed budget based marketing mechanism for the resource management [12]. Auctioning has long been an important aspect of many economies. Indeed, it provides a fair trading environment as a decentralized structure, are easier to implement than other economic models and respect the autonomy of resource owners [13]. In Another variant of auctioning fixed budget mechanism efficiency and fairness of the allocation of resources at equilibrium is evaluated through the measures of 'utility uniformity' and 'envy -freshness'. The grid resource
allocation model is based on Continuous Double Auctions (CDA). Different scheduling strategies have been analyzed which can be applied by the user to execute workflows in such an environment, and try to identify the general behavioral patterns that can lead to a fast and cheap workflow execution [14].

In history based pricing model consumers and producers determine their bid and Ask prices using a sophisticated history-based dynamic pricing strategy and the auctioneer follows a discriminatory pricing policy which sets the transaction price individually for each matched buyer-seller pair. The pricing strategy presented generally simulates human intelligence in order to define a logical price by local analysis of the previous trade cases. Here we employ a continuous double auction protocol as an economic-based approach to allocate idle processing resources among the demanding nodes [15].

In the economic model the center point is the interaction between grid users and providers. While most market models have been based on auctions, commodity market models have also been interesting research topic. Markets are considered to be based on commodity where applications can treat computational and storage resources as interchangeable and not as specific machines and disk systems. Obviously, prices are a key element of this model [16],[17],[18].

An alternative approach to market based and economy based model is finance based model. Various grid resources such as memory, storage, software, and compute cycles are seen as individual commodities and pricing of the resources is done in isolation and in combination of various resources. A quality of service (QoS)-profit equilibrium model has been proposed for pricing grid resources that is based on finance concepts [19].

Two shortcomings in a grid economic environment has been identified in [20]. The first shortcoming is that there are no standards for pricing schemes, caused by a large difference in the units that are traded (e.g. CPU cycles or virtual clusters) in grid computing. The second shortcoming is the lack of a model for managing the pricing of intangible elements (e.g. software applications) and computational elements (e.g. virtual machines, which comprise resources such as CPU, memory, disk space, network bandwidth).

B. A pricing scheme adapted to the particular nature of grid services

This paper further presents a pricing service for grid computing services, which resolves the shortcomings by introducing a general pricing scheme for informational and computational elements. We describe the functional requirements, architecture, and the interfaces of the pricing service. The pricing service allows expressing the proposed general pricing scheme as an XML document, which can be linked to Service Level Agreements (SLA). Contrary to other proposals on pricing, the pricing service is separated from the functionality of metering, accounting, and payment.

Various kinds of solutions to grid resource discovery have been suggested, including centralized and hierarchical information server approaches. However, both of these approaches have serious limitations in regard to scalability, fault tolerance, and network congestion. To overcome these limitations, indexing resource information using a decentralized, for instance peer to peer, network model has been actively proposed in the past few years.[21]

A new infrastructure called Grid Bank provides services for accounting of the grid resources thus filling the gap of this needed services. The support of computational economy and accounting services can lead to a self-regulated accountability in grid computing. This paper presents requirements of grid accounting and different economic models within which it can operate and proposes a Grid Accounting Services Architecture to meet them.[22]

Practically, we intend to focus on a particular salient attribute of typical grid services which represents in our case an intangible element of perceived value by the user. This makes our approach original compared to previous papers on grid pricing. The salient attribute developed in our model corresponds to the idea that priorities of jobs executed on the grid are most of the time satisfied. Indeed services provide users with benefits that are perceived with more or less value. Consequently, we assume that the management of priorities in the execution of jobs is perceived as an important element of value. In logistic terms, the quality of this management corresponds in the model to our service level.

To be able to obtain a pricing strategy leading to the best priority management of jobs executed, we have developed a mathematical programming model. It aims at minimizing inconveniences due to mismanagement of priorities while taking into account grid capacity constraints. As probabilities of a given job being put "on hold" exist, we treat capacity constraints of the grid as MDPs. In the following section, we present the detailed model that we have developed for illustrative purposes.

IV. THE MODEL

A. Description of the model

The grid is shared by $U$ different users described by the variable $u = 1, \ldots, U$. The jobs are classified in $J$ different categories according to the level of urgency. Categories are described by the variable $j = 1, \ldots, J$. Category $j = 1$ includes very urgent jobs and category $j = J$ includes jobs that are not urgent at all. The grid can process jobs with $P$ different priority levels. The highest priority is $p = 1$ whereas the lowest priority is $p = P$.

For each user, jobs arrive randomly with exponential distribution. Let $\alpha_{ju}$ be the parameter of the distribution for the job's category $j$ and for user $u$. The job's size is random and follows also an exponential distribution. Let $\beta_{ju}$ be the parameter of the distribution for the job's category $j$ and for user $u$.

Each user is rewarded at each period with a certain amount of monetary units that permits him to pay the services provided by the grid. Let $\mu_{u}$ be the amount received at each period by user $u$. This repartition should reflect the
contribution of user \( u \) to the grid (i.e. number of computers, it specialist, etc... offered by user \( u \) to the grid’s community).

The services provided by the grid are charged according to the CPU time used and priority level \( p \) requested. Let \( \pi_p \) be the price per period for jobs with priority \( p \).

The objective of the grid manager is to provide the best service to the users with the actual capacity of the grid. The service quality is measured by a penalty function, the lowest the function the highest the quality. This function increases each time a job is not completed on time. This function can, for example, measure the percentage of jobs done with a delay. The grid manager has to find the optimal prices \( \pi_p \) in order to have the lowest penalty function.

The objective of the user is to minimize his own penalty function choosing the right priority for each job. They are two possible actions for a user. First, when a new job arrives he has to decide which level of priority will be chosen. Secondly, when a job is in the queue, he can decide to change the level of priority of the job. Of course each decision is taken knowing the load of work of the grid. The set of all possible actions is denoted with \( A \).

**B. The model’s equations**

This model is described by a Markov Decision Process (MDP) where a budget constraints are added. The only possibility to solve this enriched MDP is to use the linear programming approach.

For each user \( u \), let \( x(u,j,p) \) be the number of jobs of category \( j \) with priority \( p \) that are currently in the system. Denote with \( S = \{ x(u,j,p) \} \) the possible states of the system.

Let \( q(s,\tilde{s},a) \) be the generator of the MDP. For the cases \( s \neq \tilde{s} \), \( q(s,\tilde{s},a) \) is computed from the corresponding \( \alpha_{ju} \) and \( \beta_{ju} \). For the other cases, we have \( q(s,\tilde{s},a) = -\sum_{s,j} q(s,\tilde{s},a) \). It is important to realize that the main difficulty in modeling a MDP is to compute its generator. For example, the capacity constraints of the grid are included in the generator. Modeling capacity constraint in a MDP model is less intuitive than modeling capacity constraints in a linear programming model.

Let \( \Pi(s,u) \) be the price charged to user \( u \) for his jobs that are currently treated by the grid. It depends of course on the prices \( \pi_p \).

The linear program associated with this enriched MDP is

\[
\min \sum_{s,a} C(s,a) \cdot Y(s,a)
\]

subject to

\[
\sum_{s,a} q(s,\tilde{s},a) \cdot Y(s,a) = 0 \quad \forall \tilde{s} \in S.
\]

\[
\sum_{s,j,a} q(s,\tilde{s},a) \cdot Y(s,a) = 1.
\]

\[
\sum_{s,a} \Pi(s,u) \cdot Y(s,a) \leq \mu_u \quad u = 1, \ldots, U.
\]

In this model, \( Y(s,a) \) are the steady state probabilities. The first equations represent the flow constraints and the second equation is the normalization of the probabilities. The last inequalities represent the budget constraints for each user.

As \( S \) has to be finite, we impose reflexion conditions on the boundaries. The interested readers can find a detailed description of the modeling methods offered by the MDP paradigm in [1] and [2].

If the prices \( \pi_p \) are taken as a variable, the model is a quadratic program with a non positive definite matrix. In this case, standard softwares cannot solve the model. To avoid this problem, we consider the prices as parameters. In this case, the model is a linear program that is easily solved if not too big. As the model is not convex in \( \pi_p \), we use a grid method in order to find the global optimum.

**V. NUMERICAL ILLUSTRATION**

The aim of the present numerical results is to show the efficiency of the method. The model used is very small and is not describing accurately our real problem of grid resources allocation. We are presently starting to collect data and will be soon able to present a realistic model describing accurately our real life problem.

In this small model, we assume to have two categories of jobs: urgent and not urgent ones. We have two kind of priorities: high and low priority. We suppose that all different jobs need the same number of CPUs to be run. The difference between big and small jobs is therefore only the CPU time requested. This assumption is valid if all jobs are parallelizable in the same way and are not infinitely parallelizable. The maximum number of jobs that the grid can process simultaneously without congestion is 3. In case of congestion, jobs with high priority are processed first. In case of congestion, only 3 jobs can be processed and if there are more than 3 jobs with high priority, the last coming jobs are send in the queue. Doing so, jobs are always processed at full speed or send in the queue.

In our real life problem, we have 20 different universities or faculties. For this numerical illustration, we assume to have two different kinds of users. The first kind of users expect a lot of urgent jobs that are on average small whereas the second users expect less jobs that are on average big. We suppose that half universities are from the first kind of users whereas the rest are from the second kind.

In this numerical experiments the objective is to maximize the service level, which is the percentage of jobs that are done without delay.

Table I indicates the main parameters employed in the instance developed for the model presented in the previous section. Table II presents the optimal solution found for different prices. Prices are given in monetary units per period. We ran the model for more prices but show only few interesting results.

We see distinctly that the price policy has an effect on the service level. In this example, the maximal service level can be attained with several price policies. More than the results, this small model show the applicability of this method.
TABLE I

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha_{11})</td>
<td>0.02</td>
</tr>
<tr>
<td>(\alpha_{21})</td>
<td>0.05</td>
</tr>
<tr>
<td>(\alpha_{12})</td>
<td>0.2</td>
</tr>
<tr>
<td>(\alpha_{22})</td>
<td>0.1</td>
</tr>
<tr>
<td>(\beta_{11})</td>
<td>0.005</td>
</tr>
<tr>
<td>(\beta_{21})</td>
<td>0.005</td>
</tr>
<tr>
<td>(\beta_{12})</td>
<td>0.2</td>
</tr>
<tr>
<td>(\beta_{22})</td>
<td>0.1</td>
</tr>
<tr>
<td>(\mu_1)</td>
<td>1</td>
</tr>
<tr>
<td>(\mu_2)</td>
<td>1</td>
</tr>
</tbody>
</table>

TABLE II

<table>
<thead>
<tr>
<th>Price for high priority jobs</th>
<th>Price for low priority jobs</th>
<th>Jobs on time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>97.8%</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>97.8%</td>
</tr>
<tr>
<td>1.2</td>
<td>0.2</td>
<td>91.8%</td>
</tr>
<tr>
<td>1.2</td>
<td>0.4</td>
<td>90.2%</td>
</tr>
<tr>
<td>1.4</td>
<td>0.6</td>
<td>81.9%</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>80%</td>
</tr>
</tbody>
</table>

However, we must emphasize that MDP are subject to the "curse of dimensionality". For bigger models, the state's set \(S\) can become too big. In this case it is necessary to reduce the size of the model, applying, for instance, decomposition or parallel processing method [23].

VI. CONCLUSION AND FUTURE RESEARCH

We propose a new approach to share optimally the resources of a grid based on a pricing scheme that aims at maximizing intangible factors of perceived value by its users. We first explain the particular nature of grid services and main issues regarding their pricing. Then, we present the main findings of market resource based management adapted to the context of shared computing grids. Finally, we develop a new model to price grid services.

In this model, we retain for simplicity a single objective which corresponds to the respect of priorities set for job executions. We thus assume that priority management is perceived by individual users as a prominent benefit provided by grid services. In order to take into account stochastic priorities, we include in the mathematical programming problem MDP logistic constraints. Indeed these latter constraints make the assumption that a job can switch among different priority states. For illustrative purposes, we present a simplified instance of the model. We are in the process of adapting it for a real size grid.

This research is thus still in its infancy. As future work, we intend to develop model instances based on real data. Indeed, this research is part of a grid project named EZ-Grid and Interreg that reunites several universities from Switzerland and France.

In terms of modeling, we also intend to explore several research directions. First, in the present model, we aim to obtain an optimum for a single user. We might also adopt a broader view and instead provide an equilibrium for the whole grid. That would mean that we must include in the model a multiple user structure. This can be done with the integration of game theoretic equations in the model. A first case is when the grid actors are competing without cooperation. This case calls for Non-Cooperative Game Theory and corresponds to a strict competitive market. In this particular case, the mathematical model falls into the category of Nash-Cournot equilibria. Although quite complex, solution processes exist. The model used in this second approach is very similar to the previous model, except that the integer program is replaced by a variational inequality.

The second case corresponds to the situation where actors in the grid are competing but cooperation is possible. It then calls for Game and Negotiation Theories. In most grid applications, interactions among players are in fact cooperative. Indeed, non-profit collaborations are often set up by universities whose main goal is to develop new knowledge regarding grids. So we will also investigate the cooperative game scheme. Assuming that players are entering a cooperative game, the mathematical framework and solution process is completely different. Moreover, regarding the state of the art, we consider that techniques still need to be developed to properly fit the case.

To sum up, we propose an innovative method to tackle the problematic of pricing for the grid and will implement them on the platform that will be set up in the framework of this project.

VII. ACKNOWLEDGMENTS

This work has been sponsored by the project Virtual EZ-Grid from the swiss national fund AAA/SWITCH.

REFERENCES


