Research Report

How Prices can be set to allocate Grid Computing Resources in a Financial Service Institution

GRID COMPUTING IS AN IT CONCEPT TO SHARE COMPUTING RESOURCES AMONG DEPARTMENTS AND USERS THAT REDUCES IT COSTS AND PROVIDES COMPUTING RESOURCES DYNAMICALLY WHEN THEY ARE NEEDED. RESOURCE MARKETS ARE AN EFFECTIVE MECHANISM TO REGULATE THE RESOURCE SHARING, BUT THE MOST OFTEN USED AUCTIONS ARE COMPLEX. WE HAVE DEVELOPED A STEPWISE APPROACH TO HELP FIRMS OFFERING INTERNAL GRID COMPUTING SERVICES TO SET TRANSPARENT BUT EFFECTIVE PAY-PER-USE PRICING SCHEMES AS AN ALTERNATIVE TO AUCTIONS.

Markus Lilienthal

Oliver Hinz

Introduction

Grid computing allows the sharing of resources such as processing power, storage, memory, and other services. By connecting many computing resources (not necessarily only in one data center), the Grid becomes a virtual supercomputer, which allows for a better utilization of otherwise idle resources.

The way how the resources are connected is highly standardized similar to standards for the Internet, whereas the resources themselves (also similar to the Internet) can be arbitrarily diversified. According to Information Systems literature, Grids are supposed to reduce IT costs drastically, thus contributing to Green IT developments, and offer a much more dynamic way to deliver IT resources wherever they are needed (Foster and Kesselman, 2003).

Some see the future of Grid Computing – very similar to the Internet – as one globally connected Grid of millions of computers. However, for some enterprises, in particular in the financial services industry, outsourcing of computing is not an option due to privacy concerns or provision by law. To take advantage of the capabilities of Grids, such enterprises therefore install in-house Grid computing solutions.

For example, Wachovia, the fourth-largest bank in the United States based on total assets, already deployed a Grid thus allowing applications such as parallelized portfolio evaluation to draw computing power from a pool of 10,000 processors spread across numerous cities in the United States and the United Kingdom. The potential benefits are immense, considering that in 2010, North American banks have spent more than $56 billion on IT (Jegher, 2011), cost savings of even a few percent already account for billions of dollars.

Pricing of Grid Resources

Whenever a commodity is shared by many, a mechanism that matches and regulates demand and supply becomes necessary. For more traditional, internal IT resources, probably the most common mechanisms are either to define fixed allowances for all participants, or direct cost allocation, where the departments are charged the average per unit cost.

Both mechanisms are not optimal, even for more traditional IT resources. A market based on real or virtual money could provide the needed flexibility, as the market participants may decide for themselves when they want to consume what type of resource.

Auctions are known as the most effective pricing mechanisms in these settings. Their ability to regulate demand and supply dynamically is extremely high. Therefore, auctions have been extensively considered as a means to allocate Grid resources. However, auctions also have shortcomings. Most of all, auctions are complex and planning reliability is limited. Grid research suggests the use of automated brokers, but complexity and deficits in planning reliability still remain to some extent.

In contrast to all these, the most common pricing schemes we observe in everyday life are posted, non-dynamic tariffs, such as flat fees or pay-per-use prices. The main reason for their appeal and popularity is their simplicity. They are easy to understand and reliable, however, not as efficient as auctions.

With our research, we demonstrate how IT departments can set pay-per-use prices for Grid resources that differentiate users by their needs without the complexity and unreliability of auctions.

Our New Approach of Pricing Computing Resources

We have developed a five-steps approach to determine such pay-per-use pricing schemes for Grid Computing resources (Figure 1). The aim of the scheme is the differentiation of users with respect to their performance needs while complexity is kept low.

Figure 1: Our five-steps approach

We divide users and resources into user segments and resource classes. The aim of the procedure is to obtain a tariff for each user segment (respectively resource class), whereby each user can freely decide which tariff to choose.

The idea behind the segmentation is a strategy of repurposing resources over their lifetime. Throughout the lifetime of a resource, the computational power decreases in relation to the market standard. In a traditional setting with dedicated resources, the resources are either over- or underutilized during most of their lifetime. In a Grid environment, resources can be simply reassigned stepwise to lower classes before they are eventually written off. The system enables all resource consumers to always choose the resource class that fits their requirements.

The five-steps (Figure 1) are as follows.

Step 1: Measuring Willingness-to-Pay
In a first step, we determine the individual willingness-to-pay (WTP) for each consumer at certain speed levels (at least four). WTP is the maximum amount of money the user is willing to pay for the service. Based on these price preferences at discrete points, we can estimate a continuous WTP function.

Step 2: Identifying user segments
After the determination of the WTP the next step is to identify the different user segments. To identify the segments and their number we apply statistical clustering, a method that assigns each respondent to one cluster.

Step 3: Defining resource classes
Resources are classified such that the size of the resource class corresponds to the size of the user segments.

Step 4: Averaging within segments
Having determined the segments, the next step is to compute the WTP by segment as an expected value.

Step 5: Finding tariffs
In the last step, we determine optimal tariffs. There are as many tariffs as user segments or resource classes (one-to-one relationships between tariff, resource class and user segment). Because the consumers may choose freely, the challenge is to determine tariffs such that they choose their supposed tariff. Users always choose the tariff where the difference between WTP and their costs is maximized. This behavior is known as utility maximization. The tariffs should furthermore cover the costs (not necessarily in every segment, but in total).

Empirical Study

As a proof-of-concept, we conducted a survey study in a large European bank that is planning to switch from dedicated servers for single business units to an enterprise Grid. The empirical study follows the five-steps. The sample comprises 21 project leaders and business unit heads with their own budget responsibility.

The CTO office identified about 80 leading employees in Great Britain, Germany and Singapore who held suitable positions with budget responsibilities for IT. Out of this population, 21 business unit managers and project leaders with a budget responsibility for IT resources were willing to participate. The CTO office evaluated this subsample as representative and the results of the study as meaningful.

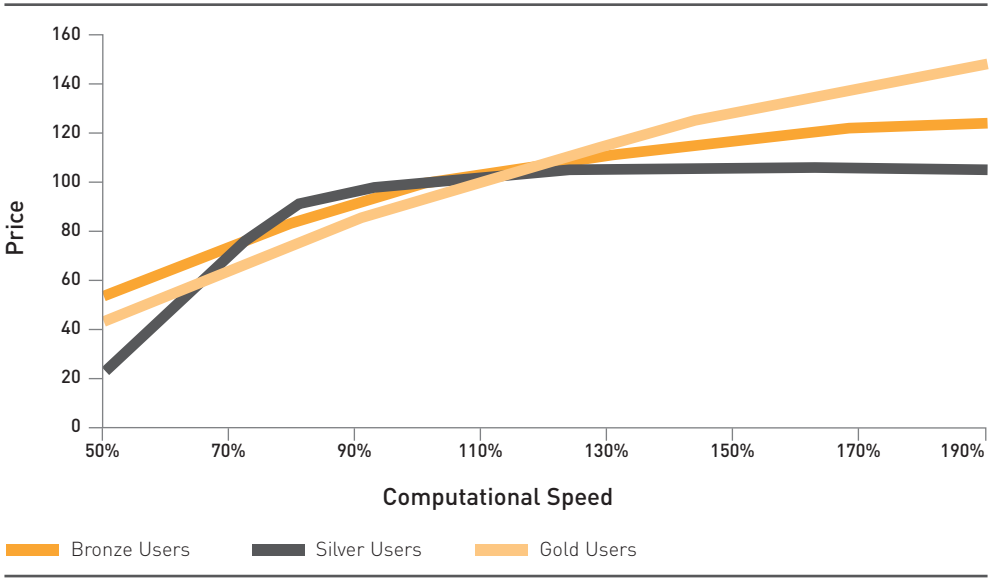

Figure 2: The three user segments in the empirical study

Results

The result of the clustering procedure is displayed in Figure 2. We identified three robust user segments in our empirical study, which we name Gold users, Silver users and Bronze users. Gold users are characterized by a high WTP for high speed resources. Silver users are neither willing to pay for greater computational power nor are they willing to switch to slower resources for a small discount. Bronze users are willing to switch to slower server classes if they receive a sufficient discount, but are also willing to pay moderately higher prices for additional power. We observe that the Gold users are willing to pay much less than double price for double speed, thus provision of those resources would likely be unprofitable from a purely economic point of view and given dedicated computing resources for this group. However, in a Grid system with our pricing scheme, it will be possible to provide access to such resources anyway.

The obtained prices are shown in Figure 3. Based on our survey results, it is possible to assign prices to each resource class such that Gold users prefer Gold resources, Silver users prefer Silver resources and Bronze users prefer Bronze resources.

	Average Power (in class)	WTP Gold Users	WTP Silver Users	WTP Bronze Users	Optimal Pay-per-Use Price
Gold Server	160%	132%	104%	118%	132%
Silver Server	101%	93%	99%	97%	99%
Bronze Server	64%	57%	56%	68%	68%

Figure 3: The suggested tariffs

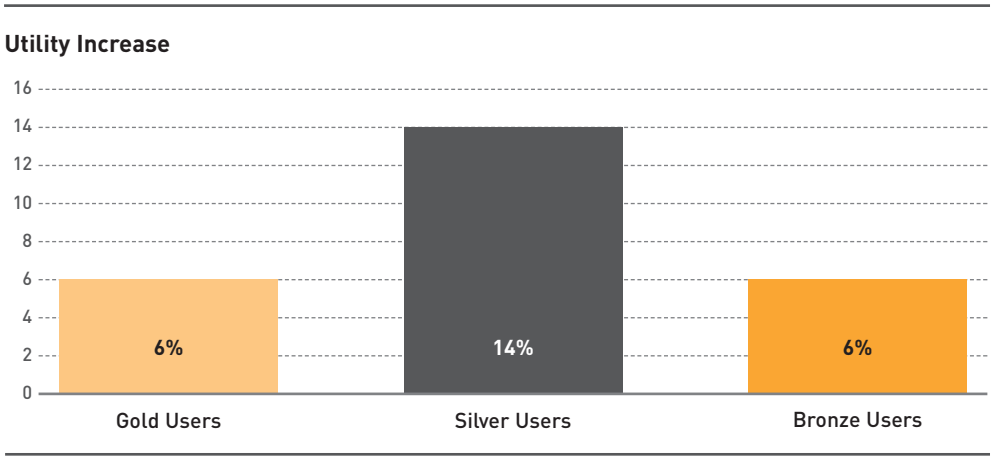

Figure 4: Utility benefit keeping costs fixed

Benefits of the Tariffs

In our study we analyze two benefit scenarios: a cost-neutral scenario and a power-neutral scenario. That means, savings are either used to increase utility while keeping costs fixed, or to reduce costs by delivering the same computational performance as before. The reference scenario for both cases is dedicated resources with direct cost allocation.

We conclude that, by introducing tariffs, the enterprise can either increase the utility by 9% on average (see Figure 4 for details by user segment) while the costs are kept unchanged, or save 7% of the costs without losing any computational power when compared to the benchmark of dedicated resources.

From a managerial perspective, it might also be appealing to apply a mixture of both, such that the cost reduction comes along with an increased performance, which is easier to communicate. Depending on the examined scenario, the entire enterprise will observe either increased utility or reduced costs.

Having only three different price levels, the suggested tariff structure is very simple and easy to understand. It allows the IT department to repurpose hardware continuously: The IT department purchases new servers initially for the Gold segment and later reassigns them to the Silver segment. At the end of the servers' lifetime, the IT department assigns the servers to the Bronze segment.

Conclusion

The Grid technology itself helps to realize untapped cost-saving opportunities that result from idle resources. Moreover, the management of Grids may provide another opportunity, because it will allow enterprises to move from flat fees that cover total costs for dedicated servers to pay-per-use tariffs. Additionally, our method helps IT management to post prices such that incentives are set to move jobs to repurposed, slower servers.

We found that all our participants (internationally dispersed business unit managers and project leaders) are willing to shift jobs to slower servers if incentives in form of lower prices are set accordingly.

However, not all consumers of IT resources are alike. We clearly identified three different segments in our proof-of-concept study. One segment, which we call Gold, had a very high willingness to pay for high-performance computing. The second segment, Silver, derived the most utility from standard servers, whereas the third segment, which we call Bronze, was willing to shift its jobs to slower servers if the IT management was willing to give a discount of about 32% compared to the Silver servers.

The overall costs for the enterprise can be reduced by 7% by repurposing older servers, or the utility to the business units can be increased by 9% at stable costs.

We pinpoint the advantages of our pricing scheme:

- We achieve a fair pricing where all users (business units) can individually decide how to spend their budget,
- Business units can rely on easy to predict expenses,
- The scheme increases utility and/or cuts costs,
- IT investments have a clear life-time cycle and are neither over- nor underutilized.

References

Foster, I.; Kesselman, C.:
The Grid 2. Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco 2003 (2nd ed.).

Jegher, J.:
IT Spending in Banking: A North American Perspective. Celent market research report, January 10th, 2011.