Extended Resource Management Using Client Classification and Economic Enhancements

Tim PÜSCHEL[1,2], Nikolay BORISSOV[1], Dirk NEUMANN[1],
Mario MACÍAS[2], Jordi GUITART[2], Jordi TORRES[2]
[1]*Institute of Information Systems and Management (IISM)*
Universität Karlsruhe (TH), Englerstr. 14, 76131 Karlsruhe, Germany
Tel: +49 721 608 - 83 70, Fax: +49 721 608 - 83 99,
Email: {pueschel, borissov, neumann}@iism.uni-karlsruhe.de
[2]*Barcelona Supercomputing Center - Technical University of Catalonia (UPC),*
c/ Jordi Girona, 29, Barcelona, 08034, Spain
Tel: +34 93 413 40 49, Fax: +34 93 413 77 21,
Email: {tim.pueschel, mario.macias, jordi.guitart, jordi.torres}@bsc.es

Abstract: Commercialization of Grid resources will become more and more important as utility computing and the deployment of Grids gains momentum. This results in the necessity to not only base Grid components on technical aspects, but also to include economical aspects in their design. This paper presents a framework that links technical and economical aspects to the management of computational resources. Economic enhancements like dynamic pricing and client classification are introduced based on a technical resource management environment and positioned within this resulting in a proposed architecture for an Economically Enhanced Resource Manager (EERM). The introduced approach is evaluated considering various economic design criteria and example scenarios.

1. Introduction

In many cases IT applications – such as data mining, portfolio analysis and video stream analysis – are characterized by the fact that they have a strongly varying demand for resources like processors and storage. To accommodate peak load times, it is necessary to maintain an adequate IT-infrastructure. During off-peak times these resources mainly remain idle. With increasing global competition, enterprises are forced to cut down costs dramatically and therefore strive for trimming down costs for IT infrastructure [3].

This need gave rise to the vision of utility computing [15] where computer resources can be accessed dynamically in analogy to electricity and water. Utility computing becomes more powerful if more resource providers add their resources to the Grid. It is thus the utmost objective to attract more providers.

With state-of-the-art technology, this assimilation into the Grid is hampered, as the local resource managers facilitating the deployment to the Grid resources are not designed to incorporate economic issues (e.g. price). They plainly adhere to technical parameters that define when jobs are scheduled. In the easiest case, the local resource managers employ a First-in-First-out algorithm for scheduling – ignoring all economic factors.

In recent times, several research projects have started to develop price-based resource management components that support the idea of utility computing. Those approaches are entirely devoted to scheduling by utilizing the price mechanism.

In addition, resource management is much more comprehensive than just scheduling. For example Service-Level-Agreement (SLA) management is also part of resource management that is oftentimes omitted in economic approaches. This plays for instance a

role when deciding which already ongoing jobs to cancel in overload situations in order to maintain system stability. To achieve better performance in the commercialization of distributed computational resources, decisions about the supplied resources and their management therefore should be based on both on a technical and on an economic perspective [9].

Technical resource management systems typically offer the possibility to include priorities for user groups. In purely price-based schedulers it is not possible to distinguish important from unimportant partners, as only current prices matters for the allocation.

2. Objectives

The objectives of this work are to motivate and introduce economical enhancements to resource management and present an architecture comprising these enhancements. We will motivate that client classification should be integrated into economically enhanced resource management systems. Essentially, there are two main reasons to do so: First, client classification allows the inclusion of long-term oriented relationships with strategically important customers. Second, client classification can be used as an instrument of revenue management, which allows skimming off consumer surplus. The main contribution of this paper is to show how technical parameters can be combined into an economically enhanced resources management that increases revenue for the local resource sites.

The remainder of the paper is structured as follows: Section 3 presents a motivational scenario, section 4 related work. Section 5 explains the economic enhancements and the mechanism of client classification. Section 6 gives an overview of the goals and the architecture of the EERM. Subsequently section 7 contains an example scenario and a short evaluation of the proposed mechanisms. Finally, section 8 concludes the paper and describes our future work.

3. Motivational Scenario

Suppose a large service provider maintains an IT-Infrastructure, whose free resources are offered to external users. The service provider already has a number of clients but depending on the time specific resource usage, it has fluctuating spare capacity. Therefore, the service provider offers its excessive resources over a Grid market to find new clients and optimize the resource utilization and thus its revenue. To retain the good relations with current clients and encourage regular use of its services, the provider offers special-agreements to preferred client. Preferred clients (“Gold-clients”) receive preferences when submitting jobs and obtain reservation on certain share of the provider’s resources.

For clients who do not want to entrust their data to just any arbitrary Grid service provider, this scenario is a very interesting option. They can have preferred client contracts with a few selected Grid service providers. This increases their chances to access the services of a trusted provider on demand. Preferred clients are also awarded with further benefits, such as better service levels or price discounts, from the providers as additional incentives to use their resources.

At the same time the proposed mechanisms are not exclusive, so they leave the option to use other Grid resources of other providers in case of missing free capacity.

4. Related Work

Requirements on quality of service (QoS) functionalities by the management of disperse computational resource are considered into the *Globus Architecture for Reservation and Allocation* (GARA) [6]. The incorporated components are a first step to achieving end-to-end QoS guarantees by introducing advanced reservation of resources.

QoS can also be achieved by introducing risk management into the Grid, which is elaborated in [5]. Their proposal allows modeling and managing the risk that the service level agreement (SLA) cannot be fulfilled. This allows taking the risk of SLA failure into account when deciding on prices and penalties.

Another approach for autonomic quality of service aware resource management is based on online performance models [10]. The authors introduce a framework for designing resource managers able to predict the impact of a job in the Grid performance and adapt resource allocation in such a way that service level agreements (SLAs) can be fulfilled.

Elements of client classification in Grids such as price discrimination, based on customer characteristics, are explored in [1] and [11]. Chicco et al. describe data- mining algorithms and tools for client classification in the electricity Grids [4] but concentrate on methods for finding groups of customers with similar behaviors.

Further related work in client classification includes a framework for admission control on e-commerce websites that prioritizes user sessions based on predictions about the user's intentions to buy a product [14].

5. Economic Enhancements and Client Classification

One key requirement of commercial resource managers is to offer QoS regarding the job execution, such as guarantees about the available resources and execution time. The first objective of an economic enhanced resource manager is to maximize its revenue, e.g. by allocating as many jobs as possible. However overload situations can lead to reduced overall performance [12] and break QoS agreements between the provider and clients.

To avoid this, the resource manager needs information about the current utilization of the offered resources as well as information about the required resource capacity of the incoming jobs. While the information about the current utilization could be delivered from monitoring services, the job's execution time and thus its required capacity according to the job's execution-deadline is difficult to estimate. Kounev et al. [10] propose a mechanism to estimate the influence of a job's execution on the utilization through online performance models. The agreed QoS also should be met when some of the computational resources failure. This requires to keep an adequate percentage of the offered resources free, so failure situations could be handled transparent to the client and theirs job's QoS still met. Where such a buffer is not possible or desired it is crucial to at least meet as many SLAs as possible and thereby minimize the negative effect of failure situations.

To this end we propose the introduction of a Job Cancellation and Suspension mechanism. Where predictions are inaccurate or problems in the Grid lead to reduced capacity some jobs are cancelled or suspended to ensure other jobs meet their SLA. Deciding which jobs to cancel should ensure that overall revenue is maximized, i.e. the cancellation penalties and the loss of revenue due to cancelled jobs are minimized.

As another enhancement we introduce dynamic pricing based on technical as well as economic factors. Resource prices can be based on the current utilization of the Grid, the impact a job has on the utilization of the Grid, client classification as well as other factors such as current supply and demand. For example when the resource utilization is very high or an incoming job leads to a high utilization a higher price is charged.

A client is a user or an application which is willing to allocate and consume distributed computational resources. The following factors can be used to differentiate client classes:

- *Price discrimination*: is an economic factor proposed in the past by different authors. One example is the idea of using Grid miles [1] [11], in analogy to frequent flyer miles. In general certain clients can be given price discounts on reservation or final prices.
- *Job prioritization* is another option to differentiate clients. We differentiate two types of job-priorities – strict and soft priorities.

- *Strict priority* means that jobs from clients with priority are always preferred over clients without priority. Thus there is no real competition between the different classes of clients.
- *Soft priority* means jobs from clients with priority are generally preferred but clients without priority have the chance to outbid prioritized clients.

- *Reservation of resources* is important for clients who want to ensure that they always have a certain capacity at their disposal.
- The last introduced discrimination factor is *quality of service (QoS),* which results in versioning of the service as known from pricing theory [16]. Offering different levels of QoS for different classes of clients is possible by modeling them in the SLA.

Based on the described desired properties of the resource manager as well as the abovementioned client classification factors we introduce a framework of economic enhanced resource management.

6. Economically Enhanced Resource Management

Beside the specified economic enhancements regarding the client classification the EERM-mechanism has to satisfy common economic design criteria explained in the first subsection. To allow the integration of client classification, associated economic enhancements, as well as the economic design criteria we propose and describe a framework of Economically Enhanced Resource Manager (EERM).

6.1 Economic Design Criteria

The EERM has to satisfy following economic design criteria proposed in [2] and [17]:

- *Individual rationality*: The provider must benefit from using the EERM and the client should benefit from choosing a provider using the EERM. For the provider this benefit could be a higher or more predictable revenue, lower risk (e.g. of paying penalties) and better client retention. For the client this benefit could be a higher ratio of acceptance of important jobs, lower prices, better service levels or preferred acceptance of jobs.
- For the criterion of *incentive compatibility* it is important to choose the characteristics of the mechanism in such a way that the clients report their true requirements. This avoids strategic behavior, e.g. with the aim to influence the client classification.
- *Revenue maximization*: The objective of the resource providers is to maximize their revenue, which is one of the economic characteristics of the EERM.
- Client Classification adds some additional *computational complexity*. Depending on the policies that are chosen winner determination has to be slightly adapted. It, however, does not introduce any NP-hard problems into the mechanism and the additional computational cost should be limited.
- Another criterion is *efficiency*. A mechanism is called allocative efficient if it maximizes the sum of individual utilities.

6.2 Model of the EERM

The main goals of the EERM are to:

- link technical and economical aspects of resource management
- establish more precise price calculations for resources, taking usage of the Grid, performance estimations and business policies into account and
- strengthen the economic feasibility of the Grid.

In this work we focus on presenting the features of the EERM that can be related to client classification.

Figure 1 shows the EERM Architecture. The *Economy Agent* first receives a request from a market agent, checks whether the job is technically and economically feasible and

calculates a price for the job based on the client's class, resource availability, pricing policies as well as predictions of future job executions from the estimator component.

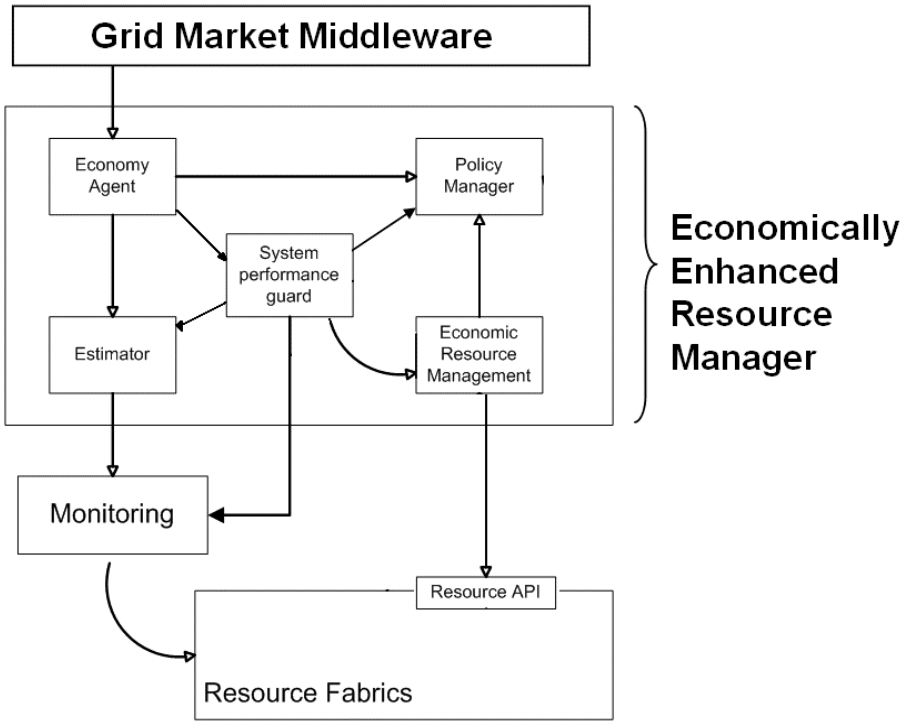

Figure 1: EERM Architecture

The *Estimator* component calculates the expected impact on the utilization of the Grid.

The *System Performance Guard* (SPG) monitors the performance of the provider's supplied resources and ensures the accepted SLAs. If there is a risk that one or more SLAs cannot be fulfilled the SPG can take decisions to suspend or cancel jobs to ensure the fulfillment of the SLAs and maximize overall revenue.

To keep the EERM adaptable, the *Policy manager* stores and manages policies for client classification, job cancellation and suspension. Policies are formulated in Semantic Web Rule Language (SWRL) [8]. An example for a policy based rule is like the following:

$$ClientClass(?clientclass) \wedge sameAs(?clientclass, "Standard") \wedge$$

$$Utilization(?utilization) \wedge InsideUtilizationRange(?utilization, "70\% - 100\%")$$

$$\Rightarrow RejectJob$$

This policy express that if the utilization of the Grid is between 70% and 100% and the client classification of a job is "Standard" the job is not accepted. This implies that in this case only jobs with other classifications are accepted.

The *Economic Resource Manager* is responsible for the communication with the local resource managers and influences the local resource management to achieve a more efficient global resource usage.

The EERM interacts with various other components, including the Grid Market Middleware, a Monitoring component and Resource Fabrics. The Grid Market Middleware represents the middleware responsible for querying prices and offering of resources.

The *Monitoring* is responsible for monitoring the state and the performance of the Grid.

Resource Fabrics enables the low level access to the Grid-resources, e.g. via Condor [13] or Globus [7].

Figure 2 shows the sequence of a Job Cancellation due to Grid performance problems. The Monitoring informs the System Performance Guard about problems on the grid. The

SPG proofs policies, requests necessary information, chooses the jobs to be cancelled and interacts finally with the Economic RM to initiate the cancellation of the respective jobs.

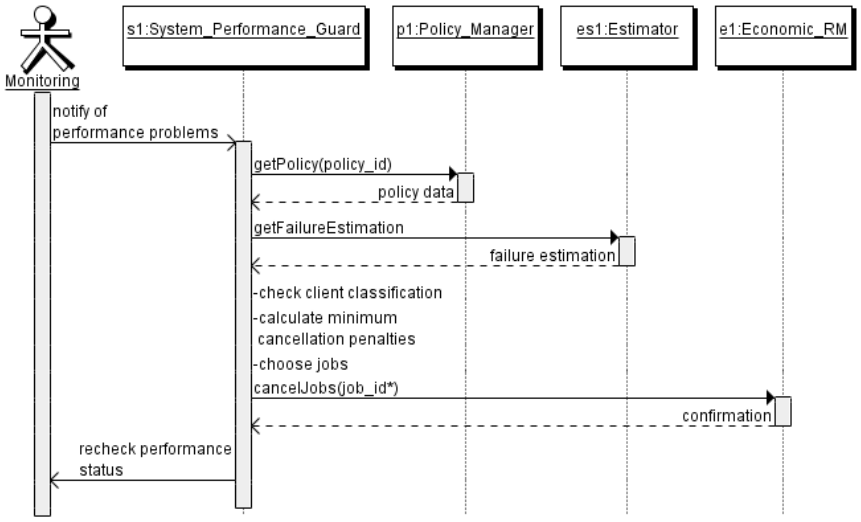

Figure 2: Sequence Diagram of Job Cancellation

7. Evaluation

For the example scenario we have several assumptions:

- We assume that the system only receives information about jobs that become available in the following time period.
- A Gold-client only uses the provider if he can launch jobs with capacity requirements between 30 units and 60 units per period.
- The total capacity of the provider is 100.
- The jobs shown in Table 1 will be available during the run.

First we consider a scenario without EERM (Case I). In this case any job is accepted if there is enough capacity left to fulfill it. As can be seen in Table 2 Jobs A, B, C, G, H, L, M, O, P, and Q are accepted, the other jobs cannot be accepted due to capacity constraints. This results in total revenue of 1349 and an average utilization of 89.7.

Table 1: Example Data with Available Jobs During the Run

Jobs:	Start	End	Capacity/t	Client Class	Price
A	1	3	55	Standard	330
B	1	5	24	Standard	180
C	1	7	20	Standard	140
D	2	4	20	Standard	120
G	4	7	20	Standard	160
H	4	9	15	Standard	135
L	6	8	30	Standard	90
M	6	9	12,5	Standard	50
O	8	10	20	Standard	90
P	9	10	21	Standard	84
Q	9	10	30	Standard	90
R	2	6	30	Gold	375
S	5	8	30	Gold	300
T	7	10	7,5	Gold	75
U	7	9	20	Gold	150
V	9	10	20	Gold	100

`

Then we introduce a fixed reservation of 60% of resources for the Gold-client to ensure his requirements are fulfilled (Case II). Now total revenue is 1400 and average utilization of 71.2. Even though the utilization is lower an increase in revenue can be achieved.

Table 2: Allocation Example Data with Available Jobs During the Run

Case	Jobs completed	Revenue	Avg Utilization
I	A, B, C, G, H, L, M, O, P, Q	1349	89.7
II	C, D, R, S, M, T, U, O, V	1400	71
III	A, R, G, H, S, T, U, V	1625	73.5

In the third case the policy is to accept only jobs from the Gold-client if the job would result in utilization higher than 70%. In case this is not sufficient to fulfill the requirements of the Gold-client, jobs from Standard-clients are stopped. This is a policy that can be used in situations as described in the motivational scenario. In this case it is not necessary to stop any jobs and the policy results in total revenue of 1625 and an average utilization of 73.5. This is a significant increase in revenue.

8. Conclusions

In this paper we motivated client classification and further economical enhancement for resource management. We presented factors and technical parameters that can be used for these enhancements to increase revenue for the local resource sites. Furthermore we introduced the preliminary architecture for an Economically Enhanced Resource Manager integrating these enhancements. Due to the general architecture and the use of policies and a policy manager our approach is can be adapted to a wide range of situations.

We evaluated our approach considering economic design criteria and using an example scenario. The evaluation of our first model shows that the proposed economic enhancements firstly enable maximizing provider's benefit and secondly strengthen the relationship with business clients.

The next steps will include refinement of the architecture as well as the implementation of the EERM. During and following this process, further evaluation of the system will be done, e.g. by testing the system and running simulations. Another issue that requires further consideration is the autonomous generation of business policies for the EERM.

Acknowledgments

This work is supported by the Ministry of Science and Technology of Spain and the European Union (FEDER funds) under contract TIN2004-07739-C02-01 and Commission of the European Communities under IST contract 034286 (SORMA).

References

[1] Buyya, R. *Economic-based distributed resource management and scheduling for Grid computing*. PhD thesis, Monash University, 2002.

[2] Campbell, D. E. *Resource Allocation Mechanisms*. Cambridge University Press, London, 1987.

[3] Carr, N. *The End of Corporate Computing*. MIT Sloan Management Review (46:3), pp 66-73, 2005.

[4] Chicco, G., Napoli, R. and F. Piglione. *Comparisons Among Clustering Techniques for Electricity Customer Classification.* IEEE Transactions on PowerSystems, 21(2): pp 933– 940, 2006.

[5] Djemame, K., I. Gourlay, J. Padgett, G. Birkenheuer, M. Hovestadt, O. Kao, and K. Voß. *Introducing risk management into the Grid.* In *The 2nd IEEE International Conference on e-Science and Grid Computing (eScience2006)*, page 28, Amsterdam, Netherlands, 2006.

[6] Foster, I., C. Kesselman, C. Lee, B. Lindell, K. Nahrstedt, and A. Roy. *A distributed resource management architecture that supports advance reservations and co-allocation.* In *Proceedings of the 7th International Workshop on Quality of Service (IWQoS 1999)*, pp 62–80, London, UK, 1999.

[7] Globus Project. http://www.globus.org [April 2007].

[8] Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, B. Grosof and M. Dean. *SWRL: A Semantic Web Rule Language Combining OWL and RuleML* - http://www.w3.org/Submission/SWRL/ 2004.
[9] Kenyon, C. and G. Cheliotis. *Grid resource commercialization: economic engineering and delivery scenarios.* Grid resource management: state of the art and future trends, pp. 465–478, 2004.
[10] Kounev, S., R. Nou, and J. Torres. *Using QPN to add QoS to Grid Middleware*. UPC Research Report: UPC-DAC-RR-CAP-2007-4, 2007.
[11] Newhouse, S., J. MacLaren, and K. Keahey (2004). *Trading Grid services within the uk e-science Grid.* Grid resource management: state of the art and future trends, pp. 479–490, 2004.
[12] Nou, R., F. Julià, and J. Torres (2007). *Should the Grid middleware look to self-managing capabilities?*. The 8th International Symposium on Autonomous Decentralized Systems (ISADS 2007). pp 113–122, Sedona, Arizona, USA, 2007.
[13] Litzkow, M., M. Livny, and M. Mutka. *Condor - A Hunter of Idle Workstations*. In 8th International Conference of Distributed Computing Systems (ICDCS), pp. 104-111, IEEE CS Press, Los Alamitos, CA, USA, 1988.
[14] Poggi, N., T. Moreno, J. Berral, R. Gavaldà, J. Torres. *Web Customer Modeling for Automated Session Prioritization on High Traffic Sites*. Proceedings of the 11th International Conference on User Modeling. Corfu, Greece, 2007.
[15] Rappa, M.A. *The utility business model and the future of computing services.* IBM SYSTEMS JOURNAL, VOL 43, NO 1, pp. 32-42, 2004.
[16] Varian, H. *Versioning Information Goods*, University of California, Berkeley, 1997.
[17] Wurman, P. R. *Market structure and multidimensional auction design for computational economies.* PhD thesis, University of Michigan, 1999.