ARE UTILITY, PRICE, AND SATISFACTION BASED RESOURCE ALLOCATION MODELS SUITABLE FOR LARGE-SCALE DISTRIBUTED SYSTEMS?

XIN BAI, LADISLAU BÖLÖNI, AND DAN C. MARINESCU

*School of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL 32816-2362, USA
Email: (xbai, lboloni, dcm)@cs.ucf.edu*

HOWARD JAY SIEGEL

*Department of Electrical and Computer Engineering
and Department of Computer Science
Colorado State University
Fort Collins, CO 80523-1373, USA
Email: HJ@ColoState.edu*

ROSE A. DALEY AND I-JENG WANG

*Applied Physics Laboratory
Johns Hopkins University
11100 Johns Hopkins Road Laurel, MD 20723-6099, USA
Email: (Rose.Daley, I-Jeng.Wang)@jhuapl.edu*

In this paper, we discuss a resource allocation model that takes into account the utility of the resources for the consumers and the pricing structure imposed by the providers. We show how a satisfaction function can express the preferences of the consumer both regarding the utility and the price of the resources. In our model, the brokers are mediating among the selfish interests of the consumers and the providers, and societal interests, such as efficient resource utilization in the system. We report a simulation study on the performance of the model.

1. Introduction

Resource management in a large-scale distributed system poses serious challenges due to the scale of the system, the heterogeneity and inherent autonomy of resource providers, and the large number of consumers and the diversity of their needs. Individual resource providers are likely to have

different resource management objectives and pricing structures. In this case, direct negotiation between resource providers and consumers is very inefficient. We need a broker to mediate access to resources from different providers. A broker is able to reconcile the selfish objectives of individual resource providers who want to maximize their revenues, with the selfish objectives of individual consumers who want to get the most possible utility at the lowest possible cost, and with some global, societal objectives, e.g., to maximize the utility of the system.

To formalize the objectives of the participants, we use: (i) a consumer *utility function*, $0 \leq u(r) \leq 1$, to represent the utility provided to an individual consumer, where r represents the amount of allocated resources; (ii) a provider *price function*, $p(r)$, imposed by an individual resource provider, and (iii) a consumer *satisfaction function*, $s(u(r), p(r))$, $0 \leq s \leq 1$, to quantify the level of satisfaction of an individual consumer that depends on both the provided utility and the paid price.

The consumer utility function could be a sigmoid⁷

$$u = u(r) = \frac{(r/\omega)^\zeta}{1 + (r/\omega)^\zeta}$$

where ζ and ω are constants provided by the consumer, $\zeta \geq 2$, and $\omega > 0$. Clearly, $0 \leq u(r) < 1$ and $u(\omega) = 1/2$.

The provider price could be a linear function of the amount of resources:

$$p = p(r) = \xi \cdot r$$

where ξ is the *unit price*.

A consumer satisfaction function takes into account both the utility provided to the consumer and the price paid. For a given utility, the satisfaction function should increase when the price decreases and, for a given price, the satisfaction function should increase when the utility increases. A candidate satisfaction function is⁶:

$$s(u, p) = 1 - e^{-\kappa \cdot u^\mu \cdot p^{-\epsilon}} \quad (1)$$

where κ , μ , and ϵ are appropriate positive constants.

Several systems, such as Nimrod/G³, Rexec/Anemone⁴, and SETI@home⁵, use market based models for trading computational resources. In this paper, we consider a model where the allocation of resources is determined by their price, their utility to the consumer, and by the satisfaction of the consumer.

2. A Utility, Price, and Satisfaction Based Resource Allocation Model

Consider a system with n providers offering computing resources and m consumers. Call $\mathcal{R}$ the set of providers and $\mathcal{U}$ the set of consumers. Consider provider R_j , $1 \leq j \leq n$, and consumer U_i , $1 \leq i \leq m$, that could potentially use resources of that provider.

Let r_{ij} denote the resource of R_j allocated to consumer U_i and let u_{ij} denote its utility for consumer U_i . Let p_{ij} denote the price paid by U_i to provider R_j . Let t_{ij} denote the time U_i uses the resource provided by R_j . Let c_j denote the resource capacity of R_j , i.e., the amount of resources regulated by R_j .

The term “resource” here means a vector with components indicating the actual amount of each type of resource:

$$r_{ij} = (r_{ij}^1, r_{ij}^2, \dots, r_{ij}^l)$$

where l is a positive integer and r_{ij}^k corresponds to the amount of resource of the k -th type. The structure of r_{ij} may reflect the rate of CPU cycles, the physical memory required by the application, the secondary storage, and so on.

The utility of resource of the k -th type provided by R_j for consumer U_i is a sigmoid:

$$u_{ij}^k = u(r_{ij}^k) = \frac{(r_{ij}^k / \omega_i^k)^{\zeta_i^k}}{1 + (r_{ij}^k / \omega_i^k)^{\zeta_i^k}}$$

where ζ_i^k and ω_i^k are constants provided by consumer U_i , $\zeta_i^k \geq 2$, and $\omega_i^k > 0$. Clearly, $0 < u(r_{ij}^k) < 1$ and $u(\omega_i^k) = 1/2$.

The overall utility of resources provided by R_j to U_i could be:

- the product over the set of resources provided by R_j , i.e., $u_{ij} = \prod_{k=1}^l u_{ij}^k$, or

- the weighted average over the set of resources provided by R_j , i.e., $u_{ij} = \frac{1}{l} \sum_{k=1}^l a_{ij}^k u_{ij}^k$, where a_{ij}^k values are provided by consumer U_i .

We consider a linear pricing scheme $p_{ij}^k = \xi_j^k \cdot r_{ij}^k$, though more sophisticated pricing structures are possible. Here ξ_j^k represents the unit price for resource of type k provided by provider R_j . The amount consumer U_i pays to provider R_j for a resource of type k is $p_{ij}^k \times t_{ij}$. The total cost for consumer U_i for resources provided by provider R_j is

$$p_{ij} = \sum_{k=1}^l p_{ij}^k \times t_{ij}$$

Based on Equation 1, we define the degree of satisfaction of U_i for a resource of the k -th type provided by provider R_j as

$$s_{ij}^k(u_{ij}^k, p_{ij}^k) = 1 - e^{-\kappa_i^k u_{ij}^k \mu_i^k (p_{ij}^k / \phi_i^k)^{-\epsilon_i^k}}, \quad \kappa_i^k, \phi_i^k, \mu_i^k, \epsilon_i^k > 0$$

where μ_i^k and ϵ_i^k control the sensitivity of s_{ij}^k to utility and price; ϕ_i^k and κ_i^k are normalization constants; ϕ_i^k is a reference price; and $\kappa_i^k = -\log \alpha$, with α a reference value for the satisfaction function. Detailed information about these parameters can be found in Bai.¹

The overall satisfaction of consumer U_i for resources provided by R_j could be:

- the product over the set of resources provided by R_j , i.e., $s_{ij} = \prod_{k=1}^l s_{ij}^k$, or
- the weighted average over the set of resources provided by R_j , i.e., $s_{ij} = \frac{1}{l} \sum_{k=1}^l b_{ij}^k s_{ij}^k$, where b_{ij}^k values are provided by consumer U_i .

We consider a *provider-broker-consumer model* that involves a broker B . In this model, the amount of resources to be allocated is determined according to a *target utility* (denoted as τ), i.e., the broker allocates an amount of resources such that the utility of each type of resource to the consumer reaches this τ value. The broker also has “societal goals” and attempts to maximize the average utility and revenue, as opposed to providers and consumers that have individualistic goals. To reconcile the requirements of a consumer and the candidate providers, a broker chooses a subset of providers such that the satisfaction is above a threshold and all providers in the subset have equal chances to be chosen by the consumer. We call the size of this subset *satisficing size*, and denote it as σ . Detailed information about the model can be found in Bai et al.²

Several quantities are used to characterize the resource management policy for broker B and its associated providers and consumers:

- The average hourly revenue for providers. The revenue is the sum of revenues for all of its resource types. This average is over the set of all providers connected to broker B .
- The consumer admission ratio. This ratio is the number of admitted consumers over the number of all consumers connected to B . A consumer is admitted into the system when there is a provider able to allocate some of the resources requested by the consumer, otherwise the consumer is dropped.
- The average consumer overall utility. This average is over the set of all admitted consumers connected to broker B .
- The average consumer overall satisfaction. This average is over the set of all admitted consumers connected to broker B .

3. A Simulation Study

We simulate a system of 100 clusters and one broker. The number of nodes of these clusters is a random variable normally distributed with the mean of 50 and the standard deviation of 30. Each node is characterized by a resource vector containing the CPU rate, the amount of main memory, and the disk capacity. For example, the resource vector for a node with one 2 GHz CPU, 1 GB of memory, and a 40 GB disk is $(2GHz, 1GB, 40GB)$. Initially, there is no consumer in the system. Consumers arrive with an inter-arrival time exponentially distributed with the mean of 2 seconds.

The parameters of the utility function of consumers, i.e., u_{ij}^k , are uniformly distributed in the intervals shown in Table 1. The CPU rate, memory space, and disk space of a request, r_{ij}^k , are exponentially distributed with the mean of $2GHz$, $4GB$, and $80GB$, and in the range of $[0.1GHz, 100GHz]$, $[0.1GB, 200GB]$, and $[0.1GB, 1000GB]$, respectively.

Table 1. The parameters for the simulation are uniformly distributed in the intervals displayed in this table.

Parameter	CPU	Memory	Disk
ξ	[5, 10]	[5, 10]	[5, 10]
ω	[0.4, 0.9]	[0.5, 1.5]	[10, 30]
κ	[0.02, 0.04]	[0.02, 0.04]	[0.02, 0.04]
μ	[2, 4]	[2, 4]	[2, 4]
ϵ	[2, 4]	[2, 4]	[2, 4]
ϕ	[40, 60]	[80, 120]	[1800, 2200]

The *demand-capacity ratio* for a resource type k is the ratio of the amount of resources requested by the consumers to the total capacity of resource providers for resource type k , $\sum_j c_j^k$. In our model, the consumers do not provide the precise amount of resources needed, they only specify their utility function. In the computation of the demand-capacity ratio, for each consumer and each resource, it is assumed that for the requested r_{ij} value the corresponding u_{ij} value is 0.9. The *demand-capacity ratio* vector for all resource types is $\eta_j = (\eta_j^1 \ \eta_j^2 \ \dots \ \eta_j^l)$. To simplify the interpretation of the results of our simulation we only consider the case when $\eta_j^1 = \eta_j^2 = \dots = \eta_j^l = \eta$. The service time t_{ij} is exponentially distributed with the mean of λ seconds. By varying the λ value we modify demand-capacity ratio so that we can study the behavior of the system under different loads.

For a multi-dimensional resource, we let the overall utility be the product of the utility of all types of resource, and we let the overall satisfaction

be the product of the satisfaction of all types of resource.

We investigate the consumer admission ratio, the average hourly revenue, the average consumer satisfaction, and the average consumer utility for different target utility (Figures 1) and satisficing size (Figure 2) under various scenarios of demand-capacity ratio (Figures 3). We also compare the system performance of our scheme for several σ values with a *random strategy* where we randomly choose a provider from the set of all providers, without considering the satisfaction function. We study the evolution in time. In each case, we run the simulation 50 times and show the average value and a 95% confidence interval.

Figure 1 (a) shows that when $\tau = 0.8$, $\tau = 0.85$, and $\tau = 0.9$, the consumer admission ratio is approximately 1.0, and the three plots overlap with each other. When $\tau = 0.95$, during the transient period some consumer requests are dropped. As time goes on, the consumer admission ratio increases. More consumers can be admitted into the system due to the resource fragmentation^a. In the steady state the admission ratio is 1. Figure 1 (b) shows that the average hourly revenue increases during the transient period, decreases due to the resource fragmentation, and then reaches a stable value. The larger is τ , the more resources are allocated to consumers, and the higher is the average hourly revenue. Figure 1 (c) shows that the average consumer satisfaction increases during the transient period and then reaches a stable value. The average consumer satisfaction is higher when τ is smaller; the smaller is τ , the more consumers can be admitted by resource providers with cheaper prices and these consumers experience higher satisfaction. Figure 1 (d) shows that the average consumer utility decreases during the transient period due to the resource fragmentation and then reaches a stable value. The average consumer utility is lower when τ is smaller.

Figure 2 (a) shows that the consumer admission ratio is approximately 1.0 for all cases. Figure 2 (b) shows that the average hourly revenue increases during the transient period, decreases due to resource fragmentation, and leads to a stable value. A small value of σ limits the number of choices the broker has and this restriction leads to lower average hourly revenues. The larger is σ , the higher is the average hourly provider revenue. The random strategy, which corresponds to the maximum value of $\sigma = |\mathcal{R}|$

^aResource fragmentation is an undesirable phenomena; in our environment the amount of resources available cannot meet the target utility value for any request and an insufficient amount of resources is allocated.

Figure 1. Consumer admission ratio (a), average hourly revenue (b), average consumer satisfaction (c), and average consumer utility (d) vs. time (in seconds) for $\sigma = 1$, $\eta = 1.0$, and $\tau = 0.8, 0.85, 0.9$, and 0.95 .

has the highest average hourly provider revenue. Figure 2 (c) shows that the average consumer satisfaction increases during the transient period and then leads to a stable value. The average consumer satisfaction is higher when σ is smaller. Indeed, when $\sigma = 1$ we direct the consumer to that resource provider that best matches the request. When we select at random one provider from the set of all providers we observe the lowest average consumer satisfaction - we have a high probability to select a provider that is not with the highest satisfaction. Figure 2 (d) shows that the average consumer utility drops during the transient period due to the resource fragmentation and then reaches a stable value. The average consumer utility is lower when σ is smaller. The random strategy has the highest average consumer utility; when σ is larger consumers have a better chance to get resources according to the τ values.

Figure 3 (a) shows that when η is set to 0.25, 0.50, or 0.75, the system

Figure 2. Consumer admission ratio (a), average hourly revenue (b), average consumer satisfaction (c), and average consumer utility (d) vs. time (in seconds) for $\tau = 0.9$, $\eta = 0.5$, and $\sigma = 1, 10$, and 20 . For the random strategy, $\sigma = |\mathcal{R}| = 50$.

is capable of handling all requests and the corresponding plots overlap with each other. When $\eta = 1.0$ some requests are dropped. As time goes on, the consumer admission ratio increases due to resource fragmentation. During the steady state the consumer admission ratio is 1. Figure 3 (b) shows that the average hourly revenue increases during the transient period, and then decreases to reach a steady value. The larger is η , the higher is the average hourly revenue. The average consumer satisfaction drops during the transient period, increases due to resource fragmentation, and then converges to a steady value, as shown in Figure 3 (c). The smaller is η , the earlier the system reaches the steady state and the higher is the average consumer satisfaction. The average consumer utility drops during the transient period and then reaches a steady value, as shown in Figure 3 (d). The smaller is η , the earlier the system reaches the steady state and the higher is the average consumer utility.

Figure 3. Consumer admission ratio (a), average hourly revenue (b), average consumer satisfaction (c), and average consumer utility (d) vs. time (in seconds) for $\tau = 0.9$ and $\sigma = 1$.

4. Conclusions

Economic models are notoriously difficult to study. The complexity of the utility, price, and satisfaction based models precludes analytical studies and in this paper we report on a simulation study. The goal of our simulation study is to validate our choice of utility, price, and satisfaction functions, to study the effect of the many parameters that characterize our model, and to get some feeling regarding the transient and the steady-state behavior of our models. We are primarily interested in qualitative rather than quantitative results, and we are interested in trends rather than actual numbers. It is too early to compare our model with other economic models proposed for resource allocation in distributed systems, but we are confident that a model that formalizes the selfish goals of consumers and providers, as well as societal goals, has a significant potential. This is a preliminary study that cannot provide a definite answer to the question posed in the title of

the paper. Our intention is to draw the attention of the community to the potential of utility, price, and satisfaction based resource allocation models.

The function of a broker is to monitor the system and set τ and σ for optimal performance. For example, if the broker perceives that the average consumer utility is too low, it has two choices: increase τ or increase σ . At the same time, the system experiences an increase of average hourly revenue and a decrease of average consumer satisfaction. We note that while the utility is always increasing with the amount of allocated resources, the satisfaction also takes into account the price paid and exhibits an optimum at a certain level of resources. Increasing the resources beyond the optimum will still increase the utility but yield lower satisfaction, because the additional utility was paid an unjustifiably high price.

The simulation results shown in this paper are consistent with those in Bai et al.², where we use a much simpler model based upon a synthetic quantity to represent a vector of resources.

5. Acknowledgments

This research was supported in part by National Science Foundation grants MCB9527131, DBI0296035, ACI0296035, and EIA0296179, the Colorado State University George T. Abell Endowment, and the DARPA Information Exploitation Office under contract No. NBCHC030137.

References

1. X. Bai. *Coordination, Matchmaking, and Resource Allocation for Large-Scale Distributed Systems*. PhD thesis, University of Central Florida, 2006.
2. X. Bai, L. Bölöni, D. C. Marinescu, H. J. Siegel, R. A. Daley, and I.-J. Wang. A brokering framework for large-scale heterogeneous systems. *To be presented at the 15th Heterogeneous Computing Workshop*, Rhodes Island, Greece, April 2006.
3. R. Buyya, D. Abramson, and J. Giddy. Nimrod/g: An architecture of a resource management and scheduling system in a global computational grid. In *Proc. of the 4th Int. Conf. on High Performance Computing in the AsiaPacific Region*, volume 1, pages 283–289, 2001.
4. B. Chun and D. Culler. Market-based proportional resource sharing for clusters. Technical report, University of California, Berkeley, September 1999.
5. SETI@home. URL <http://setiathome.ssl.berkeley.edu/>.
6. H. R. Varian. *Intermediate Microeconomics: A Modern Approach*. Norton, New York, March 1999.
7. M. Xiao, N. Shroff, and E.-P. Chong. Utility-based power control in cellular wireless systems. In *INFOCOM 2001, Joint Conf. of the IEEE Computer and Communication Societies*, pages 412–421, 2001.