Dynamic Auction Mechanism for Cloud Resource Allocation

Wei-Yu Lin Guan-Yu Lin Hung-Yu Wei*
Department of Electrical Engineering, National Taiwan University
Corresponding Author*: hywei@cc.ee.ntu.edu.tw

***Abstract*—We propose a dynamic auction mechanism to solve the allocation problem of computation capacity in the environment of cloud computing. Truth-telling property holds when we apply a second-priced auction mechanism into the resource allocation problem. Thus, the cloud service provider (CSP) can assure reasonable profit and efficient allocation of its computation resources. In the cases that the number of users and resources are large enough, potential problems in second-priced auction mechanism, including the variation of revenue, will not be weighted seriously since the law of large number holds in this case.**

I. INTRODUCTION

In cloud computing, the allocation method plays a key role in managing large scale of computation capacity. Market-oriented allocation rules, which apply pricing mechanism into capacity control, are useful for the design of a more efficient algorithm. Improper allocation rules might cause the inefficiency of the system. Izakian et. al. [1] proposed a Continuous Double Auction (CDA) mechanism, resembling the market mechanism, to solve a resource allocation problem in grid computing. In this study, the centralized auctioneer distributes the resources by matching the values of both sides and the CDA mechanism ensures the allocation to be efficient.

We propose a *second-price auction mechanism*, which applies the marginal bid (the highest among the unsuccessful bids) to determine the price of the resource, for computation capacity allocation with the assistance of pricing and truth-telling mechanism. This mechanism applies the original development by Vickery [2] into the process of dynamic adjustment to the computation capacity. Our mechanism ensures the efficient allocation of computation capacity and generates good revenue to the administrator. Furthermore, the dynamic adjustment endures the variation of users' distributions and generates efficient allocation by observing only one user value.

Two contributions of our study should be addressed: I. The introduction of peak/off-peak concept into the resource allocation problem and II. The systematic analysis containing background and float tasks. Traditionally, the analysis of resource allocation problem applies the concept of *fixed* demand, which does not vary over time. While in this work, the varying demand can be described by our peak/off-peak concept more exactly and therefore the cloud service provider can further improve both system efficiency and its own revenue. In our framework, the cloud service provider not only distributes out resources but also possesses its own background task. The introduction of this modeling may improve the profit of the service provider and the performance of the system.

This work is supported by the Ministry of Economic Affairs in Taiwan under Grant 98-EC-17-A-02-S2-0127.

II. SYSTEM MODEL

We construct a real-time model consisting of two periods with n cloud users and a cloud service provider (CSP). The CSP has two tasks: performing time-insensitive background computing and distributing resource to the cloud users in the dynamic process. If the total input into the background task excess the threshold, the CSP will gain a fixed amount of value. The CSP will also sell its residual resources to the cloud users after decideing how much resource shall be distributed to the background task. Thus, we model the CSP's utility function and resource constraint as follows:

$$\Pi^{SP} = \pi_B(r_1^B + r_2^B) + \sum_{t=1}^{2} \pi_{Vt}(r_t^V) \tag{1}$$

$$r_B^V + r_t^V = R \quad \text{for all t=1,2} \tag{2}$$

Π^{SP} is the CSP's total revenue composed of the value generated by background task, π_B, and the revenue collected from the users, π_{Vt} at time t. r_t^B is the total resource allocated into background computation at time t, and r_t^V is the total resource distributed to the cloud users. The value of background computation is denoted as $\pi_B(r_1^B + r_2^B)$ and

$$\pi_B = \begin{cases} V_B & \text{if } r_1^B + r_2^B \geq \underline{r}^B \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

To simplify our analysis of users' time-variant resource demand, we assume the cloud users only demand variable computing capacity. It means the demand of each user is determined by its own task at certain period. However, the demand of the users varies in peak and off-peak periods. For generalization, we assume each user demand one unit of capacity.

The proposed a mechanism is based on sealed-bid auction. In the beginning of each period, the users submit their bids to the CSP. The CSP then collects all the bids and determines the price. Assume the decision rule of the CSP on background task is to simply divide the task equally into two periods. In other words, the CSP will provide capacity $k = R - \underline{r}^B/2$ to cloud users in each period, and the price is determined by the $(k+$

$1)^{th}$ highest bid. It means the resource will be distributed to the first k^{th} highest bidders under the price of the $(k+1)^{th}$ highest bid. Based on the essential property of this mechanism, truth-telling in the users' bids holds since their payments do not rely on their own bid [3]. With this property, we may reduce the resource allocation problem into an ordering problem. The rule of resource allocation is to observe the ordered values of the cloud users. The advantage of the system is the simplification of the CSP's decision rule and the clear-cut allocation rule.

III. VALUE VARIATION AND REVENUE INFERIORITY PROBLEM

We construct a value matrix to analyze the payoff of the CSP. The value matrix can be expressed as the bids made by users ordered form high to low:

$$\pi_B = \begin{pmatrix} \vec{V_1} \\ \vec{V_2} \end{pmatrix} = \begin{bmatrix} V_{11} & \cdots & V_{1n} \\ V_{21} & \cdots & V_{2n} \end{bmatrix} \tag{4}$$

The price vectors of period 1 and 2 are $(v_{1(k+1)}, v_{2(k+1)})$, and the revenue vector CSP collected from the users $(kv_{1(k+1)}, kv_{2(k+1)})$ is nearly the same as the producer revenue in competitive market. However, this mechanism does not ensure profit maximization due to its truth-telling property under constraints.

To examine the potential weakness of this mechanism, such as possible strategic deviation and the revenue inferiority [4], we assume the bid values of each user is a sampling from a given common distribution, random variable X. Define Y as the event that N values are sampled from X and x is the $(K+1)^{th}$ larger sample of the N ones.

$$f_X(X=x|Y) = \frac{P_r(Y|X=x)f_X(X=x)}{\int_0^\infty P_r(Y|X=x)f_X(X=x)\mathrm{d}x} \tag{5}$$

$$P_r(Y|X=x) = \frac{\begin{pmatrix} N \\ K \end{pmatrix}(1-F(x))^K F(x)^{N-K-1}}{\int_0^\infty \begin{pmatrix} N \\ K \end{pmatrix}(1-F(x))^K F(x)^{N-K-1}\mathrm{d}x} \tag{6}$$

We derive the likelihood function as following:

$$\begin{aligned} & f_X(X=x|Y) \\ = & \frac{(1-F(x))^K F(x)^{N-K-1} f_X(X=x)}{\int_0^\infty (1-F(x))^K F(x)^{N-K-1} f_X(X=x)\mathrm{d}x} \end{aligned} \tag{7}$$

In our simulation, we assume X is a Gaussian distribution with mean 30 and variance 3. Fig.1 shows the price distribution given K allocated resources and N total bidders($N = 2K+1$). The pink line, as the probability density function of X, is original bidder value distribution. As we expected, when $N = 2K+1$ holds, the peak value of each curve with different N is the same as $E(X)$. (This result can be observed directly from eq.(7) and it applies to normal and uniform X). Moreover, the variation of profit from peak value decreases obviously as number of bidders increases. In other words, by auctioning reources to about half number of bidders, the mean of user distribution can be specified. Furthermore, even if number of bidders is small ($N = 3, K = 1$), the variation in profit is still not as serious as original distribution X.

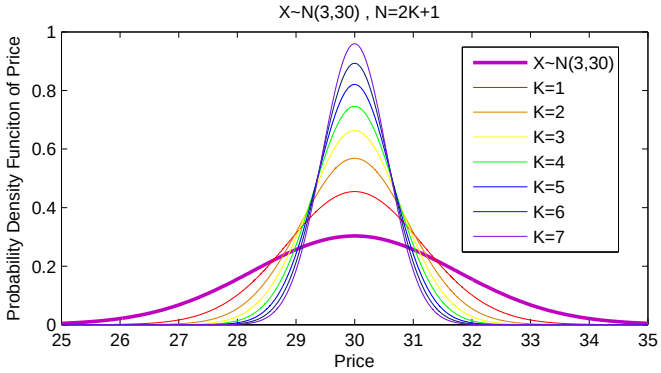

Fig. 1. Revenue distribution

By the simulation result, we conclude that when number of users and the resource are large enough, this mechanism will not only overcome the possible strategic deviation but also the revenue inferiority and generate revenue pretty close to the market-clearing price. Even when the number of resources and users are not large enough, we may also apply the same condition.

IV. CONCLUSION AND FUTURE WORKS

The main contribution of this paper is developing a new resource allocation algorithm by applying auction method into the resource allocation problem in cloud computing. We propose a theoretical framework to cope with the capacity distribution under cloud computing framework. Under the proposed mechanism, we ensure the efficient allocation of capacity under simple decision rule and generate appropriate revenue to the CSP. We also examine potential drawbacks of this mechanism proposed by the literature and successfully found the condition underneath. To extend the application both in the slope and scale, the same task schedule may be applied to the cloud users in the future works since it enables the CSP to build more pricing criteria and thus improving the efficiency of the system. Moreover, it is also possible to develop a dynamic adjustment mechanism on the basis of our finding. Since we assume there are peak and off-peak demands on the capacity, the modification can improve not only the efficiency of the system but also the revenue generated for the CSP.

REFERENCES

[1] H. Izakian, et al., "An auction method for resource allocation in computational grids," *Future Gener. Comput. Syst*, Vol. 26, pp. 228-235, 2010.

[2] W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *The Journal of Finance*, Vol. 16, pp. 8-37, 1961.

[3] B. Awerbuch, et al., "Reducing truth-telling online mechanisms to online optimization," *presented at the Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*,San Diego, CA, USA, 2003.

[4] T. Sandholm, "Issues in computational Vickrey auctions," *Int. J. Electron. Commerce*, Vol. 4, pp. 107-129, 2000.