Users' time preference based stochastic resource allocation in cloud spot market: Cloud provider's perspective

Research in Progress

Anik Mukherjee[1], R P Sundarraj[1], Kaushik Dutta[2]

[1]Department of Management Studies, Indian Institute of Technology, Madras
Sardar Patel Road, Chennai, Tamil Nadu 600036 – India.
{anikit.jgec@gmail.com; rpsundarra}@iitm.ac.in
[2]Department of Information Systems and Decision Sciences, Muma College of Business,
University of South Florida, Tampa, Florida, United States of America
duttak@usf.edu

Abstract. Cloud Computing spot markets have enabled the users to make use of the spare computing capacities of the cloud providers at a relatively cheaper price which in turn has given the providers such as Amazon and Google an opportunity to earn extra money by auctioning-off the underutilized resources. However, resource availability is a problem in the spot market owing to spot-price fluctuations. Ignoring the customer's preference is one of the potential reasons behind this. In this paper, we propose a time preference (value of service at different points of time) based stochastic integer linear programming model to allocate the cloud resources among the cloud users with a view to maximizing the revenue of cloud providers from the spot-market.

Keywords: Cloud Computing · Resource Allocation · Spot Market · Time Preference · Stochastic Programming.

1 Introduction

According to the National Institute of Standards and Technology, "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." [NIST 2016]. In fact, cloud computing is an online marketplace where users with less computing capacity can make use of others' computing resources shared over the internet. There are mainly four types of cloud computing facilities available in the market namely On-Demand, Reserve, Dedicated Host and On-Spot1. Each of the computing resources varies in terms of facilities provided by the cloud providers. First three types of the resources allow the cloud users to

1 https://aws.amazon.com/ec2/pricing/

access the virtual machines[2] for a longer period of time whereas On-Spot users can use the cloud resource on an hourly basis.

This paper focuses on the Spot market, which was first introduced by Amazon in the year 2009 to make better use of computing resources available in the cloud computing market [Amazon 2016]. Spot market works like electricity market where a number of users bid for a resource, and resource is allocated to the winner [Muli et al. 2013, Song et al. 2012]. For example, in the case of Amazon AWS, users of the spot instances bid for a resource, and Amazon dynamically set a spot price depending on the incoming bid and the available resources. If the user's bid exceeds the spot price of the instance, the resource is assigned to the user. Thus, spot markets provide a mechanism for buying and selling of computing services in a form of Cloud Instance. Cloud users act as a buyer as they need to procure computing resources. Like other online markets, spot markets also influence buyers' decision making while buying resources. In this context, one of the most important factors is the willingness to buy a resource at a particular time. To elaborate on this, we can say that a user may want to access one resource at time t or (t+1). But, (s)he prefers to get the resource at t to (t+1) i.e. he should be willing to pay more at t for that resource than the time-period (t+1). This behavior of customers in online marketplaces has been studied extensively in the literature of Intertemporal Choice Theory [Frederick et al. 2002], the study of trade-offs between cost/benefits of goods/ services and delivery at different times. In the context of spot market of cloud computing, we introduce this concept of the trade-off between time and cost as Time-preference where the cloud providers capture the time preferences of the interested users at the different point of time. Based on the captured preferences, cloud providers allocate the resources among the users. We term this special kind of market as an Extended Spot Market which considers Time Preference of the users as additional information during spot resource allocation. In the previous work of Mukherjee et al. (2016), authors have already mathematically proved that additional information such as Time Preference in spot resource allocation improves the cloud providers' revenue from the traditional spot allocation followed in Amazon [Mukherjee et al. 2016]. In this paper, we present a model and solution methodology to allocate cloud resources among the interested cloud users based on the time preference. Our paper is structured as follows: Section 2 presents a review of the relevant literature. Problem definition is included in Section 3, while Sections 4 and 5 detail the Mathematical Model and Solution Approach respectively. Implications and concluding remarks are available in Sections 6.

2 Literature Review

In the past, researchers have done extensive work on the revenue maximization and cost minimization for the cloud providers and cloud users in the context of resource allocation. Research on these domains can be broadly classified into the following two categories as follows:

[2] Note, the word Virtual Machine and Resource are used interchangeably in this paper.

Cloud Services

Cloud providers provide four types of services namely On-Spot, On-Demand, Reserved and Dedicated Host. Cloud users are charged based on the services they use. Researchers have identified that the cloud providers always try to maximize their revenue from the cloud services. On the contrary, cloud users tend to minimize the cloud computing cost incurred by them. So, algorithms have been proposed in the literature to address this kind of problem. As an example, Chaisiri et al. presented a stochastic programming and Benders decomposition based cloud resource provisioning technique to minimize the cost of cloud consumers [Chaisiri et al. 2012] whereas Toosi et al. employed a dynamic pricing based auction mechanism to generate near-optimal profit of the cloud provider in spot market resource allocation [Toosi et al. 2016]. In the context of revenue maximization, Alzhouri et al. have shown that the dynamic pricing scheme in spot market can help the cloud organizations utilize the spare resources (VMs) more efficiently [Alzhouri et al. 2017]. Toosi et al. devised dynamic programming based allocation algorithm to maximize the overall revenue of the cloud providers by allocating Virtual Machines to three different cloud service markets namely on-demand, on-spot and reservation [Toosi et al. 2015].

Cloud Auction & Pricing

In the realm of auction and pricing, researchers have proposed various schemes and methodology for the betterment of existing techniques. In Ijakian et al., a double auction based resource allocation mechanism is presented to allocate resources among the users for better resource utilization, profit allocation, and execution rate [Ijakian et al. 2010]. Few researchers have argued that the fixed price models, as adopted by Amazon EC2, Microsoft Azure [Amazon 2016, Azure 2015] etc., do not guarantee economically efficient resource allocation [Zaman and Daniel 2013]. Rather application of combinatorial auction with prior knowledge of user's demand in cloud resource allocation generates better revenue of the cloud providers by maximizing the user's utility with higher resource utilization [Zaman and Daniel 2013]. However, this article does not talk about the truthfulness of the bidders. Toosi et al. proposed a back-propagation based price formation algorithm to maximize the market surplus as well as truthful bidding by the users [Toosi et al. 2016].

In the context of bidding, user's response to the Quality of Service offered by the cloud providers plays a key role in resource allocation. De et al. considered user's patience while allocating the VMs to the different users [De et al. 2016]. Their experiment shows that the proposed strategy reduces the resource allocation cost and improves the Quality of Service (QoS) level.

However, the research on the time-preference based cloud resource allocation is almost scant in the literature. We pick up on this idea and try to address the gap by developing a time-preference based mathematical model to allocate resource in cloud spot-market.

3 Problem Definition

Set of jobs (henceforth called cloud users) arrive at the cloud spot marketplace (say Amazon Web Service) to access the cloud resources as per their needs/ requirements. Cloud providers ask the cloud users for their time preferences for the next t time periods (say 5 time periods). The rate of arrival jobs is purely stochastic[3] in nature. The problem can be explained as follows:

First, the cloud provider captures the time preferences of the users/ customers arriving at time period t_1 for the next t time periods (i.e. up to t_1+t) and allocates resource for the t_1^{th} period. In the next time period (i.e. at t_2^{th} or $(t_1+1)^{th}$), an additional number of users arrive at the same cloud marketplace to access the cloud resources. Hence, the cloud provider needs to capture the time preferences of these customers for the next t time periods (i.e. up to t_2+t where $t_2>t_1$). This process continues for each time period as soon as a new customer arrives at the cloud marketplace. On the contrary, the cloud provider needs to allocate resources to the existing customers at each time period on a rolling basis depending on the availability of the resources. Note, the number of available resource changes over time. So, the resource allocation should be done in such a way that the overall revenue of the cloud provider is maximized given that the waiting time of the customer to get the preferred resource is not exceeded by a certain threshold. However, the spot market is a highly unstable market. The choice of a resource to bring in the spot market from the reserve instances may influence the overall revenue of the cloud provider as the power consumption cost by a resource may vary over time. Thus, selecting the right resource for the current spot market is a key decision to be taken by the cloud provider. In this paper, we aim to develop a solution methodology to help the cloud provider maximize the overall revenue from the cloud spot market by selecting the right resource for the right users/ customers.

4 Stochastic Programming Model

As mentioned earlier, few parameters such as arrival rate of the customers etc. are stochastic in nature. Thus, Stochastic Programming Model will be a good option to address the real-life spot market resource allocation problem mentioned in this paper. In past, the researchers have captured the stochastic parameters using various distributions such as Normal distribution, Poisson distribution, Pareto distribution [Beloglazov and Buyya 2010, Alzhouri et al. 2017, Javadi et al. 2011]. Hence, we decide to use these distributions to develop the Stochastic Programming model. The detailed description of the model is given below.

[3] Note, stochastic means the number of users bidding for a resource at a time is changing over time.

4.1 Model Description

The objective of the model is to maximize the revenue of the cloud provider by calculating the difference between the amount of money earned from the users' time preferences and the operational cost of the resources [Equation 1]. In the context of cloud computing, we assume that the operational cost[4] is proportional to the power consumption. Further, we consider the cloud resources (Virtual Machines) as the key component of the allocation mechanism. Hence, we have added two constraints related to the resource namely resource requirement by the user as well as the resource availability [Equation 2]. The first part of the Equation 2 refers to the Requirement Constraint which indicates the total number of resources required by the cloud users at that time-period whereas the second part refers to the total number of available resources at that time-period. However, the cloud users too play a key role in the revenue maximization of the cloud provider. As the spot-market is highly competitive in nature, all the cloud users may not be able to get their preferred resources as soon as entering the market. They may have to wait for some time to get hold of the resource. In the Equation 3, we have captured the user's willingness to wait for a resource. Finally, we need to check the quality of resource present in the spot market. Equation 4 guarantees that the best revenue generating resources are brought into spot market by the cloud provider [Note, the term 'best' refers to the virtual machines capable of accommodating more number of cloud users with lesser operational cost]. In this paper, we are assuming that all the non-stochastic input parameters such as operational cost, waiting time of the users etc. are known beforehand.

4.2 Formulation

In the context of cloud computing, we have considered the user's time preference [$Pref_i^{(tr)}$] and the power consumption rate [P(U)] by the available resources as stochastic parameters. The term stochastic refers to the uncertain nature of the parameter. To explain it further, we can say that the power consumption rate depends on the resource utilization which varies over time. Again, the number of users may vary each time-period, so does the preferences [Alzhouri et al. 2017]. According to Beloglazov and Buyya, the power consumption rate can be modeled as Normal Distribution [Beloglazov and Buyya 2010] whereas many researchers suggested that the arrival rate of customers follow Poisson distribution [Alzhouri et al. 2017]. Similarly, the resource availability as well as requirements in the spot market changes with time. In this case, resource availability follows Pareto distribution and resource requirement follows Poisson distribution [Javadi et al. 2011]. A detailed description of the model is given below:

Allocation Constraint — *A number of resources allocated to the users should lie between the required number of resources and available number of*

[4] Operational cost can be found on the product of the power consumption (by the Virtual Machines) derived from the power model proposed by [Beloglazov and Buyya 2010] and per-unit power consumption cost.

	resources. (Equation 2).
Waiting-Time Constraint	*Waiting time of the cloud users should be within a specified time duration. (Equation 3).*
Resource Constraint	*Which resources to bring in the spot-market? (Equation 4)*

Notations

Input Variables

$Pref_i^{(tr)}$	*Time-Preference of i^{th} user at t^{th} time-period for r^{th} resource.*
$P(U)$	*The power consumption of a resource as a function CPU-Utilization.*
$Cost_r$	*Per-unit power consumption cost of the resource.*
$Required(t)$	*The stochastic demand of number resources by the cloud users.*
$Available(t)$	*Stochastic supply of a number of resources by the cloud provider.*
$wait(ir)$	*Waiting time of the cloud users to access the resource.*
k	*The fraction of the power consumed by an idle-server.*
P_{max}	*Maximum power (e.g. 250W) usage by a cloud resource.*
$U(t)$	*Utilization of resource at time t.*
$\overline{U_r}$	*Mean of CPU Utilization of each resource r.*
$S_{U_r}^2$	*The variance of CPU utilization of each resource r.*

DecisionVariables

$X_i^{(tr)}$	*1, i^{th} user is allotted r^{th} resource at t^{th} time-period.*
	0, otherwise
Y_r	*1, r^{th} resource type is selected.*
	0, Otherwise

Objective

$$Max \quad E\left[Pref_i^{(tr)} X_i^{(tr)}\right] - E[P(U) Cost_r Y_r] \qquad \textbf{(1)}$$

SUBJECT TO CONSTRAINTS,

$$E(Required(t) Y_r) \leq \sum_{ir} X_i^{(tr)} \leq E(Available(t) Y_r) \quad , \forall_t \qquad \textbf{(2)}$$

$$\sum_t t\, X_i^{(tr)} \leq wait(ir) \quad , \forall_{i,r} \qquad \textbf{(3)}$$

$$\sum_r Y_r = E(Available(t)) \quad , \forall_t \qquad \textbf{(4)}$$

$$X_i^{(tr)} \in \{0,1\} \qquad \textbf{(5)}$$

$$Y_r \in \{0,1\} \qquad \textbf{(6)}$$

Where,

$$P(U) = k\, P_{max} + (1-k) P_{max}\, U(t) \qquad \textbf{(7)}$$

$$U(t) \sim Normal - Distribution\left(\sum_{r=1}^{m} \overline{U_r}, \sqrt{\sum_{r=1}^{m} \sqrt{S_{U_r}^2}}\right) \qquad \textbf{(8)}$$

$$Available(t) \sim \text{Pareto-Distribution(Scale, Shape)} \qquad \textbf{(9)}$$

$$Required(t) \sim \text{Poisson-Distribution(mean)} \quad \textbf{(10)}$$

$$Pref_i^{(ir)} \sim \text{Poisson-Distribution(mean)} \quad \textbf{(11)}$$

$$wait(tr) \sim \text{Poisson-Distribution(mean)} \quad \textbf{(12)}$$

5 Solution Approach

As discussed in the previous section, the mathematical model presented in this paper is stochastic in nature which is practically very difficult to solve in real-time. In the past, researchers have presented various approaches to solving stochastic programming. As an example, the detailed methodologies to solve linear programming under uncertainty have been addressed in the literature [Powell 2014, Minoux 2007, Dantzig 1955, Birge 1997]. In this section, we have presented a solution methodology in the context of spot market cloud resource allocation. To avoid the complexity of the problem, we have considered a special case of our model outlined before. In this regard, we have considered only one stochastic variable namely resource availability (available) and rest are kept deterministic. Further, we assume that the available resources are identical to each other. Hence, the resource utilization and power consumption cost are constant. So, we have discarded the second expression (pertaining to operational cost) from the objective function as well as the resource index r from the mathematical model. On the other hand, as we have considered identical resources, we do not need to have a separate decision variable to choose the best resource among all. Hence, the constraint pertaining to choose a resource is not required in the problem. Thus, we decide to remove Equation 4 from the mathematical model. Finally, we can rewrite the Stochastic Programming model as follows:

$$Max \quad Pref_i^{(t)} X_i^{(t)} \quad \textbf{[Problem P1]} \quad \textbf{(13)}$$

Subject to Constraints,

$$\sum_i X_i^{(t)} \leq Available(t) \quad , \forall_t \quad \textbf{(14)}$$

$$\sum_t t\, X_i^{(t)} \leq wait(i) \quad , \forall_i \quad \textbf{(15)}$$

$$X_i^{(t)} \in \{0,1\} \quad \textbf{(16)}$$

To solve the problem (P1), we have adopted the methodology namely Stochastic Integer Linear Programming with Uncertainty in Right Hand Side proposed by Gabrel et al. 2010. As already mentioned earlier, the key objective of our work is to present a candidate solution methodology of the spot market resource allocation problem. Hence, the methodology proposed in Gabrel et al.'s paper is expected to work well in our problem as we are also dealing with uncertainty on the right-hand side of the constraint [i.e. $Available(t)$ of Equation 14]. According to [10], we can consider that the number of resources varies in the interval $[\underline{a}, \overline{a}]$ such that $Available(t)$ can take values from the defined closed interval. [Gabrel et al. 2010] have shown that the optimum solution of the problem can be found in polynomial time by including the interval constraint in the

linear program. After the inclusion of interval constraint [Equation 18], the math model becomes:

$$Optimize \quad \text{Problem P1} \tag{17}$$

Subject to Constraints,

$$\underline{a} \leq Available(t) \leq \overline{a} \quad , \forall_t \tag{18}$$

So, we need to solve this deterministic Integer Linear Program[5] for cloud resource allocation. To explain it further, if we add all the expressions which have been discarded earlier in this section for the sake of simplicity (provided the operational cost is known beforehand), the problem becomes a deterministic equivalent of the stochastic programming proposed in Section 4.2. This methodology gives us an initial direction to solve the basic Stochastic Programming for spot market cloud resource allocation.

6 Research Implications and Conclusions

The problem we have considered here is a resource allocation problem in the cloud computing domain. In our exhaustive search of the literature, we have not found any cloud resource allocation paper using the time-preference of the users. As this is a research-in-progress paper, we have proposed a candidate solution methodology to solve this problem. In future, we will extend this research further by considering multiple stochastic parameters simultaneously.

The key contribution of the paper is the incorporation of user's time preference in cloud computing paradigm. As this is a stochastic problem, it is very difficult to get a good solution of this problem in real time. So, our next aim is to come up with theorems/ online algorithm for efficient resource allocation which will maximize the revenue of the cloud providers with better resource allocation capability in the cloud spot market. In addition to it, we aim to extend this problem by defining a market place which will help both the cloud users as well providers gaining better revenue by exchanging resources among different cloud providers across different zones (inter-zonal and intra-zonal transfers).

References

1. Amazon 2016. https://aws.amazon.com/ec2/purchasing-options/.
2. Alzhouri, F., Anjali, A., Yan, L., and Ahmed, S. B.: Dynamic Pricing for Maximizing Cloud Revenue: A Column Generation Approach. 18th International Conference on Distributed Computing and Networking, 22. ACM, (2017).
3. Azure 2015. https://azure.microsoft.com/en-in/pricing/
4. Beloglazov, A., and Buyya, R.: Adaptive threshold-based approach for energy-efficient consolidation of virtual machines in cloud data centers. MGC@ Middleware. (2010).

[5] Note, we need to solve a series of Linear Program to solve an Integer Linear Program. Hence, the methodology proposed by Gabriel et al., 2010 seems to work well in our problem.

5. Birge, J. R.: State-of-the-art-survey—Stochastic programming: Computation and applications. INFORMS journal on computing 9, 2, 111-133(1997).
6. Chaisiri, S., Bu-Sung, L. and Dusit, N.: Optimization of resource provisioning cost in cloud computing. IEEE Transactions on Services Computing 5.2, 164-177(2012).
7. Dantzig, G. B.: Linear programming under uncertainty. Management science 1, 3-4: 197-206(1955).
8. De, A., Marcos, D., Carlos, H. C., Marco, A. N. and Renato, L. C.: Impact of user patience on auto-scaling resource capacity for cloud services. Future Generation Computer Systems 55, 41-50(2016).
9. Frederick, S., Loewenstein, G., O'donoghue, T.: Time discounting and time preference: a critical review. J Econ Lit 40(2),351–401(2002).
10. Gabrel, V., Cécile, M., and Nabila, R.: Linear programming with interval right hand sides. International Transactions in Operational Research 17, 3, 397-408(2010).
11. Izakian, H., Ajith, A., and Behrouz, T. L.: An auction method for resource allocation in computational grids. Future Generation Computer Systems 26, 2, 228-235(2010).
12. Javadi, B., Derrick, K., Jean-Marc, V., and David, P. A.: Discovering statistical models of availability in large distributed systems: An empirical study of seti@ home. IEEE Transactions on Parallel and Distributed Systems 22, 11, 1896-1903(2011).
13. Minoux, M.: Robust LP with Right-Handside Uncertainty, Duality and Applications. (2007).
14. Mukherjee, A., Sundarraj, R. P., and Dutta, K.: On Considering Customer's Short-term Preferences for Resource Allocation in Cloud Computing Spot Markets. Workshop on Information Technology and Systems (WITS 2016) (2016).
15. Muli, B. Y., Assaf, S., Orna, A., Muli, B., Assaf, S., and Dan, T.:. Deconstructing Amazon EC2 spot instance pricing. ACM Transactions on Economics and Computation (1:3), pp16:1-16, 20 (2013).
16. NIST 2016.: http://www.nist.gov/itl/csd/cloud-102511.cfm
17. Powell, W. B.: Clearing the jungle of stochastic optimization. In Bridging Data and Decisions, 109-137. Informs(2014).
18. Song, Y., Murtaza, Z., and Kang-Won, L.: Optimal bidding in spot instance market. INFOCOM, 2012 Proceedings IEEE. IEEE, (2012).
19. Toosi, A. N., Kurt Vanmechelen, K. R. and Buyya, R.: Revenue maximization with optimal capacity control in infrastructure as a service cloud markets. IEEE transactions on Cloud Computing 3, 3, 261-274 (2015).
20. Toosi, A. N., Kurt Vanmechelen, F. K., and Buyya, R.: An auction mechanism for cloud spot markets. ACM Transactions on Autonomous and Adaptive Systems (TAAS) 11, 1, 2(2016).
21. Zaman, S. and Daniel, G.: A combinatorial auction-based mechanism for dynamic VM provisioning and allocation in clouds. IEEE Transactions on Cloud Computing 1, 2, 129-141(2013).