Profit maximization model for cloud provider based on Windows Azure platform

Chaisiri, Sivadon; Lee, Bu-Sung; Niyato, Dusit

2012

Chaisiri, S., Lee, B. S., & Niyato, D. (2012). Profit maximization model for cloud provider based on Windows Azure platform. In Proceedings of 2012 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, pp.1-4.

https://hdl.handle.net/10356/94751

https://doi.org/10.1109/ECTICon.2012.6254333

Profit Maximization Model for Cloud Provider Based on Windows Azure Platform

Sivadon Chaisiri[†], Bu-Sung Lee[†‡], and Dusit Niyato[†]
[†] School of Computer Engineering, Nanyang Technological University, Singapore
[‡] HP Labs Singapore, 1 Fusionopolis Way, 14th Floor Connexis (South Tower), Singapore 138632
Email: {*schaisiri,ebslee,dniyato*}*@ntu.edu.sg*

***Abstract*—This paper studies a cloud computing market where a cloud provider rents a set of computing resources from Windows Azure operated by Microsoft. The cloud provider can integrate value-added services to the resources. Then, the services can be sold to customers, and the cloud provider can earn a profit. Moreover, the cloud provider could save much cost and increase higher profit with the 6-month subscription plan offered by Windows Azure. However, the maximization of profit is not trivial to be achieved since the amount of the customers' demand cannot be perfectly known in advance. Consequently, the subscription plan could not be optimally purchased. To deal with such a maximization problem, the paper proposes a stochastic programming model with two-stage recourse. The numerical studies show that the model can maximize the profit under the customers' demand uncertainty.**

NOMENCLATURE

$\mathcal{I}$	Set of compute instances, $i \in \mathcal{I}$
$\mathcal{J}$	Set of cloud services, $j \in \mathcal{J}$
Ω	Set of scenarios, $\omega \in \Omega$
F	Monthly subscription fee ($ per base unit)
C_i	On-demand price of instance i ($ per hour)
P_j	Selling price for cloud service j ($ per hour)
$\alpha_{j\omega}$	Customers' demand (hours) for cloud service j under scenario ω
θ_i	Equivalent ratio of compute instance i
π_ω	Probability of scenario ω
L	Length of subscription for one base unit (hours)
x	Number of base units
$y_{ij\omega}$	Number of hours utilized from base units for service j under scenario ω
$z_{ij\omega}$	Number of hours of compute instance i purchased with on-demand price for service j under scenario ω

I. INTRODUCTION

Windows Azure is a cloud computing platform of Microsoft where cloud computing users can host applications [1], [2]. Windows Azure provides different classes of compute instances to meet users' requirements. A compute instance is comparable to a server bundled with a certain scalable platform. In particular, Windows Azure deploys a specialized operating system for the platform.

Windows Azure provides two purchasing options, namely pay-as-you-go (i.e., on-demand option) and 6-month subscription plan (i.e., subscription option).[1] The on-demand option can be purchased without any commitment. Thus, a compute instance can be dynamically provisioned at any moment that the instance is needed. In contrast, the subscription option needs to be purchased in advance with a 6-month subscription term. However, with a cheaper price, the subscription option can greatly reduce the total cost incurred to users for a long-term computing usage.

In this paper, a cloud computing market is studied where a cloud provider rents compute instances from Windows Azure. Then, value-added services can be built on the rented compute instances. The cloud provider can earn a profit by selling the cloud services to customers. Cloud services could be video streaming, online game, MapReduce platform, web/application hosting, and financial analysis services. Thus, the cloud computing model of the cloud provider can be software-as-a-service (SaaS) and/or platform-as-a-service (PaaS) [2].

Although the cloud provider can increase the profit by purchasing the subscription option, the optimal number of subscribed compute instances is not trivial to be known. In other words, to gain the maximum profit can be a major challenge. The reason is that the customers' demand is not precisely known in advance. To deal with this demand uncertainty, this paper proposes a profit maximization model based on stochastic programming with two-stage recourse [6]. With the uncertainty, the proposed model can be used to obtain the number of subscribed compute instances such that the expected profit is maximized. The numerical studies are performed to evaluate the model. The results show that the model can effectively maximize the profit under the demand uncertainty.

II. RELATED WORK

In [3], a method based on integer programming was proposed to rent resources located in cloud computing. An algorithm to rent additional servers in cloud computing was proposed in [4] to accommodate workloads in a local cluster. In [5], an auto-scaling method was proposed and evaluated in Windows Azure. The method applies integer programming for resizing the number of instances purchased with the on-demand option. The method did not consider the subscription option which can significantly reduce the cost. The methods in [3]–[5] cannot guarantee the optimal solution under demand uncertainty.

To deal with the uncertainty, a stochastic programming model [6] for renting resources in cloud computing was studied in [7]. The objective of the model is to minimize the expected resource provisioning costs incurred to customers. In contrast, to mainly focus on a cloud provider's perspective, this paper applies stochastic programming to maximize the

[1] The detail of Windows Azure mentioned in this paper is based on that in [1] with the latest update on February 6, 2012.

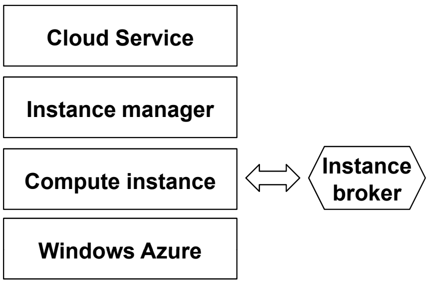


Fig. 1: System model of the cloud provider's computational service.

provider's profit. In addition, this paper also proposes a system model based on Windows Azure which is an efficient cloud computing platform [8] for the cloud provider to build value-added cloud services for customers.

III. Proposed Profit Maximization Model

A. System Model and Assumption

The architectural system model of a cloud provider's computational service is depicted in Fig. 1. The model consists of four layers, namely cloud service, instance manager, compute instance, and Windows Azure. The top three layers are manageable by the cloud provider whereas the bottom layer is operated by Microsoft. The cloud service layer represents a set of cloud services (i.e., set $\mathcal{J}$) offered by the cloud provider to the customers. The cloud provider rents computational resources from Windows Azure to operate the cloud services. To simplify the model, it is assumed that only CPU time of compute instances is the only one type of resource considered in the model. Thus, costs incurred by other resources and services (e.g., storage, service bus, caching, SQL Azure, the Internet traffic for the data transfer, etc.) are ignored. It is assumed that all compute instances are installed with a set of software required by all cloud services.

To efficiently manage compute instances, the instance manager layer offers main functions which are assumed to be available, for example, distributed scheduling, load balancing, accounting and billing, and monitoring functions. The compute instance layer provides a pool of compute instances rented from Windows Azure. Windows Azure layer supplies the platform to host compute instances.

Windows Azure provides 5 classes of compute instances (i.e., set $\mathcal{I}$) as shown in Table I. Each type features different (virtual) hardware specification and on-demand price. The 6-month subscription and on-demand options are both considered in the model. Although the subscription option is the 6-month term, the model determines the option in a month-by-month basis. Let the monthly subscription length (i.e., L) be 750 hours [1]. It is assumed that Windows Azure does not limit the number of hours that the cloud provider can purchase, and the rented compute instances are always available to the cloud provider.

The subscription option includes 750 hours of Small instance, i.e., 1 *base unit* [1]. Users can increase the number of base units. The monthly subscription is \$71.99 per base unit (i.e., F). The 750-hour size of a base unit can be converted to the number of hours for any compute instances. Windows Azure defines the *equivalent ratio* denoted by θ_i for the conversion as presented in Table I. This equivalent ratio is useful for selecting appropriate compute instances to efficiently accommodate cloud services. In this model, for each cloud service, the customers' demand represents the number of hours (i.e., $\alpha_{j\omega}$) which will spend in Small instance.

TABLE I: Compute instances offered by Windows Azure.

Instance	Specification (CPU / RAM / storage)	\$ per hour	θ_i
Extra Small	1 GHz / 768 MB / 20 GB	0.04	1
Small	1.6 GHz / 1.75 GB / 225 GB	0.12	1
Medium	2 x 1.6 GHz / 3.5 GB / 490 GB	0.24	2
Large	4 x 1.6 GHz / 7 GB / 1,000 GB	0.48	4
Extra Large	8 x 1.6 GHz / 14 GB/ 2,040 GB	0.96	8

In Fig. 1, the instance broker (IB) is available in the compute instance layer. IB is responsible for making a decision to purchase compute instances. The main contribution in this paper is to develop an optimization model for IB.

The decision of IB consists of 3 decision variables, i.e., x, $y_{ij\omega}$, and $z_{ij\omega}$. Variable x denotes the number of base units of the subscription option which needs to be purchased in advance. To deal with the demand uncertainty, variables $y_{ij\omega}$ and $z_{ij\omega}$ are considered as *recourse actions*. According to observed *scenario* $\omega \in \Omega$, the recourse actions state the number of hours to be taken from base units (i.e., $y_{ij\omega}$) and also the number of hours to be purchased with the on-demand option (i.e., $z_{ij\omega}$). A scenario represents possible demand (i.e., $\alpha_{j\omega}$). Let Ω_j denote the set of scenarios of demand for cloud service j. A multivariate set of scenarios for every cloud service can be obtained through the Cartesian product, namely $\Omega = \prod_{j\in\mathcal{J}} \Omega_j$.

Finally, scenario $\omega \in \Omega$ can be represented as a random vector, i.e., $\xi(\omega) = \big(\alpha_{ij_1\omega}, \alpha_{ij_2\omega}, \ldots, \alpha_{i|\mathcal{J}|\omega}\big)$. It is assumed that the discrete probability distribution of Ω associated with respective probabilities (i.e., π_ω) is available in the model.

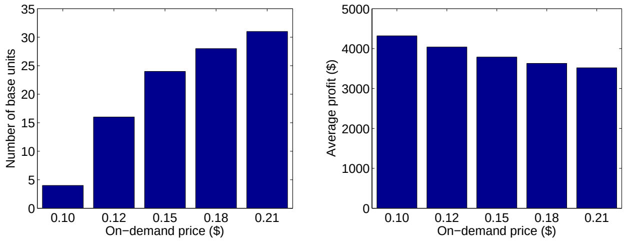
B. Stochastic Optimization Model

This paper proposes a stochastic programming model with two-stage recourse [6] for the instance broker, namely

$$\begin{aligned}
\text{Maximize:} \quad & \mathbb{E}\big[\mathscr{Q}[x,\omega]\big] - F\,x && (1)\\
\text{subject to:} \quad & x \in \{0,1,\ldots\} && (2)\\
& y_{ij\omega}, z_{ij\omega} \geq 0, \forall i \in \mathcal{I}, j \in \mathcal{J}, \omega \in \Omega && (3)\\
& \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \theta_i y_{ij\omega} \leq Lx, \forall \omega \in \Omega && (4)\\
& \alpha_{j\omega} = \sum_{i\in\mathcal{I}} \theta_i \big[y_{ij\omega} + z_{ij\omega}\big],\\
& \qquad \forall j \in \mathcal{J}, \omega \in \Omega. && (5)
\end{aligned}$$

In (1), the objective function is to maximize the cloud provider's profit. $\mathbb{E}[\cdot]$ denotes the expected value of profits incurred by every scenario $\omega \in \Omega$ where function $\mathscr{Q}(x,\omega)$ denotes the maximization problem, given the value of variable x and scenario ω, as defined as follows:

$$\mathscr{Q}[x,\omega] = \max_{y_{ij\omega}, z_{ij\omega}} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \Big[P_j\ y_{ij\omega} + (P_j - C_i) z_{ij\omega}\Big]. \quad (6)$$



(a) Number of base units

(b) Average profit

Fig. 2: Impact of on-demand price.

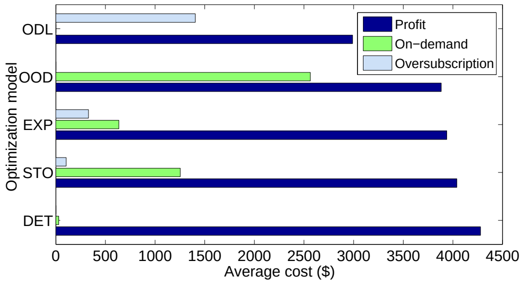


Fig. 3: Cost comparison among different models.

Suppose Ω has *finite support*. That is, Ω has the finite number of scenarios, and each scenario ω is described by respective probability, i.e., $0 \leq \pi_\omega \leq 1$ and $\sum_{\omega\in\Omega}\pi_\omega = 1$. Then, the stochastic programming model in (1)-(5) can be transformed into a deterministic equivalent model which is a mixed integer linear programming model. That is, given a probability distribution of Ω, $\mathbb{E}\big[\mathscr{Q}[x,\omega]\big]$ in (1) can be redefined as follows:

$$\mathbb{E}\big[\mathscr{Q}[x,\omega]\big] = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{\omega\in\Omega}\pi_\omega\Big[P_j\ y_{ij\omega} + (P_j - C_i)z_{ij\omega}\Big]. \quad (7)$$

The model consists of the following constraints. Constraints in (2) and (3) indicate that the variables take the values from the sets of non-negative integers and non-negative real numbers, respectively. Constraint in (4) controls the amount of utilizable hours of base units. Constraint in (5) states that the number of hours offered by the cloud provider has to meet the customers' demand.

The proposed model can be implemented and efficiently solved with a traditional optimization solver, e.g., GNU Linear Programming Kit and IBM ILOG CPLEX Optimizer.

IV. Performance Evaluation

A. Parameter Setting

To evaluate the performance of the optimization model derived in (1)-(5), the proposed model and other compared models are implemented and solved by GAMS/CPLEX [9].

For experimental parameters, the actual prices to rent compute instances in Windows Azure are applied. To simplify the experiment, only Small instance is used in the evaluation. Two cloud services are evaluated, namely J_1 (e.g., web hosting service) and J_2 (e.g., application hosting service). The selling prices for J_1 and J_2 are $0.20 and $0.40 per hour, respectively. Let the demand (i.e., $\alpha_{j\omega}$) for the cloud service vary in the interval [1000, 16000] hours, and 16 scenarios are considered for each demand (i.e., $|\Omega| = 256$ scenarios). The discrete probability distributions of demand for J_1 and J_2 are assumed to be uniform and exponential distributions, respectively.

B. Numerical Studies

1) Impact of on-demand prices: The impact of on-demand prices on the decision of the proposed model is investigated. It is assumed that the on-demand price can be later adjusted by Microsoft without noticing the cloud provider in advance, while the subscription fee and selling prices of cloud services are fixed. The on-demand price is varied in the interval [$0.10, $0.21]. As shown in Fig. 2(a), the increment of on-demand price results in subscribing more number of base units. Since the on-demand price is more expensive, the model purchases the fixed-price subscription option. In contrast, the higher on-demand price clearly decreases the profit as shown in Fig. 2(b). Note that the average profit mentioned in this experiment is calculated by a developed simulation. That is, the simulation generates a number of possible scenarios, and then an average profit is obtained given the profit under generated scenarios.

2) Comparison among different models: Next, the different profit maximization models are evaluated, namely the proposed stochastic programming (STO) derived in (1)-(5), deterministic-demand (DET), only-on-demand (OOD), on-demand-less (ODL), and expected-value of uncertainty (EVU) models. DET is a deterministic optimization model in which the demand is assumed to be precisely known in advance. OOD and ODL are considered as deterministic optimization models as well. OOD can instantly apply the on-demand option at the moment when a scenario is observed. Hence, OOD does not require the subscription option. In contrast, ODL solely determines the worst-case scenario to hedge peak demand by applying the subscription option without later purchasing the on-demand option. EVU is a well-known optimization model based on Jensen's inequality for dealing with the uncertainty [12]. To address the demand uncertainty issue, EVU uses the expected-value of demand of each cloud service as the fixed demand. Then, a deterministic linear programming model can be formulated and solved given the expected-value of demand.

In Fig. 3, the average profit, on-demand cost (i.e., cost of purchased on-demand option), and oversubscription cost (i.e., cost of unused hours of base units) incurred by each compared model are presented. Again, the simulation is developed to obtain the average costs given a set of generated possible scenarios. Clearly, DET yields the best solution, since the demand applied in DET is assumed to be perfectly known. DET yields the highest profit; however, DET is not applicable when the demand uncertainty is commonly involved. ODL yields the least profit, since ODL oversubscribes base units to completely avoid the on-demand option. Hence, the oversubscription cost is the highest, and such a cost decreases the profit margin. Since OOD solely purchases the more expensive on-demand option, the yielded profit is poor. EVU performs well in which both on-demand and oversubscription costs greatly decrease. However, the profit of EVU (i.e., \$3,938.17) is still less than that of STO (i.e., \$4,040.05) since EVU considers only the expectation of demand. Theoretically, EVU yields the worse solution than that of STO [12]. In particular, STO takes the probability distribution of all scenarios into the optimization model. STO yields the highest profit when the demand uncertainty is regarded.

TABLE II: Cost comparison among different numbers of base units.

#	\$Subscribe	\$On-demand	\$Over	\$Profit
4	\$287.96	\$2,206.28	\$2.15	\$3,951.07
12	\$863.88	\$1,549.33	\$52.58	\$4,032.11
15	\$1,079.85	\$1,325.43	\$89.45	\$4,040.04
16	**\$1,151.84**	**\$1,253.43**	**\$103.85**	**\$4,040.05**
17	\$1,151.84	\$1,184.09	\$120.38	\$4,037.39
20	\$1,439.80	\$984.65	\$176.82	\$4,020.86
30	\$2,159.70	\$440.61	\$461.55	\$3,845.00

As presented in Table II, other different solutions given different numbers of base units (shown in the first column) are compared with STO as well. A simulation is developed to generate scenarios and evaluate costs incurred by purchasing the fixed number of base units. Different average costs are evaluated (as shown in the column headers), i.e., subscription (\$Subscribe), on-demand (\$On-demand), oversubscription (\$Over), and profit (\$Profit) costs. In Table II, the subscription of 16 base units (as highlighted) is the same solution as that of STO (i.e., optimal solution). It is observed that a solution of the number of base units close to 16 has the average costs converging to that of STO.

Although the simulation (i.e., brute-force search) used in this experiment can be applied to obtain the optimal number of base units, the computational complexity of the simulation is higher than that of STO solved by GAMS/CPLEX. That is, the simulation performs several iterations to evaluate the numbers of base units given a set of generated scenarios. With the larger numbers of scenarios (i.e., $|\Omega|$) and cloud services (i.e., $|\mathcal{J}|$), the simulation could take a longer time to evaluate the candidate numbers of base units. In terms of the computational performance, STO performs well. For this parameter setting, the total execution time for solving STO by GAMS/CPLEX is less than one second, while the simulation takes longer than a few ten seconds on the test machine with 2.93 GHz quad-core processor and 4 GB of RAM. The small number of parameters used in the experiment might be the main reason that the computational time of STO is very small. The scalability of STO given larger problem sizes will be investigated in the future work.

V. Conclusion

The cloud computing market has been studied in this paper where the cloud provider builds cloud services on Windows Azure platform operated by Microsoft. To maximize the cloud provider's profit, the paper has considered the customers' demand uncertainty by formulating and solving the stochastic programming model with two-stage recourse. The results show that the proposed model can maximize the profit under the demand fluctuation. Other than Windows Azure, this optimization model could be modified and applied to other similar cloud computing platforms as well, e.g., Amazon EC2 [10] and GoGrid [11].

This paper is the preliminary study of a trading mechanism framework for (small and medium) cloud providers that sell cloud services hosted in other large cloud providers' platforms. For the future work, the complete framework will be proposed. That is, a stochastic programming with multi-stage recourse [7] will be derived to maximize the profit for multiple time stages, rather than the 2 stages addressed in this paper. In addition, the game theory approaches addressed in [13] will be applied to the framework to define appropriate selling prices of cloud services which can maximize the profit based on the supply-and-demand volume of cloud services. Finally, the framework will be practically implemented and deployed in a real cloud computing market.

Acknowledgment

This work was done in the Parallel and Distributed Computing Centre at the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. This work was supported by the projects "User and Domain Driven Data Analytics" and "Design and Analysis of Cloud Computing for Data Value Chain: Operation Research Approach", granted by A*STAR Thematic Strategic Research Programme [14].

References

[1] Windows Azure, http://www.windowsazure.com/

[2] B. P. Rimal, E. Choi, and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems," in *Fifth Int. Joint Conference on INC, IMS and IDC*, pp. 44-51, Aug. 2009.

[3] R. Aoun, E. A. Doumith and M. Gagnaire, "Resource Provisioning for Enriched Services in Cloud Environment," in *IEEE Second Int. Conference on Cloud Computing Technology and Science*, pp. 296-303, 2010.

[4] M. Mattess, C. Vecchiola, and R. Buyya, "Managing Peak Loads by Leasing Cloud Infrastructure Services from a Spot Market," in *12th IEEE Int. Conference on High Performance Computing and Communications*, pp. 180-188, Sept. 2010.

[5] M. Mao, J. Li, and M. Humphrey, "Cloud Auto-scaling with Deadline and Budget Constraints," in *IEEE/ACM Int. Conference on Grid Computing*, pp. 41-48, Oct. 2010.

[6] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag Newyork, Inc., 1997.

[7] S. Chaisiri, B. S. Lee, and D. Niyato, "Optimization of Resource Provisioning Cost in Cloud Computing," *IEEE Transactions on Services Computing*, 2011.

[8] Z. Hill, J. Li, M. Mao, A. Ruiz-Alvarez, and M. Humphrey, "Early Observations on the Performance of Windows Azure," in *Proc. of the 19th ACM Int. Sym. on High Performance Distributed Computing*, 2010.

[9] GAMS Solvers, http://www.gams.com/solvers/index.htm

[10] Amazon EC2, http://aws.amazon.com/ec2/

[11] GoGrid, http://www.gogrid.com/

[12] J. L. Higle, "Chapter 1: Stochastic Programming: Optimization When Uncertainty Matters," Tutorials in Operations Research, INFORMS, 2005.

[13] D. Niyato, S. Chaisiri, and B. S. Lee, "Economic Analysis of Resource Market in Cloud Computing Environment," in *IEEE Asia-Pacific Services Computing Conference*, pp. 156-162, Dec. 2009.

[14] Thematic Strategic Research Programme, http://www.a-star.edu.sg/astar/Research/FundingOpportunities/GrantsSponsorships/ThematicStrategicResearchProgramme/tabid/247/Default.aspx