



Cloud Service Pricing Strategy Based on Service Level Agreement with Default Compensation

Xing Wan, Fu-zan Chen^(✉), and Min-qiang Li

College of Management and Economics, Tianjin University, Tianjin, China
fzchen@tju.edu.cn

Abstract. The emergence of cloud computing has completely changed the entire IT world. There is no doubt that making a profit is a cloud service provider (CSP)'s ultimate goal. To make a profit, a CSP needs a reasonable pricing strategy. At the same time, in order to increase the attractiveness of cloud service, CSPs tend to write default clauses in the service level agreement (SLA) and the default cost will affect the profit of CSPs. To determine an optimal pricing strategy, this paper builds some bi-level programming models of CSPs and users. The basic model does not consider default compensation and the default cost is included in the extended model. Finally, the comparison results of the two models prove that it is significant for the CSP to set default compensation clauses.

Keywords: Bi-level programming · Cloud service · Default compensation · Pricing strategy

1 Introduction

In the 21st century, cloud services have developed rapidly in terms of technical services and market size. The global scale has increased from \$68.3 billion in 2010 to \$305.8 billion in 2018. According to Gartner's forecast, by 2020 this figure will exceed \$400 billion.

Hence, it is crucial for a cloud computing providers (CSP) to make a reasonable pricing strategy for cloud computing services. The configuration of cloud service resources is an important factor of pricing [1]. It is because resource allocation directly affects the operating cost. In addition, relevant research indicates that it is more reasonable to adopt a dynamic and tiered pricing strategy [2].

To better coordinate the transaction cooperation between cloud service providers, CSPs usually sign service level agreement (SLA) with users. Generally, SLA includes an agreement on the price of the service, the level of service, and the amount of compensation for breach of contract. The latest version of the Alibaba Cloud Elastic Cloud Service (ECS) Service Level Agreement, which came into force on February 1st, 2018, specifically defines service availability level indicators and compensation plans. One of the clauses is "For a single ECS instance, if the service availability is less than 99.95% but equal to or higher than 99%, the user is compensated for 10% of the monthly service fee."

The default compensation item will inevitably lead to certain default costs, which will thus affect the profit of the CSP. Obviously, this impact is worthy of deep study. The default issue can be treated as whether a transaction is generated [3]. However, using quantitative methods to analyze the cost of default is more realistic.

The main contribution of this paper consists of three parts:

- This paper focuses on the analysis of the default compensation for cloud service pricing. Therefore, this study established two comparative bi-level programming models.
- This paper uses quantitative methods to analyze the cost of default. Furthermore, the response time is defined as the service quality indicator and could be calculated by the queuing theory model.
- This paper clearly provides the optimal pricing formula and optimal resource allocation formula for the basic model. Moreover, the evaluation between the basic model and the extended model shows it is valuable for CSP to set default clauses.

The rest of this paper is organized as follows. This paper discusses related work on cloud service pricing in Sect. 2. The problem description and two comparative bi-level programming models are presented in Sect. 3. The cost function and the user utility function are defined in Sect. 4. In Sect. 4, this paper lists the optimal pricing formula and the optimal resource allocation formula for the basic model and shows the comparison results of the two models. The conclusion is given in Sect. 5.

2 Related Work

In the previous research on cloud service pricing, scholars have established pricing models from different perspectives. Jin *et al.* propose the cloud service pricing optimization scheme on a fine-grained scale [4]. And they calculate the maximum value of the social welfare so that both parties are satisfied. Wang *et al.* established a key parameter determination model including price and service quality index based on multiple game relationships among big data reporting organizations, cloud service organizations, and customers [5]. It can be seen that cloud service pricing is an essential game process between CSPs and users.

For the game problem of cloud service pricing, Cardellini *et al.* propose the optimal pricing strategy based on the relationship between SaaS and IaaS [6]. They studied two different IaaS provider pricing strategies: the first assumes that the IaaS provider sets a unique price; the second assumes that the IaaS provider can set different prices for different customers. For each pricing strategy, the literature confirms the existence of game equilibrium.

Most of the previous related literature use simple linear functions to calculate the cloud service cost. Chiang *et al.* consider the cost of hardware resource, energy consumption, buffer, user opportunity and congestion in the cost efficiency analysis of cloud services [7]. The total cost is a linear combination function for each cost. When studying the cost model of cloud services, Liu *et al.* even consider energy consumption factors, involving physical voltage, capacitance, and frequency. But the total cost is

also a linear function of the decision variable of the number of servers [8]. What utility function should be used depends on the definition of “quality of service” in the research.

There are a variety of existing user utility functions. Sumanta *et al.* construct a linear user utility function in the study of cloud service pricing, which is a linear combination of a vector composed of utility positive factors and a vector composed of negative factors [9]. Tang *et al.* establish a logarithmic user utility function for service quality in their research, based on the law of diminishing marginal utility in economics [10].

Given the problem of cloud service default, most of the previous literature chose response time as a service level indicator to investigate default and adopted a variety of processing methods. Zhang *et al.* use the default condition as a constraint on the model to ensure that the response time of the service could not exceed the time specified in the agreement [3]. Sen *et al.* construct the resource request model of the service system based on the first-come-first-served MMC model [11]. They propose that the probability of completing the request within the time limit should be greater than the probability set in SLA. Macías *et al.* propose that users can pay a certain tip to allow cloud service providers to guarantee the quality of service [12].

3 Bi-level Programming Model for Pricing Cloud Service

This paper studies the optimal pricing strategy for real-time cloud services considering SLA in a monopolistic environment, that is, the game relation between one CSP and multiple users.

CSP takes profit maximization as the decision objective, and its decision variable is price and resource allocation. Users take the utility maximization as the decision objective, and the decision variable is to choose the service or not.

The detailed game process is as follows:

1. In the beginning, the CSP provides the users with service information, including price, promised response time and default compensation clauses.
2. The users decide whether to select the service according to the service information provided and their demand for the urgency of the service. Then, they issue purchase instructions to the supplier.
3. The CSP assigns service queues to users based on their purchase instruction, and the users start to wait in line for the service.
4. The users accept the provider’s cloud service and feedback on the status of the service to the CSP.
5. The CSP judges whether there is a breach of contract on the basis of the users’ feedback. If there is a default, the users will be compensated according to the clauses of default.

According to the problem description above, this paper takes the decision of CSP as the upper decision and the decision of the user as the lower decision.

To strongly prove the value of the compensation clauses set in the SLA, this paper establishes two comparative models. The basic model does not consider default compensation and the default cost is included in the extended model.

In the basic model, CSP aims at maximizing its profit π , taking the price p and the number of configured servers NS as decision variables. When there are n users purchasing the services, the CSP's profit is the part of the revenue from the users minus the operating cost C of the cloud service. The CSP's programming is expressed as (1).

$$\max \pi(NS, p) = pn - C \tag{1}$$

Assuming there are M users in the cloud service market, the utility of user j ($j = 1, 2, \dots, M$) is a function U of the price and promised response time PRT in the SLA. Each user's goal is to maximize their utility u_j . This study assumes that the CSP only provides one service, and users are free to choose whether to purchase the service or not. The users' programming is expressed as (2) and (3).

$$\max u_j(x_j) = U_j(PRT, p)x_j \tag{2}$$

$$\text{s.t. } x_j = 0 \text{ or } 1 \tag{3}$$

Equation (3) means that when x_j is equal to 0, user j does not purchase the service, and when x_j is equal to 1, user j purchases the service. It can be inferred that user j would purchase the service only if the utility function value is greater than zero.

In the basic model of this study, the promised response time PRT is not directly treated as a decisive variable. Since there is no default clause, the CSP should consciously make the promised response time PRT as equal as possible to the actual response time RT of each user. It should be noted that RT is a random variable.

Because it is a real-time cloud service, this paper uses the queuing theory model to analyze the response time of the system. Furthermore, the system can be viewed as an MM1 model in which multiple service stations are connected in parallel [4]. This paper assumes that each server is homogeneous. The service rate μ of the system represents the number of users each server can serve in a unit of time. The arrival rate of the system represents the number of users entering the system per unit time. In this study, the arrival rate can be directly reflected by the occupied market size n . Thus, the promised response time PRT can take the average value of the actual response time RT , which is expressed as (4).

$$PRT = E(RT) = \frac{1}{\mu - \frac{n}{NS}} \tag{4}$$

In the extended model, the CSP needs to consider the issue of default costs DC . The CSP's profit is equal to the revenue minus the operating cost and the default cost. Meanwhile, the promised response time PRT is a decisive variable, not necessarily equal to the average value of the actual response time RT . Anyhow, the value of PRT should not be exaggerated. This paper assumes that the minimum value of PRT is the reciprocal of the service rate μ . So, the CSP's programming in the extended model is expressed as (5) and (6).

$$\max \pi(NS, PRT, p) = pn - C - DC \tag{5}$$

$$\text{s.t. } PRT \geq \frac{1}{\mu} \tag{6}$$

The users' programming in the extended model is the same as that in the basic model, which is expressed as (2) and (3).

In the extended model, whether a default occurs or not depends on whether the actual response time RT exceeds the promised response time PRT . If the actual response time of user j exceeds the promised response time, the CSP needs to provide the user with default compensation and bear the cost.

However, according to the fact and relevant literature, the compensation provided by CSP cannot be infinite. In other words, there should be an upper limit α for compensation. When the actual response time increases to a certain extent, the default cost paid by the CSP does not increase.

Thus, the default function is defined as (7) and (8). And the relationship between default cost of user j and the response time is shown in Fig. 1.

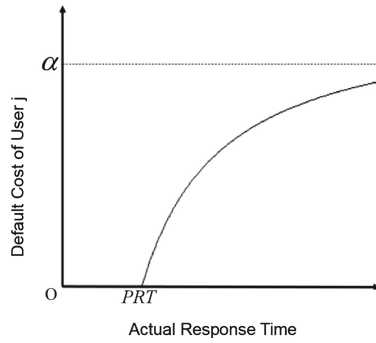


Fig. 1. The relationship between default cost of user j and the response time

$$DC = \sum DC_j \tag{7}$$

$$DC_j(RT) = \begin{cases} 0, & RT \leq PRT \\ \alpha(1 - \frac{PRT}{RT}), & RT > PRT \end{cases} \tag{8}$$

Both in the basic model and the extended model, the CSP needs to consume a lot of power resources, maintain the servers regularly, and pay for relevant personnel. These would make the CSP bear operating costs.

This paper assumes that each cost in the operation is proportional to the number of servers NS . For each server, the operating cost is a . The operating cost function is defined as (9).

$$C = aNS \quad (9)$$

For the lower decision of the models, cloud services with lower prices and better quality are more popular with the user community. In the eyes of users, the smaller the promised response time is, the better quality service would be. So, the utility function is defined as (10).

$$U = v_0 - \delta_j PRT - p \quad (10)$$

In (10), v_0 indicates users' basic utility, which is the same for each user. δ_j represents the sensitivity of user j to the response time. The higher δ_j means the higher time requirements of user j . This paper assumes that δ_j obeys a uniform distribution of zero to a positive number b .

4 Results and Analysis

4.1 Optimal Decision of the Basic Model

User j would purchase the service only if the utility function value is greater than zero and δ_j obeys a uniform distribution of zero to a positive number. Therefore, according to (10), the occupied market size n could be expressed as (11).

$$n = \frac{v_0 - p}{PRT} \quad (11)$$

Then, according to (4) and (11), n is solved as (12).

$$n = \frac{\mu NS(v_0 - p)}{NS + v_0 - p} \quad (12)$$

If the optimal price p^* and the optimal number NS^* of servers exist, the partial derivatives of the profit function with p and NS are both equal to 0.

The second derivative test proves that the optimal price and the optimal must exist. Finally, the optimal number NS^* of servers and the optimal price p^* of the basic model are obtained as (13) and (14).

$$NS^* = (\sqrt{1 + v_0} - 1)(\sqrt{a\mu\sqrt{1 + v_0}(\sqrt{1 + v_0} - 1)} - 1) \quad (13)$$

$$p^* = NS^* + v_0 - \sqrt{NS^*(NS^* + v_0)} \quad (14)$$

4.2 Comparative Evaluation

Based on the queuing theory, the probability density function of the customer's stay time is expressed as (15).

$$f_{RT}(t) = \begin{cases} (\mu - \frac{n}{NS})e^{-(\mu - \frac{n}{NS})t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{15}$$

In terms of probability theory, the calculation method of the default cost expectation of user j in the extended model is expressed as (16).

$$E(DC_j) = \int_{-\infty}^{+\infty} f_{RT}(t) \cdot DC_j(t) dt \tag{16}$$

Then, the profit function of the extended model is transformed into (17).

$$\pi = pn - aNS - \alpha n \int_{PRT}^{\infty} (\mu - \frac{n}{NS})e^{-(\mu - \frac{n}{NS})t} (1 - \frac{PRT}{t}) dt \tag{17}$$

Equation (17) is not an elementary function. Therefore, this paper has to use the numerical method to solve the extended model. The parameter settings are listed in Table 1.

Table 1. Parameter settings

Parameter	Value
μ	2
a	1
v_0	3

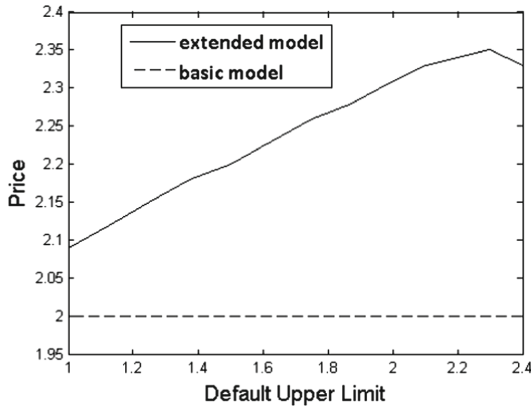


Fig. 2. The optimal price in the cases of different values of the default upper limit α

Next, the enumeration method is used to search and evaluate the optimal solution of the extended model with the default upper limit α taking multiple values.

Finally, the parameters are substituted into (13) and (14) to evaluate the basic model, and the results of the two models are compared and analyzed.

Figures 2, 3 and 4 show the optimal solution of the basic model and the extended model when α takes different values. Figures 5 and 6 present the occupied market size, cost and profit when the decisive variables are optimal.

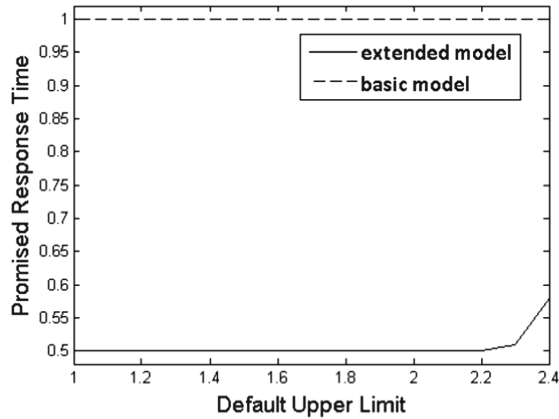


Fig. 3. The promised response time in the cases of different values of the default upper limit α

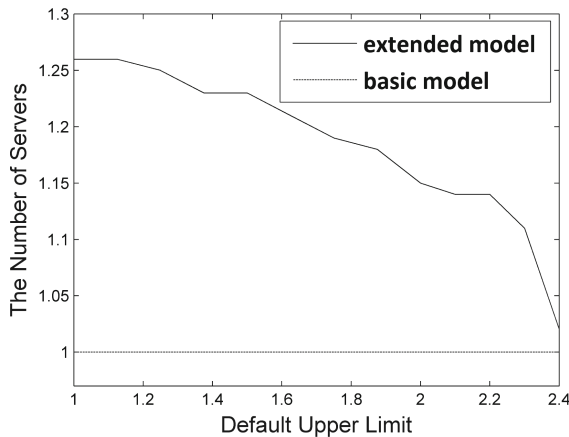


Fig. 4. The number of serves in the cases of different values of the default upper limit α

When α is lower than 1.9 (i.e. less than 82% of the corresponding price), the profit in the extended model is higher than the profit of the base model. In this case, the price and the occupied market size in the extended model are both greater than those in the basic model, though the number of servers (the operating cost) and the default cost are also more.

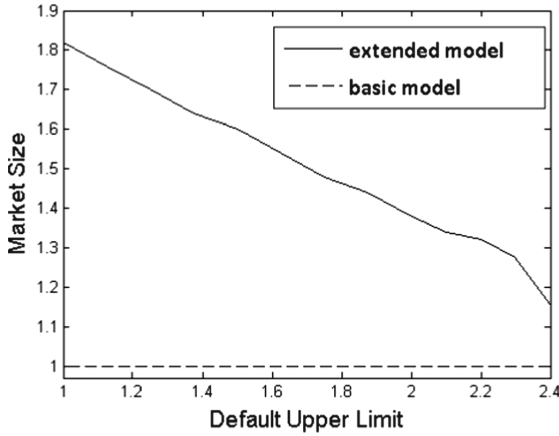


Fig. 5. The occupied market size when the decisive variables are optimal in the cases of different values of the default upper limit α

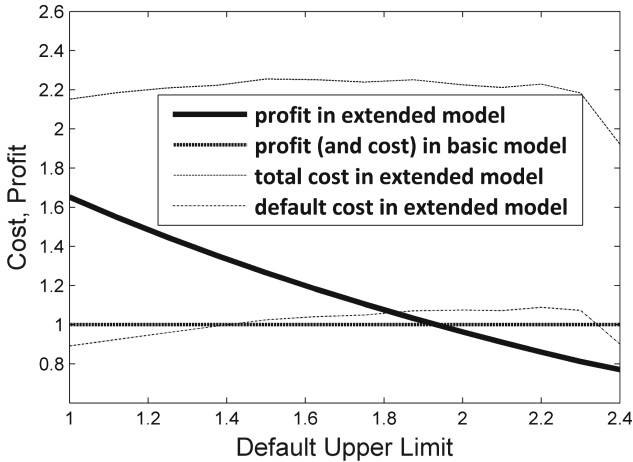


Fig. 6. The profit when the decisive variables are optimal in the cases of different values of the default upper limit α

When α is higher than 1.9 (i.e. more than 82% of the corresponding price), the profit in the extended model is less. However, due to the very short *PRT*, the CSP still occupies a greater market size in the extended model. At worst, the market size in the extended model has increased by 15%, from 1 to 1.15.

According to the results, it is significant for the CSP to set default compensation clauses in SLA. Such a strategy for cloud services is beneficial for the CSP to expand its market size, and this strategy would even give the CSP a chance to increase its profit. When the default upper limit is low, if the CSP adopts the default compensation

strategy, it would gain more profit. It is because the price and the occupied market size get improved obviously, whereas the costs only increase slightly. When the default upper limit is high, the default compensation strategy still makes the CSP occupy a larger market size thanks to the adjustable promised response time.

5 Conclusion and Future Work

This article studies the optimal pricing strategy for real-time cloud services considering SLA with default compensation in a monopolistic environment.

In the model section, the article establishes two comparative bi-level programming models (the basic model and the extended model) based on the game relation between one CSP and multiple users. The basic model does not consider default compensation and the default cost is included in the extended model. Simultaneously, this paper provides the default cost function, the operating cost function, and the utility function for the models. For the default cost function, this study chooses the response time as the service quality index, which is calculated by the queuing theory model.

In the analysis section, the article first analyzes the basic model by using analytical methods to figure out the optimal pricing formula and optimal resource allocation formula. Then, the article compares the evaluation between the basic model and the extended model. The results show that it is better for the CSP to set default compensation clauses. The default compensation strategy gives CSP a great opportunity to increase its profit. Even though the profit level does not rise, the market size would expand.

In the future, we plan to study the tiered pricing strategy for cloud services. Also, we are considering adding the expectation of default cost to the utility function.

Acknowledgment. The work was supported by the general program of national natural science foundation of China (No. 71771169) and the key program of national natural science foundation of China (No. 71631003).

References

1. Tanaka, M., Murakami, Y.: Strategy-proof pricing for cloud service composition. *IEEE Trans. Cloud Comput.* **4**(3), 363–375 (2016)
2. Pan, W., Yu, L., Wang, S.: Dynamic pricing strategy of provider with different QoS levels in web service. *J. Netw.* **4**(4), 228–235 (2009)
3. Zhang, Z., Tan, Y., Dey, D.: Price competition with service level guarantee in web services. *Decision Support Systems* **47**(2), 93–104 (2009)
4. Jin, H., Wang, X., Wu, S., Shi, X.: Towards optimized fine-grained pricing of IaaS cloud platform. *IEEE Trans. Cloud Comput.* **3**(4), 436–448 (2015)
5. Wang, J., Liu, A., Zhang, S.: Key parameters decision for cloud computing: insights from a multiple game model. *Concurrency Comput. Pract. Experience* **29**(7), e4200 (2017)
6. Cardellini, V., Valerio, V.D., Presti, F.L.: Game-theoretic resource pricing and provisioning strategies in cloud systems. *IEEE Trans. Services Comput.* **99**, 1 (2016)

7. Chiang, Y., Ouyang, Y., Hsu, C.: Performance and cost-effectiveness analyses for cloud services based on rejected and impatient users. *IEEE Transactions on Services Computing* **9** (3), 446–455 (2016)
8. Liu, C., Li, K., Li, K., Buyya, R.: A new cloud service mechanism for profit optimizations of a cloud provider and its users. *IEEE Trans. Cloud Comput.* **99**, 1 (2017)
9. Sumanta, B., Soumyakanti, C., Megha, S.: Pricing cloud services—the impact of broadband quality. *Omega* **50**, 96–114 (2015)
10. Tang, L., Chen, H.: Joint pricing and capacity planning in the IaaS cloud market. *IEEE Trans. Cloud Comput.* **99**, 57–70 (2017)
11. Sen, S., Raghu, T.S., Vinze, A.: Demand heterogeneity in it infrastructure services: modeling and evaluation of a dynamic approach to defining service levels. *Inf. Syst. Res.* **20**(2), 258–276 (2009)
12. Macías, M., Guitart, J.: SLA negotiation and enforcement policies for revenue maximization and client classification in cloud providers. *Future Gener. Comput. Syst.* **41**, 19–31 (2014)