Should Opportunists Be Encouraged? Optimal Decisions in Hybrid Cloud Service Systems

Sheng Zhu, Jinting Wang, and Wei Wayne Li, *Senior Member, IEEE*

***Abstract*—This paper investigates a hybrid service system with a cloud server and an in-house server. We consider two different scenarios: a hybrid service system with orbit space and a hybrid service system without orbit space. In the hybrid service system with orbit space, customers who fail to enter the cloud server can choose to join the in-house subsystem or to enter an orbit space and retry the cloud server. An admission control mechanism based on queue-length limitation is adopted to adjust whether the cloud service resources are open to customers. When the cloud server cannot be accessed immediately, some customers send their jobs to the in-house subsystem, while others (called opportunists) try to send their jobs to the cloud server again. We obtain the optimal queue-length limitation for a given retrial rate. The service provider and customers are different stakeholders, and their market forces are also different. Therefore, it is more realistic to explore the game relationship between them by using dynamic game theory. We can also explore the joint optimums of the queue-length limitation and the retrial rate in the framework of the Stackelberg game. Finally, by comparing with the hybrid service system without orbit space, we discuss the significance of the existence of orbit space, and gain management insights. It is found that the existence of opportunists may benefit the service provider, although they significantly harm social interests, regardless of whether they are cooperative or non-cooperative; therefore, opportunists are encouraged in some situations. Numerical analysis shows that adding a retrial orbit to a hybrid cloud service system with certain input parameters may even more than triple the service provider's revenue.**

Key words: Cloud service; Queueing-game; Optimal decision; hybrid service system; Stackelberg game.

I. Introduction

The cloud provides people with low-delay services, but using the cloud also leads to a variety of security risks (listed by Brender and Markov [1]), such as information security, data location, and so on. In March 2015, while repairing the XEN bug, cloud service providers such as Amazon AWS, IBM SoftLayer, Amazon Linode, and Rackspace suffered multiple host restarts. Amazon AWS has to suspend nearly 10% of the business of its cloud hosts. Security accidents have made potential users very cautious about the cloud service. According to RightScale survey data, although 88% of enterprises use the public cloud, 68% of them run less than 20% of their enterprise applications in the cloud. Specially, for risk-sensitive enterprises, they always send their enterprise applications to a mixed service system including cloud server from external providers and their traditional servers (called *in-house servers* in this paper).

In real-life situations, web service companies usually provide computing resources for customers. Considering the security risk and the efficiency of cloud service, web service companies often adopt a mixed service system with a cloud server and an in-house server, in which customers who fail to enter the cloud server can choose to join the in-house subsystem or to enter an orbit space and retry the cloud server. The cloud service resources are purchased from external cloud service providers. It is generally agreed in the contract that the arrival rate of jobs sent to the cloud shall not exceed a fixed value. The web service companies must ensure that the contract is not breached by limiting the effective arrival rate, but customers always hope to access the cloud and to be served at full speed. Therefore, an admission control needs to be designed by the web service companies. Due to the admission control, a portion of customers are diverted to the in-house subsystem. However, some customers are opportunistic; if the cloud server cannot be immediately accessed, they prefer to suspend their jobs that remain in an orbit space, where service requests do not require sorting and system management, and only necessary storage space needs to be provided. After a while, they try again at a certain retrial rate to enter the cloud. In this paper, we consider the optimal admission control, the joint optimal decisions of the web service company and customers in the framework of a dynamic game theory, and the rationality of the existence of opportunists.

The hybrid service system proposed in this paper is similar to a two-tier service system. Tuohy et al. [16] first introduced the two-tier service system, and recent works on this topic include Guo, Lindsey and Zhang [7], Hua, Chen and Zhang [10], among others. Different from previous literature, we study a two-tier service system with an orbit space. Our model is the same as Zhu, Wang and Li [20], in which they focused on customers' equilibrium strategies and a socially optimal retrial rate. In this paper, we study optimal admission control, joint optimal decisions, and the rationality of the existence of opportunists.

In a pioneering study on admission control, Spencer et al. [15] believed that the total arrival rate of jobs sent to the

Corresponding author: Jinting Wang (e-mail: jtwang@cufe.edu.cn).

This material is partially based upon work supported by the National Natural Science Foundation of China under Grant Nos. 72371259, 71871008 and the Emerging Interdisciplinary Project of CUFE (21XXJC010) to Jinting Wang; by the Fundamental Research Funds for the Universities of Henan Province (Grant No. NSFRF220442), the Project of Establishing the "Double First-Class" Discipline of Surveying and Mapping Science and Technology (Grant No. GCCRC202307) to Sheng Zhu; and by the NSF under Grant Nos. 2302469 and 2318662, and by NASA under Grant No. 80NSSC22KM0052 to Wei Li.

Sheng Zhu is with School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, 454003, China (e-mail: shengzhu_ms@sina.com).

Jinting Wang is with the Department of Management Sciences, School of Management Science and Engineering, Central University of Finance and Economics, Beijing, 100081, China (e-mail: jtwang@cufe.edu.cn).

Wei Wayne Li is with the Department of Computer Science and the NSF Center for Research on Complex Networks, Texas Southern University, Houston, TX 77004, USA (e-mail: liww@tsu.edu).

cloud server varies with the queue-length capacity of the in-house subsystem, and that the arrival rate can be controlled by limiting the capacity of the in-house subsystem. This study explored the impact of the amount of information about the future of queue-length on the system's performance. In our work, we adopt the same admission control, which will be detailed in Section II. However, we focus on the optimal admission control for customers. Further, we consider the optimal queue-length limitation between customers and the web service company in terms of dynamic game theory. To the best of our knowledge, this paper is the first to study the optimal admission control of a system with a cloud server and an in-house server from the viewpoint of dynamic game theory. Interested readers can refer to Xu [18] and Xu and Chan [19] for more details about admission control.

In addition, customers are assumed to be strategic. They decide whether to join the in-house subsystem or to enter the orbit when the cloud cannot be immediately accessed due to the admission control. Therefore, our discussion on all topics cannot ignore customers' equilibrium strategies. Fortunately, it has been derived in Zhu et al. [20]. Interesting readers can refer to Naor [13], Burnetas and Economou [2], Economou and Manou [4], Engel and Hassin [5], Hassin and Haviv [8], Hassin and Snitkovsky [9], Manou, Economou and Karaesmen [12], Guo and Hassin [6], Shi and Lian [14], Wang and Zhang [17] for more details on Nash equilibrium strategy.

Customers and the web service company represent different interest groups. The interaction between them is an interesting problem. The game between them can be characterized as the Stackelberg game. Our use of dynamic game theory is inspired by Caldentey and Wein [3], which discussed the two-stage supply chain based on the Stackelberg game. Li et al. [11] also considered the Stackelberg game problem between mobile devices and edge cloud servers, and proved the existence of a Stackelberg equilibrium in the game. Different from the above works, we consider a hybrid service system that can be modeled as a queueing system with two servers and one retrial orbit, and we explore the economic phenomena based on the dynamic game. In order to obtain the optimal strategy, we need to analyze it by combining the queueing game and the Stackelberg game. It is found that the web service company may reap certain benefits from the strategically speculative behavior of the opportunists.

The main contributions of this paper are listed as follows:

- **Optimal admission control**. In real-life situations, web service companies maximize their interests by choosing the appropriate admission control. In this paper, we derive the optimal queue-length limitation in the hybrid service system with orbit space, and explore the relation between the expected total net benefit of the web service company and VPC (defined in Section II).
- **Joint optimum in dynamic game**. Customers and the web service company represent different interest groups. Based on the dynamic game between them, we develop the joint optimums of the queue-length limitation and the retrial rate within the Stackelberg game formulation, and provide the computational algorithm of the joint optimums.
- **Practical significance of the retrial orbit**. The practical significance of the retrial orbit in the system is discussed. We find under certain conditions that the existence of the retrial orbit is beneficial to the web service company, although it harms broader social interests regardless of whether customers are cooperative or non-cooperative.

The paper is organized as follows. In Section II, we provide a detailed description. Section III studies the optimal queue-length limitation given an exact retrial rate from the viewpoint of the web service company. In Section IV, we explore the joint optimums of the queue-length limitation and the retrial rate based on the Stackelberg game. Section V considers whether the web service company should permit the existence of an orbit space. Numerical analysis is provided in Section VI. Finally, conclusions are offered in Section VII.

II. Model Description and Preliminaries

We consider a hybrid service system with a cloud server and an in-house server. According to whether the system has orbit space, we divide the system into two categories: hybrid service system with orbit space (see Figure 1) and hybrid service system without orbit space (see Figure 2). The hybrid service system with orbit space is the same as the model studied by Zhu and Wang [20]. There is no difference between the hybrid service system without orbit space and the hybrid system with orbit space, except for the existence of orbit space. In order to make the paper readable, we give a brief description of the hybrid service system with orbit space.

In a hybrid service system with orbit space, the web service company provides customers with two kinds of service resources: the in-house subsystem and the cloud. The in-house subsystem is the service resource of the web service company itself, but the cloud service is that purchased by the web service company from an external cloud computing provider. Customers bring their jobs to the web service company at random and ask for service from the company. The arrivals of customers are assumed to follow a Poisson process with intensity λ. The web service provider signs a long-term cloud service contract with the external cloud computing provider, which specifies that the total arrival rate of jobs sent to the cloud within the contract time cannot exceed the stated value. VPC is the abbreviation for the value prescribed by the contract, denoted by γ. Hence, the total arrival rate of jobs sent to the cloud server cannot exceed γ. Upon the arrivals of customers, the web service company diverts their jobs through the admission controller. If the queue-length in the in-house subsystem is less than the given queue-length limitation of L, their jobs are shunted to the in-house subsystem. Once the queue-length reaches L, the arriving jobs will be sent to the cloud until the queue-length in the subsystem is lower than L again. The service rate of jobs in the in-house subsystem is assumed to be μ.

In this paper, the service time does not include the propagation delay, only considering the time from being served by the server to the end of the service. In fact, our proposed method is also feasible for considering the propagation delay scenario. The reason is listed as follows. The propagation delay depends on the specific network connection and can be seen as a specific value, so the corresponding propagation

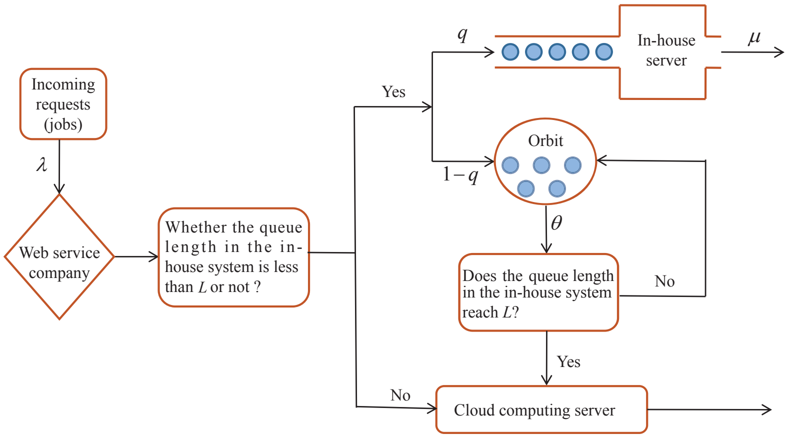

Incoming requests (jobs)
λ
Web service company
Whether the queue length in the in-house system is less than L or not ?
Yes
No
q
In-house server
μ
$1-q$
Orbit
θ
Does the queue length in the in-house system reach L?
No
Yes
Cloud computing server

Fig. 1. Illustration of the hybrid service system with orbit space (see Zhu et al. [20]).

delay cost is also a constant, denoted by C_{pd}. In this paper, the reward after being completed by the cloud is R. When considering the propagation delay situation, we only need to replace the reward after being completed by the cloud, R, with $R-C_{pd}$. In this paper, we do not consider the service times of service requests in the cloud computing center. This is because in this paper, we assume that the cloud computing capabilities are much stronger than the in-house server. Compared to the in-house server, the service time of the cloud can be almost ignored. Mathematically, we can achieve a service time of approximately zero for the cloud through standardization. In addition, in the work of Zhu et al. [20], we have also explained in detail that if the service time of the cloud is not close to zero, our proposed method is still feasible.

The cloud computing center has powerful computing capacity and scalability, but the in-house server, especially those without virtualization, are limited by their service equipment and does not have the same powerful computing capacity and scalability as the cloud. Therefore, from the perspective of customers, they are more willing to join the cloud. Customers are strategic. Some customers who can't directly access the cloud will follow the arrangement of the web service company and join the in-house subsystem. But others will not. If they find that their jobs can't directly access the cloud, they will suspend their jobs for a while and then try again to enter the cloud at a specific retrial rate. Only when the queue-length in the in-house subsystem reaches the given queue-length limitation, can opportunists immediately access the cloud once they make retrials. Otherwise, retrying jobs remain in an orbit and repeat the previous operation. Customers in the orbit are called opportunists. The inter-retrial times are exponentially distributed with retrial rate θ.

The above model can be characterized as a two-tier queueing system with a retrial space. This is because there are two servers to choose from in the system: an in-house server and a cloud server, and customers who are unwilling to join the in-house server can join the retrial space and retry entering the cloud system with certain rate. Our theoretical and numerical research will reveal that, under some circumstances, using a model with an orbit space allows the web service company to increase the net benefit, suggesting that the managers have an incentive to set up such a service mechanism.

Due to the scalability and powerful computing power of the cloud computing center, all joined service requests can be immediately serviced when allowed by the controller, so there is no customer joining a retrial space in this situation. However, when the queue length in the in-house subsystem does not reach the queue-length limitation, the controller will block service requests from joining the cloud system, and these service requests will join the in-house subsystem or enter the retrial space. Strictly speaking, customers have three strategies when the cloud server is not open due to the admission control: joining the in-house subsystem directly, joining the retrial orbit and then retrying the in-house subsystem, and joining the retrial space and then retrying the cloud. However, joining the retrial orbit and then retrying the in-house subsystem must not be the best strategy, because the expected net benefit of an arriving customer in this situation must be less than joining the in-house subsystem directly. Therefore, when the in-house subsystem does not reach the queue-length limitation, we only consider that customers either join the in-house subsystem directly or enter the retrial orbit and then retry the cloud.

Zhu et al. [20] showed that an arriving customer will enter the in-house subsystem with equilibrium joining probability $q^e(L,\theta)$ or join the orbit with probability $1-q^e(L,\theta)$ when the queue-length in the in-house subsystem is lower than a given queue-length limitation. Served jobs can be divided into three types: some jobs directly enter the cloud and are served immediately (type-1 jobs); some jobs enter the cloud from the orbit (type-2 jobs); the remainder is served by the in-house server (type-3 jobs). Customers with type-i jobs are called type-i customers, where $i=1,2,3$.

The following symbols will be used in the rest of paper. We assume that ℓ is the queue-length in the in-house subsystem, which is a random variable. Each served customer will receive the reward of R after service completion. Service delay will incur a waiting cost, and customers in the retrial space will also have to pay operational fees. Let C_i, $i=0,1,2$ be the

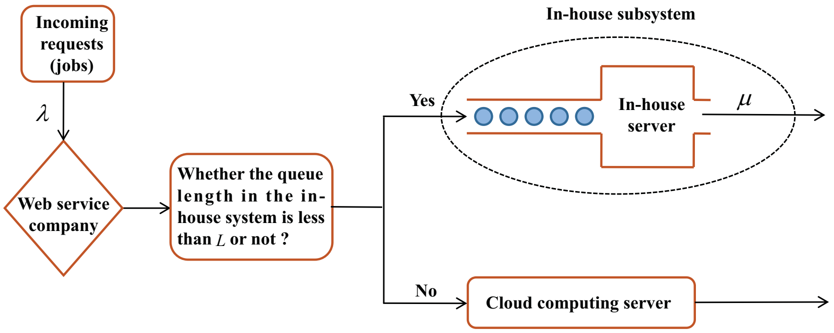


Fig. 2. Illustration of the hybrid service system without orbit space.

Table 1 *Notation and explanation*

Notation	Explanation
λ	The arrival rate of jobs.
μ	The service rate of jobs in the in-house subsystem.
θ	The retrial rate of opportunists retrying the cloud server.
γ	The VPC, i.e., the upper limit of total arrival rate of jobs sent to the cloud server.
L	the queue-length limitation.
ℓ	The queue-length in the in-house subsystem.
R	The reward after service completion
C_0	The waiting cost per unit time per job in the in-house subsystem.
C_1	The waiting cost per unit time per job in the orbit.
C_2	The operational fee of each retrial per unit time.
C_{h1}	the holding cost per time unit per job in the in-house subsystem.
C_{h2}	the holding cost per time unit per job in the orbit, respectively.
$q^e(L,\theta)$	The conditional equilibrium joining probability of joining the in-house subsystem given $\ell < L$.
$P_i(L,\theta,q^e(L,\theta))$	The probability of having i jobs in the in-house subsystem.
$N_{slow}(L,\theta,q^e(L,\theta))$	The mean number of jobs in the in-house subsystem.
$N_{orbit}(L,\theta,q^e(L,\theta))$	The mean number of jobs in the orbit.
$\bar{N}(L,\theta,q^e(L,\theta))$	The mean number of retrials before accessing the cloud successfully.
$\lambda_{otc}(\theta,q^e(L,\theta)\|\ell=L)$	The arrival rate of jobs from the orbit to the cloud under $\ell = L$.
$\Psi_1(L,\theta)$	The individual expected net benefit per unit time obtained by each type-2 customer after service completion.
$\Psi_2(L,\theta)$	The expected total net benefit per unit time of all type-2 customers.
$\Phi_{wsc}(L,\theta,q^e(L,\theta))$	The expected total net benefit per unit time of the web service company.
$\theta_i^*(L), i=1,2$	The non-cooperatively optimal retrial rate and the cooperatively optimal retrial rate, respectively.

waiting cost per unit time per job in the in-house subsystem, the waiting cost per unit time per job in the orbit and the operational fee of each retrial per unit time, respectively. C_{h1} (or C_{h2}) is the holding cost per time unit per job in the in-house subsystem (or in the orbit). The mean number of jobs in the in-house subsystem (or in the orbit) is denoted as $N_{slow}(L,\theta,q^e(L,\theta))$ (or $N_{orbit}(L,\theta,q^e(L,\theta))$). We assume that $\bar{N}(L,\theta,q^e(L,\theta))$ is the mean number of retrials before accessing the cloud successfully. Let $\lambda_{otc}(\theta,q^e(L,\theta)|\ell = L)$ be the arrival rate of jobs from the orbit to the cloud under $\ell = L$. $\Psi_i(L,\theta), i = 1,2$ denote the individual expected net benefit per unit time obtained by each type-2 customer after service completion and the expected total net benefit per unit time of all type-2 customers, respectively. The expected total net benefit per unit time of the web service company is denoted as $\Phi_{wsc}(L,\theta,q^e(L,\theta))$. Let $\theta_i^*(L), i = 1,2$ be the non-cooperatively optimal retrial rate and the cooperatively optimal retrial rate, respectively. The relevant definitions are also given in Table 1.

The input parameters $\lambda,\mu,\gamma,R,C_0,C_1,C_2$ remain fixed, and the rest can be obtained from the results in Zhu et al. [20]. In our proposed model, real-time queue lengths in the in-house subsystem need to be continuously tracked, and service requests in the in-house subsystem will be uniformly allocated by the system. The system needs to record their arrival order, track their location in real-time, and then provide services one by one in order. However, in orbit space, there is no need to record their arrival order or track their position in real-time, as they are not sorted in orbit space. Therefore, the waiting/holding costs in the orbit space differ from those in the in-house server. According to the result of [20], after replacing q in Theorem 3.1 of [20] (which denotes the joining probability) with equilibrium joining probability $q^e(L,\theta)$, we can obtain the following results.

Let $P_L(L,\theta,q^e(L,\theta))$ be the probability of having L jobs in the in-house subsystem in the equilibrium state. For the in-house subsystem, it can be considered as an $M/M/1/L$ queue, then $P_L(L,\theta,q^e(L,\theta))$ can be obtained based on the basic result of $M/M/1/L$ queue as follows:

$$P_L(L,\theta,q^e(L,\theta)) = \frac{(\rho q^e(L,\theta))^L(1-\rho q^e(L,\theta))}{1-(\rho q^e(L,\theta))^{L+1}}. \qquad (2.1)$$

Zhu et al. [20] obtained the conditional mean waiting time and the mean queue-length based on the probability generating method. By using the Little formula, the effective arrival rate from the orbit to the cloud given $\ell = L$, $\lambda_{otc}(\theta,q^e(L,\theta)|\ell = L)$, can be obtained as follows:

$$\lambda_{otc}(\theta,q^e(L,\theta)|\ell = L)$$

$$= \frac{\theta(1-(\rho q^e(L,\theta))^{L+1})H_L(\theta,q^e(L,\theta))}{(\rho q^e(L,\theta))^L(1-\rho q^e(L,\theta))D(\theta,q^e(L,\theta))}, \quad (2.2)$$

where $\rho = \lambda/\mu$, $D(\theta,q^e(L,\theta)) = |\mathbf{A}|, \mathbf{A} = (A_{i,j})_{(L+1)\times(L+1)} \in \mathbb{R}^{(L+1)\times(L+1)}$, in which $A_{1,1} = A_{k+1,k} = \lambda q^e(L,\theta), k = 1,2,\cdots,L$, $A_{L+1,L+1} = -(\theta+\mu)$, $A_{k,k} = -(\mu+\lambda q^e(L,\theta)), 2 \le k \le L$, $A_{1,2} = -A_{k,k+1} = -\mu, k = 1,2,\cdots,L$, and $H_i(\theta,q^e(L,\theta)) = |\mathbf{A}_i|$, $\mathbf{A}_i$ is $\mathbf{A}$ with its i-th column replaced by vector

$$\left[b_0,-b_1,\cdots,-b_{L-1},0\right]^T,$$

here

$$b_i = \frac{\lambda(1-q^e(L,\theta))(\rho q^e(L,\theta))^i(1-\rho q^e(L,\theta))}{1-(\rho q^e(L,\theta))^{L+1}}, 0 \le i \le L-1.$$

By using the expectation formula, the mean number of jobs in the in-house subsystem can be written as

$$N_{slow}(L,\theta,q^e(L,\theta)) = \frac{\rho q^e(L,\theta)-(\rho q^e(L,\theta))^{L+1}-(1-(\rho q^e(L,\theta))) L(\rho q^e(L,\theta))^{L+1}}{(1-\rho q^e(L,\theta))\,(1-(\rho q^e(L,\theta))^{L+1})}, \quad (2.3)$$

the mean number of jobs in the orbit is

$$N_{orbit}(L,\theta,q^e(L,\theta)) = \frac{\sum_{i=0}^{L} H_i(\theta,q^e(L,\theta))}{D(\theta,q^e(L,\theta))}, \quad (2.4)$$

and the mean number of opportunist's retrials before accessing to the cloud successfully is

$$\bar{N}(L,\theta,q^e(L,\theta)) = \frac{\theta\left(1-(\rho q^e(L,\theta))^{L+1}\right)\sum_{i=0}^{L-1} H_i(\theta,q^e(L,\theta))}{\lambda(1-q^e(L,\theta))\,(1-(\rho q^e(L,\theta))^L)\,D(\theta,q^e(L,\theta))}. \quad (2.5)$$

It should be noted that we discuss the related optimal decision problems under the condition that customers adopt the Nash equilibrium strategy in this paper. We will explore the optimal queue-length limitation. Based on the dynamic game between the web service company and customers, we also obtain the joint optimums of the queue-length limitation and the retrial rate. These results are instructive to the manager of the web service company. Further, some economic insights are also analyzed.

III. Optimal queue-length limitation

In this section, we consider the optimal queue-length limitation in the hybrid service system with orbit space from the viewpoint of the web service company. The manager of the web service company wants to maximize the company's benefits. To realize this objective, the manager has the authority to set the optimal queue-length limitation, since he/she decides the queue-length limitation. Theorem 3.1 provides the optimal queue-length limitation, and Proposition 3.2 shows that the expected net benefit per unit time of the web service company under its optimal policy will increase with the VPC of γ. All proofs are given in the Appendix.

Let $\lambda_{cloud}(L,\theta)$ be the total effective arrival rate of jobs entering the cloud system given $\ell = L$. Jobs entering the cloud can be divided into two types: (1) some jobs directly enter the cloud system when there are L jobs in the in-house subsystem upon new job arrival (arriving in the cloud from an external source); (2) some jobs enter the cloud system from the orbit (arriving in the cloud from the orbit). This means that the total effective arrival rate of jobs entering the cloud given $\ell = L$ is the sum of the arrival rate of jobs coming from external sources and the arrival rate of jobs coming from the orbit. Based on the above analysis, we get

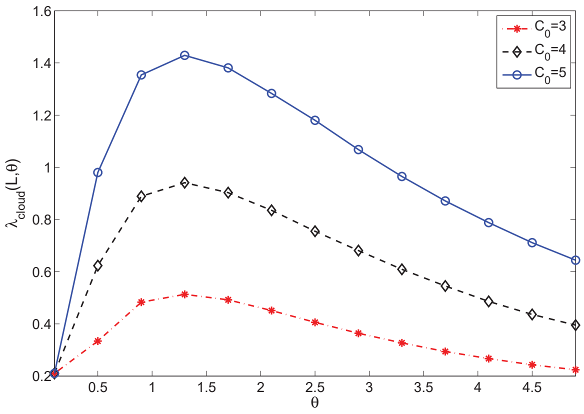


Fig. 3. The relation between $\lambda_{cloud}(L,\theta)$ and θ for $\lambda = 0.8, \mu = 1, L = 2, C_1 = 0.4, C_2 = 0.2$.

$$\begin{aligned}\lambda_{cloud}(L,\theta) =& \lambda_{otc}\left(\theta,q^e(L,\theta)|\ell = L\right) P_L(L,\theta,q^e(L,\theta)) \\ &+ \lambda P_L(L,\theta,q^e(L,\theta)),\end{aligned} \quad (3.1)$$

where $P_L(L,\theta,q^e(L,\theta))$ and $\lambda_{otc}\left(\theta,q^e(L,\theta)|\ell = L\right)$ can be determined by (2.1) and (2.2), respectively. Obviously, the total effective arrival rate of jobs entering the cloud is dependent on the retrial rate θ. Through numerical analysis (Figure 3), we easily find that the total effective arrival rate of jobs entering the cloud is weakly unimodal with respect to θ when the queue-length in the in-house subsystem reaches the queue-length limitation.

We assume that the web service company will obtain an income of P upon the arrival of each job. Note that the income often does not come from the payments of the customers, but rather from a third party, such as a company that has a contract with the web service company for advertisement services and pays the advertisement fees based on the number of jobs. Let C_{h1} and C_{h2} be the holding cost per time unit per job in the in-house subsystem and the holding cost per time unit per job in the orbit, respectively. From (2.3), the number of jobs in in-house subsystem in the equilibrium state, $N_{slow}(L,q^e(L,\theta))$, can be computed from

$$N_{slow}(L,q^e(L,\theta)) = \frac{\rho q^e(L,\theta)-(\rho q^e(L,\theta))^{L+1}-(1-\rho q^e(L,\theta))L(\rho q^e(L,\theta))^{L+1}}{(1-\rho q^e(L,\theta))\,(1-(\rho q^e(L,\theta))^{L+1})}.$$

From (2.4), we get the mean number of jobs in the orbit in the equilibrium state as follows:

$$N_{orbit}(L,\theta,q^e(L,\theta)) = \frac{\sum_{i=0}^{L} H_i(\theta,q^e(L,\theta))}{D(\theta,q^e(L,\theta))}.$$

The mean total income of the web service company per unit time comes from two parts: the income due to job arrivals and the income due to the operation fees of job retrials.

Obviously, the first part equals λP, and the second part is $C_2 N_{orbit}(L, \theta, q^e(L, \theta))\bar{N}(L, \theta, q^e(L, \theta))$. In addition, the mean total cost is the sum of the holding cost per unit time in the in-house subsystem and the holding cost per unit time in the orbit. Hence, the expected total net benefit per unit time of the web service company can be written as

$$\begin{aligned}&\Phi_{wsc}(L, \theta, q^e(L, \theta))\\ &= \lambda P + C_2 N_{orbit}(L, \theta, q^e(L, \theta))\bar{N}(L, \theta, q^e(L, \theta))\\ &- C_{h1} N_{slow}(L, q^e(L, \theta)) - C_{h2} N_{orbit}(L, \theta, q^e(L, \theta)),\end{aligned} \tag{3.2}$$

where $\bar{N}(L, \theta, q^e(L, \theta))$, $N_{slow}(L, q^e(L, \theta))$, $N_{orbit}(L, \theta, q^e(L, \theta))$ are determined by (2.3), (2.3) and (2.5), respectively. Let $L^*(\theta, \gamma)$ be the optimal queue-length limitation given the retrial rate θ and the VPC γ. We know that the queue-length limitation is decided by the manager of the web service company. Thus, the optimal queue-length limitation from the company's perspective can be obtained by maximizing the company's benefit.

Theorem 3.1 (Optimal queue-length limitation): For the given retrial rate θ and VPC γ, the optimal queue-length limitation can be computed from the following equation:

$$L^*(\theta, \gamma) = \arg\max_{L \in N} \left\{ \Phi_{wsc}(L, \theta, q^e(L, \theta)) \middle| \lambda_{cloud}(L, \theta) \le \gamma \right\}, \tag{3.3}$$

where $\lambda_{cloud}(L, \theta)$, $\Phi_{wsc}(L, \theta, q^e(L, \theta))$ are given in (3.1) and (3.2), respectively.

The web service company signs a long-term contract with an external cloud-computing provider for a fixed amount of computing resources to be consumed during the contract period. The contract stipulates that the total arrival rate of jobs sent to the cloud cannot exceed γ. Hence, the manager of the web service company must first ensure that the condition $\lambda_{cloud}(L, \theta) \le \gamma$ is satisfied and then determine the optimal queue-length limitation under this condition. Thus, (3.3) is an optimization problem with constraint, where $\Phi_{wsc}(L, \theta, q^e(L, \theta))$ can be computed from (3.2). Due to the uncomplicated objective function and constraint, we can easily obtain $L^*(\theta, \gamma)$ using Mathematica software or Matlab software. In addition, we find that γ has a key effect on the optimal queue-length limitation. The following proposition shows the relation between the expected total net benefit of the web service company and the VPC γ.

Proposition 3.2 (Monotonicity): Assume that γ_1 and γ_2 are two different VPCs, and $L^*(\theta, \gamma_1), L^*(\theta, \gamma_2)$ are the corresponding optimal queue-length limitations. If $\gamma_1 \le \gamma_2$, the following inequality holds:

$$\begin{aligned}&\Phi_{wsc}(L^*(\theta, \gamma_1), \theta, q^e(L^*(\theta, \gamma_1), \theta))\\ &\le \Phi_{wsc}(L^*(\theta, \gamma_2), \theta, q^e(L^*(\theta, \gamma_2), \theta)).\end{aligned} \tag{3.4}$$

Proof: See **A.1** in **Appendix** for the proof of Proposition 3.2.

Proposition 3.2 shows that the larger the VPC, the higher the benefit the web service company will obtain. However, as an external service resource, the cloud brings a variety of security risks, such as the leakage of the web service company's business data and customers' data. Because of this, the web service company needs to balance between utilities and risk tolerance. This is an interesting problem that is worthy of further study in future work.

IV. System Optimization Based on the Stackelberg Game

In this section, we consider the joint optimization problem in the hybrid service system with orbit space. We obtain the joint optimums of the queue-length limitation and the retrial rate based on the Stackelberg game. Specifically, Theorem 4.1 gives the joint optimums of the queue-length limitation and the retrial rate in the case that the web service company is the Stackelberg leader, and Theorem 4.2 gives the joint optimums in the case that opportunists are the Stackelberg leader. All proofs are given in the Appendix.

As we know, both customers and the web service company are selfish. They maximize their respective interests by choosing the appropriate admission control and the appropriate retrial rate. This dynamic game between them exists in real life. To maximize their interests under the other party's tactics, both parties decide on their own strategies depending on the potential strategies of the other party. In this game model, the party that decides first is referred to as the leader, while the party that decides second is referred to as the follower. The leader then modifies their choice in response to the follower's choice, and so on, until they attain Nash equilibrium. We assume that the parties' payoff functions are common knowledge and that the players know the complete history of the game thus far. The web service company and the opportunists represent different market forces in the market, so the Stackelberg game can be adopted.

Whether opportunists are cooperative or not must be determined based on specific actual situations. For example, if a web service company's service group is a specific group with a high correlation, we may consider them cooperative consumers. However, in general, customers are often noncooperative. In order to adapt our method to different market situations, we will explore the optimal strategies for each scenario. There are four cases that must be considered: (1) the manager is the Stackelberg leader and the opportunists are non-cooperative; (2) the manager is the Stackelberg leader and the opportunists are cooperative; (3) the opportunists are the Stackelberg leader and they are non-cooperative; (4) the opportunists are the Stackelberg leader and they are cooperative. $(\hat{L}_{m,1}, \hat{\theta}_{m,1})$, $(\hat{L}_{m,2}, \hat{\theta}_{m,2})$, $(\hat{L}_{c,1}, \hat{\theta}_{c,1})$ and $(\hat{L}_{c,2}, \hat{\theta}_{c,2})$ are the joint optimums of the queue-length limitation and the retrial rate in these four cases, respectively.

Theorem 4.1 (Joint optimums (a)): If the manager of the web service company is the Stackelberg leader, the joint optimums of the queue-length limitation and the retrial rate can be obtained from

$$\begin{cases} \hat{L}_{m,i} = \arg\max_{L \in \mathbf{N}} \left\{ \Phi_{wsc}\left(L, \theta_i^*(L), q^e(L, \theta_i^*(L))\right) \right. \\ \qquad\qquad\qquad \left. \middle| \lambda_{cloud}(L, \theta_i^*(L)) \le \gamma \right\}, i = 1, 2, \\ \hat{\theta}_{m,i} = \theta_i^*(\hat{L}_{m,i}), i = 1, 2, \end{cases} \tag{4.1}$$

where $\theta_i^*(L)$ can be obtained from Theorem 5.1 in [20], and $\lambda_{cloud}(L, \theta)$, $\Phi_{wsc}\left(L, \theta, q^e(L, \theta)\right)$ are given in (3.1) and (3.2), respectively.

Proof: See **A.2** in **Appendix** for the proof of Theorem 4.1.

In real-life situations, the web service company has a greater market force than the opportunists and often acts first. The opportunists observe the web service company's action, and they then make decisions based on their observations. Hence the web service company is generally the Stackelberg leader, and the opportunists are the Stackelberg follower. To extend our model to other scenarios, we also consider the case that the opportunists are the Stackelberg leader.

Theorem 4.2 (Joint optimums (b)): If the opportunists are the Stackelberg leader, the joint optimums of the queue-length limitation and the retrial rate can be obtained from

$$\begin{cases}\hat{\theta}_{c,i} = \arg \max\limits_{0<\theta<\infty} \Psi_i\big(L^*(\theta,\gamma),\theta\big), i = 1,2,\\ \hat{L}_{c,i} = L^*(\hat{\theta}_{c,i},\gamma), i = 1,2,\end{cases} \tag{4.2}$$

where $L^*(\theta,\gamma)$ is given in Theorem 3.1,

$$\Psi_1(L^*(\theta,\gamma),\theta) = R- \\ \frac{\begin{Bmatrix}(C_1+\theta C_2)\left(1-(\rho q^e(L^*(\theta,\gamma),\theta))^{L^*(\theta,\gamma)+1}\right)\\ \times\sum_{i=0}^{L^*(\theta,\gamma)-1} H_i(\theta,q^e(L^*(\theta,\gamma),\theta))\end{Bmatrix}}{\begin{Bmatrix}\lambda(1-q^e(L^*(\theta,\gamma),\theta))\left(1-(\rho q^e(L^*(\theta,\gamma),\theta))^{L^*(\theta,\gamma)}\right)\\ \times D(\theta,q^e(L^*(\theta,\gamma),\theta))\end{Bmatrix}}, \tag{4.3}$$

and

$$\Psi_2(L^*(\theta,\gamma),\theta) = \\ \frac{\begin{Bmatrix}\lambda\left(1-q^e(L^*(\theta,\gamma),\theta)\right)\left(1-(\rho q^e(L^*(\theta,\gamma),\theta))^{L^*(\theta,\gamma)}\right)\\ \Psi_1(L^*(\theta,\gamma),\theta)\end{Bmatrix}}{1-(\rho q^e(L^*(\theta,\gamma),\theta))^{L^*(\theta,\gamma)+1}}. \tag{4.4}$$

Proof: See **A.3** in **Appendix** for the proof of Theorem 4.2.

Theorem 4.1 and Theorem 4.2 imply the arithmetic logic used to compute the joint optimums of the queue-length limitation and the retrial rate. In practice, we can obtain the joint optimums in different cases using Matlab tools. How may we determine who the Stackelberg leader is? The answer can be obtained according to specific real-life situations. If the service resources provided by the web service company are scarce, the company can be regarded as the Stackelberg leader. However, if the required service resources saturate the market, we may regard opportunists as the Stackelberg leader.

V. Significance of the Existence of the Orbit

In this section, we discuss the significance of the existence of the orbit from the perspective of the web service company. According to the result of [20], the existence of opportunists (i.e., the customers that have sent jobs to the orbit) harms the greater social interest (i.e., social welfare). The term "social interest" denotes the overall benefits of all customers. In fact, the sum of consumer and enterprise benefits is often referred to as social welfare in the literature. Due to the fact that service fees are usually paid by customers to businesses and offset each other without affecting social welfare, all other literature views the overall customer benefit as social welfare. Each selfish cooperative/noncooperative opportunist sets a cooperatively/noncooperatively optimal retrial rate to maximize his/her benefit. In the absence of external constraints, opportunists always choose the optimal retrial rate to maximize their own benefits. In order to maximize the social welfare, the social planner needs to formulate relevant policies. In reality, the minority's interests are always at the expense of the majority's. As a result, we are not surprised that the existence of opportunists has a significant impact on social welfare, which is also due to their selfish behavioral strategies. In this paper, we will find another interesting phenomenon. Specifically, Theorem 5.1 and Theorem 5.2 show that from the perspective of the web service company, dealing strategically with the speculative behavior of certain customers can create more benefits for the web service company. All proofs are provided in the Appendix.

To explain this result more explicitly, we construct another model (see Figure 2 in Section II), which is the same as the model studied earlier in this paper, except that it has no orbit space. These two models are called "the model without orbit space" and "the model with orbit space", respectively. The existence of the orbit space implies that some opportunists are permitted to wait for some time and then retry to enter the cloud. But for the model without an orbit space, an arriving job can be sent only to the in-house subsystem when the queue-length in the in-house subsystem is less than the queue-length limitation. This means that no customer has a chance to become an opportunist. Indeed, if an orbit space is provided, this means that the manager of the web service company permits the existence of the opportunists.

First, we consider the model without an orbit space. When a customer sends a job to the web service company, the manager of the web service company first checks whether the queue-length in the in-house subsystem is less than the given queue-length limitation. If not, then the newly arriving job will be sent to the cloud; otherwise, it will be sent to the in-house subsystem. As we know, in the model without an orbit space, no customer has the option to postpone his/her job and retry to enter the cloud at a later time; rather, the newly arriving customer must send his/her job to the in-house subsystem when the queue-length in the in-house subsystem is less than the queue-length limitation. For the given queue-length limitation L, $N_{nos}(L)$ and $P_{nos}(i)$ denote the mean number of jobs in the in-house subsystem and the probability of having i jobs in the in-house subsystem, respectively. The in-house subsystem can be regard as the $M/M/1/L$ queue with arrival rate λ and service rate μ. From the basic result of the queueing system, we get $P_{nos}(L) = \rho^i(1-\rho)/(1-\rho^{L+1})$. In addition, we find that the model with an orbit space will degenerate to the model without an orbit space as $q = 1$. So $N_{nos}(L) = N_{slow}(L,1)$. From (2.3), we have

$$N_{nos}(L) = \frac{\rho-\rho^{L+1}-(1-\rho)L\rho^{L+1}}{(1-\rho)\left(1-\rho^{L+1}\right)}. \tag{5.1}$$

By the simple proof, we can find that $N_{nos}(L)$ is increasing in L. Let $\Phi_{nos}(L)$ be the expected net benefit of the web service company per unit time for the given queue-length limitation L. It is the mean total income per unit time minus the mean total cost per unit time. In the model without an orbit space, the mean total income per unit time is the total arrival rate λ

times P, and the mean total cost per unit time equals the mean number of jobs in the in-house subsystem $N_{nos}(L)$ times C_{h1}. Therefore, the expected net benefit per unit time of the web service company for the given queue-length limitation L is

$$\Phi_{nos}(L) = \lambda P - C_{h1} N_{nos}(L), \tag{5.2}$$

where $N_{nos}(L)$ is determined by (5.1). Obviously, $\Phi_{nos}(L)$ is decreasing in L since $N_{nos}(L)$ is an increasing function with respect to L. L^*_{nos} denotes the optimal queue-length limitation for the web service company in the model without an orbit space, and it can be obtained from

$$L^*_{nos} = \arg \max_{L \in N} \left\{ \Phi_{nos}(L) \middle| \lambda P_{nos}(L) \leq \gamma \right\}. \tag{5.3}$$

Since $\Phi_{nos}(L)$ is decreasing function with respect to L, L^*_{nos} can be rewritten as

$$L^*_{nos} = \min_{L \in N} \left\{ L : \lambda P_{nos}(L) \leq \gamma \right\}, \tag{5.4}$$

where $P_{nos}(L) = \rho^L(1-\rho)/(1-\rho^{L+1})$. From (5.4), we get L^*_{nos} as follows:

$$L^*_{nos} = \left\lceil \log_\rho \frac{\gamma}{\lambda(1-\rho) + \gamma\rho} \right\rceil. \tag{5.5}$$

As the queue-length limitation is L^*_{nos}, the net benefits of the web service company are maximized. Thus, the web service company will follow the admission control policy: If the queue-length in the in-house subsystem reaches L^*_{nos}, an arriving job will be sent to the cloud; otherwise, it will be sent to the in-house subsystem. Therefore, the mean number of jobs in the in-house subsystem can be written as $N_{nos}(L^*_{nos})$. From (5.1), we have

$$N_{nos}(L^*_{nos}) = \frac{\rho - \rho^{L^*_{nos}+1} - (1-\rho) L^*_{nos} \rho^{L^*_{nos}+1}}{(1-\rho)\left(1 - \rho^{L^*_{nos}+1}\right)}. \tag{5.6}$$

In the model without an orbit space, the expected net benefit per unit time of the web service company corresponding to the queue-length limitation L^*_{nos} is

$$\lambda P - C_{h1} N_{nos}(L^*_{nos}). \tag{5.7}$$

Secondly, we consider the model with an orbit space. As stated in Section IV, there are four cases that must be considered. The joint optimums of the queue-length limitations and the retrial rates in the four cases — $(\hat{L}_{m,1}, \hat{\theta}_{m,1})$, $(\hat{L}_{m,2}, \hat{\theta}_{m,2})$, $(\hat{L}_{c,1}, \hat{\theta}_{c,1})$, and $(\hat{L}_{c,2}, \hat{\theta}_{c,2})$ — are given in Theorem 4.1 and Theorem 4.2. We first consider the case that the manager of the web service company is the Stackelberg leader. In the model with an orbit space, the mean total income per unit time of the web service company comes from the income generated by job arrivals and the operational fees that customers in the orbit pay the web service company. The mean total incomes per unit time in cooperative and non-cooperative cases are $C_2 N_{orbit}(\hat{L}_{m,i}, \hat{\theta}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})) \bar{N}(\hat{L}_{m,i}, \hat{\theta}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i}))$ $+\lambda P$, $i = 1, 2$, respectively. The mean total holding cost per unit time is composed of two parts: the holding cost in the in-house subsystem and the holding cost in the orbit. So the mean total holding costs per unit time in cooperative and non-cooperative cases can be written as

$$\begin{aligned} &C_{h1} N_{slow}(\hat{L}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})) \\ &\quad + C_{h2} N_{orbit}(\hat{L}_{m,i}, \hat{\theta}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})), i = 1, 2. \end{aligned}$$

Therefore, when the manager of the web service company is the Stackelberg leader, the expected net benefits per unit time of the web service company in the non-cooperative and cooperative cases can be obtained from

$$\begin{aligned} &C_2 N_{orbit}(\hat{L}_{m,i}, \hat{\theta}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})) \bar{N}(\hat{L}_{m,i}, \hat{\theta}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})) \\ &+ \lambda P - C_{h1} N_{slow}(\hat{L}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})) \\ &- C_{h2} N_{orbit}(\hat{L}_{m,i}, \hat{\theta}_{m,i}, q^e(\hat{L}_{m,i}, \hat{\theta}_{m,i})) \; i = 1, 2. \end{aligned} \tag{5.8}$$

If the opportunists are the Stackelberg leader, by adopting a similar method, the expected net benefits per unit time of the web service company in non-cooperative and cooperative cases can be computed from

$$\begin{aligned} &C_2 N_{orbit}(\hat{L}_{c,i}, \hat{\theta}_{c,i}, q^e(\hat{L}_{c,i}, \hat{\theta}_{c,i})) \bar{N}(\hat{L}_{c,i}, \hat{\theta}_{c,i}, q^e(\hat{L}_{c,i}, \hat{\theta}_{c,i})) \\ &+ \lambda P - C_{h1} N_{slow}(\hat{L}_{c,i}, q^e(\hat{L}_{c,i}, \hat{\theta}_{c,i})) \\ &- C_{h2} N_{orbit}(\hat{L}_{c,i}, \hat{\theta}_{c,i}, q^e(\hat{L}_{c,i}, \hat{\theta}_{c,i})) \; i = 1, 2. \end{aligned} \tag{5.9}$$

To sum up, the expected net benefits per unit time of the web service company is given in (5.7) for the model without an orbit space, and the expected net benefits per unit time in the model with an orbit space can be computed from (5.8) and (5.9). If the former is less than the latter, it is a wise decision for the web service company to permit the existence of opportunists.

Theorem 5.1 (Existence condition of an orbit space (a)): Assume that the manager of the web service company is the Stackelberg leader. The web service company can obtain more benefits in the model with an orbit space if one of the following two cases holds:

(1) the opportunists are non-cooperative and the following inequality holds:

$$\frac{\left\{ \begin{array}{c} C_{h1}[N_{slow}(\hat{L}_{m,1}, q^e(\hat{L}_{m,1}, \hat{\theta}_{m,1})) - N_{nos}(L^*_{nos})] \\ + C_{h2} N_{orbit}(\hat{L}_{m,1}, \hat{\theta}_{m,1}, q^e(\hat{L}_{m,1}, \hat{\theta}_{m,1})) \end{array} \right\}}{\left\{ \begin{array}{c} C_2 N_{orbit}(\hat{L}_{m,1}, \hat{\theta}_{m,1}, q^e(\hat{L}_{m,1}, \hat{\theta}_{m,1})) \\ \times \bar{N}(\hat{L}_{m,1}, \hat{\theta}_{m,1}, q^e(\hat{L}_{m,1}, \hat{\theta}_{m,1})) \end{array} \right\}} < 1, \tag{5.10}$$

(2) the opportunists are cooperative and the following inequality holds:

$$\frac{\left\{ \begin{array}{c} C_{h1}[N_{slow}(\hat{L}_{m,2}, q^e(\hat{L}_{m,2}, \hat{\theta}_{m,2})) - N_{nos}(L^*_{nos})] \\ + C_{h2} N_{orbit}(\hat{L}_{m,2}, \hat{\theta}_{m,2}, q^e(\hat{L}_{m,2}, \hat{\theta}_{m,2})) \end{array} \right\}}{\left\{ \begin{array}{c} C_2 N_{orbit}(\hat{L}_{m,2}, \hat{\theta}_{m,2}, q^e(\hat{L}_{m,2}, \hat{\theta}_{m,2})) \\ \times \bar{N}(\hat{L}_{m,2}, \hat{\theta}_{m,2}, q^e(\hat{L}_{m,2}, \hat{\theta}_{m,2})) \end{array} \right\}} < 1, \tag{5.11}$$

where $N_{slow}(L, q)$, $N_{orbit}(L, \theta, q)$, $\bar{N}(L, \theta, q)$, and $N_{nos}(L)$ are given in (2.3), (2.4), (2.5) and (5.1), respectively.

Proof: See **A.4** in **Appendix** for the proof of Theorem 5.1.

If the opportunists are the Stackelberg leader and the web service company is the Stackelberg follower, the results similar to those in Theorem 5.1 can be obtained. We summarize these results in the following theorem.

Theorem 5.2 (Existence condition of an orbit space (b)): Assume that the opportunists are the Stackelberg leader. The web service company can obtain more benefits in the model with an orbit space if one of the following two cases holds:

(1) the opportunists are non-cooperative and the following inequality holds:

$$\frac{\left\{\begin{matrix} C_{h1}[N_{slow}(\hat{L}_{c,1},q^{e}(\hat{L}_{c,1},\hat{\theta}_{c,1}))-N_{nos}(L^{*}_{nos})] \\ +C_{h2}N_{orbit}(\hat{L}_{c,1},\hat{\theta}_{c,1},q^{e}(\hat{L}_{c,1},\hat{\theta}_{c,1})) \end{matrix}\right\}}{\left\{\begin{matrix} C_{2}N_{orbit}(\hat{L}_{c,1},\hat{\theta}_{c,1},q^{e}(\hat{L}_{c,1},\hat{\theta}_{c,1})) \\ \times\bar{N}(\hat{L}_{c,1},\hat{\theta}_{c,1},q^{e}(\hat{L}_{c,1},\hat{\theta}_{c,1})) \end{matrix}\right\}} < 1, \quad (5.12)$$

(2) the opportunists are cooperative and the following inequality holds:

$$\frac{\left\{\begin{matrix} C_{h1}[N_{slow}(\hat{L}_{c,2},q^{e}(\hat{L}_{c,2},\hat{\theta}_{c,2}))-N_{nos}(L^{*}_{nos})] \\ +C_{h2}N_{orbit}(\hat{L}_{c,2},\hat{\theta}_{c,2},q^{e}(\hat{L}_{c,2},\hat{\theta}_{c,2})) \end{matrix}\right\}}{\left\{\begin{matrix} C_{2}N_{orbit}(\hat{L}_{c,2},\hat{\theta}_{c,2},q^{e}(\hat{L}_{c,2},\hat{\theta}_{c,2})) \\ \times\bar{N}(\hat{L}_{c,2},\hat{\theta}_{c,2},q^{e}(\hat{L}_{c,2},\hat{\theta}_{c,2})) \end{matrix}\right\}} < 1, \quad (5.13)$$

where $N_{slow}(L,q)$, $N_{orbit}(L,\theta,q)$, $\bar{N}(L,\theta,q)$, and $N_{nos}(L)$ are given in (2.3), (2.4), (2.5) and (5.1), respectively.

Proof: See **A.5** in **Appendix** for the proof of Theorem 5.2.

Remark 5.3: According to the result of [20], the existence of opportunists significantly harms social interests. However, Theorem 5.1 and Theorem 5.2 show that sometimes the web service company can obtain more benefits in the model with an orbit space. That is to say, the existence of opportunists, regardless of whether they are cooperative or non-cooperative, may be beneficial to the web service company, even though opportunists harm the greater social interests. Specifically, if the condition of Theorem 5.1 or Theorem 5.2 holds, the web service company can obtain more benefits in the model with an orbit space, and thus the web service company should consider opening an orbit space.

Remark 5.4: If (5.10)-(5.13) do not hold, the web service company can obtain more benefits in the model without an orbit space; that is, the existence of an orbit is bad for the web service company. Hence, the web service company should not open an orbit space, and will have no opportunists in the system.

In this following, we show the algorithm of net benefits of web service company under two different models when the manager of the web service company is the Stackelberg leader (see Table 2).

VI. Numerical analysis

In this section we explore the effect of different parameters on the optimal strategy and the net benefit of the web service company through numerical analysis. The numerical experiments below all assume that the web service company is the Stackelberg leader and the opportunists may be cooperative or noncooperative. The related numerical results were obtained based on Matlab R2020b. In the following numerical experiments, except for the varying parameters, other default parameters that we use are shown in Table 3.

- **Comparisons of optimal queue-length limitations**. First, we consider the case of smaller C_0 (see Figure 4). In this case, if VPC is small (e.g., $\gamma < 0.102$ in Figure 4(a)), a larger optimal queue-length limitation will be applied under the model without orbit space, while the opposite is true for larger VPC. This is because the data security risk requirements of the web service company are higher when the VPC is smaller, and the orbit space can effectively buffer the arrival rate at the cloud server, so the model with orbit space can adopt a smaller optimal queue-length limitation. In addition, it can be observed from Figure 4(b) that the optimal queue-length limitation under the model with orbit space will be larger if C_0 is relatively large. This is because the profit that the web service company obtains from a single service request is relatively small as C_0 is relatively large. In order to maximize profits, the web service company will allow more service requests to enter the in-house subsystem while meeting risk control conditions. Therefore, setting a larger queue length limitation for the in-house subsystem in this situation is more profitable.

Table 2 *The algorithm of net benefits of web service company under two different models.*

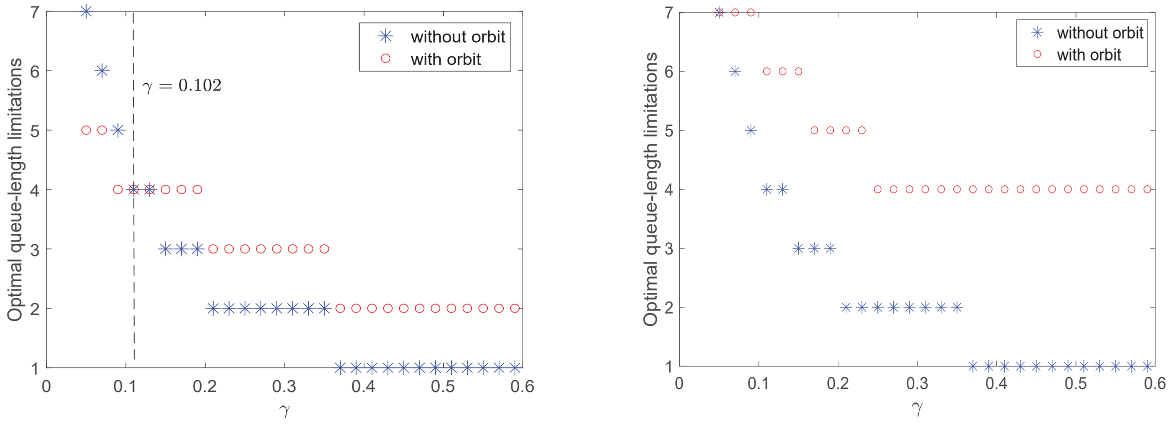
Step 1	Obtain the cooperatively and non-cooperatively optimal retrial rates $\theta_i^*(L)$ by Theorem 5.1 in [20]. Substituting $\theta_i^*(L)$ into (3.1) yields $\lambda_{cloud}(L,\theta_i^*(L))$.
Step 2	Compute the conditional equilibrium joining probability of a job joining the in-house subsystem $q^e(L,\theta_i^*(L))$ by substituting $\theta_i^*(L)$ into $q^e(L,\theta)$, where $q^e(L,\theta)$ can be obtained from Theorem 4.4 in [20].
Step 3	Obtain $\Phi_{wsc}(L,\theta_i^*(L),q^e(L,\theta_i^*(L)))$ by substituting $\theta_i^*(L)$ and $q^e(L,\theta_i^*(L)$ into (3.2).
Step 4	Compute $\hat{L}_{m,i}$ by substituting $\lambda_{cloud}(L,\theta_i^*(L))$ and $\Phi_{wsc}(L,\theta_i^*(L),q^e(L,\theta_i^*(L)))$ into (4.1). Get $\hat{\theta}_{m,i}$ by using $\hat{\theta}_{m,i}=\theta_i^*(\hat{L}_{m,i}), i=1,2$.
Step 5	Obtain the net benefits of web service company in the model with orbit space by substituting $(\hat{L}_{m,i},\hat{\theta}_{m,i})$ into (5.8).
Step 6	Compute L^*_{nos} from (5.5). Substituting L^*_{nos} into (5.7) yields the net benefits of web service company in the model without orbit space.

Table 3 *The default values of all the parameters.*

R	P	λ	μ	θ	C_0	C_1	C_2	C_{h1}	C_{h2}	γ
4	7	0.8	1	2	4	0.4	0.2	2.8	2.5	0.32

- **Effect of VPC on joint optimal strategies**. We give the joint optimums of the queue-length limitation and the retrial rate when the manager of the web service company is the Stackelberg leader. Figure 5 shows the relation between the joint optimums and VPC, which can be obtained based on the algorithm provided in Section IV. In Figure 5(a), the opportunists are non-cooperative; while the cooperative case is given in Figure 5(b). When γ is sufficiently large, the joint optimum will remain unchanged. This is because if a large VPC is set, the web service company is not sensitive to data security risks, and in this case the VPC poses little constraint on the system, so the joint optimum will not depend on it.

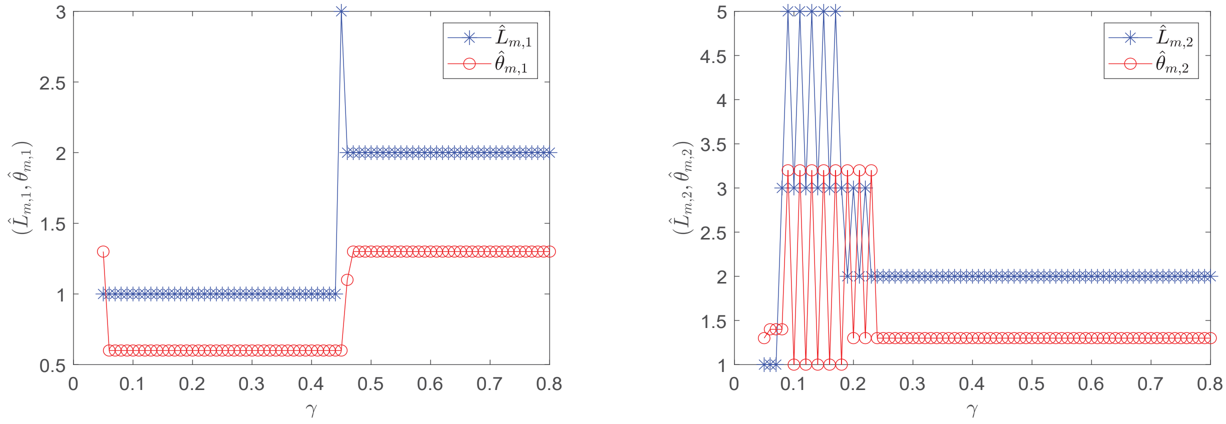
- **Comparison of net benefits under two models**. With the increase of C_{h2}, the net benefit of web service company under the model without orbit space remains constant, while



(a) $C_0 = 9$

(b) $C_0 = 21$

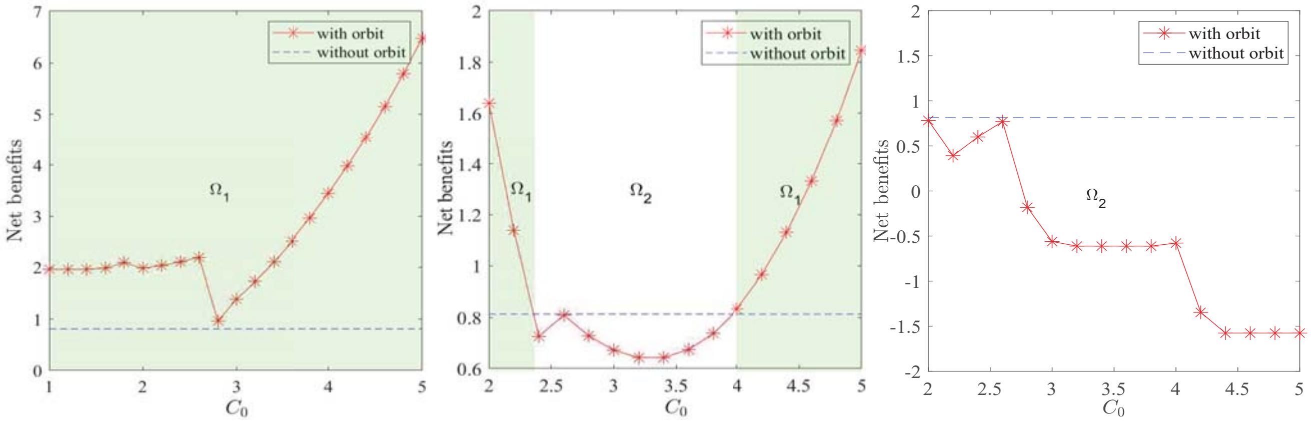
Fig. 4. Optimal queue-length limitations vs. γ for $R = 7, P = 4, \lambda = 0.8, \mu = 1, \theta = 2, C_1 = 0.4, C_2 = 0.2$.



(a) The non-cooperative case

(b) The cooperative case

Fig. 5. Net benefits of web service company vs. γ for $R = 7, P = 4, \mu = 1, C_0 = 5, C_1 = 0.4, C_2 = 0.2, C_{h1} = 2.8, C_{h2} = 2.5$.



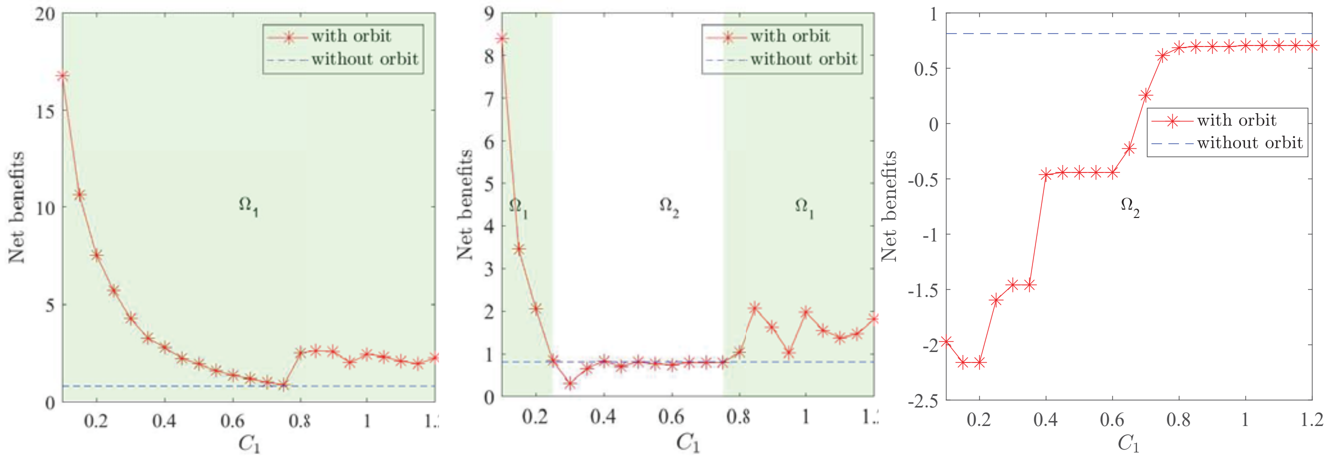
(a) $C_{h2} = 1$

(b) $C_{h2} = 2.5$

(c) $C_{h2} = 7.5$

Fig. 6. Net benefits of web service company vs. C_0 when the opportunists are non-cooperative for $R = 7, P = 4, \lambda = 0.8, \mu = 1, C_1 = 0.4, C_2 = 0.2, C_{h1} = 2.8, \gamma = 0.32$.

that under the model with orbit will gradually decrease. Figure 6 shows that when C_{h2} is small (e.g., $C_{h2} = 1$ in Figure 6(a)), the expected net benefit of the web service company under the model with orbit is always greater than that under the



(a) $C_{h2} = 1$

(b) $C_{h2} = 2.5$

(c) $C_{h2} = 6$

Fig. 7. Net benefits of web service company vs. C_0 when the opportunists are non-cooperative for $R = 7, P = 4, \lambda = 0.8, \mu = 1, \theta = 2, C_0 = 4, C_2 = 0.2, C_{h1} = 2.8, \gamma = 0.32$.

model without orbit space. However, when C_{h2} is sufficiently large (e.g., $C_{h2} = 7.5$ in Figure 6(c)), the result is exactly the opposite. In other cases, the profit curves of the web service company under these two models intersect, and there is no one model that completely dominates the other. The specific model to be adopted depends on the specific value of C_{h2}. Region Ω_1 (or Region Ω_2) in Figure 6 indicates that the service mechanism with retrial orbit (or without retrial orbit) should be adopted. Similarly, when C_1 or C_2 are small, a service mechanism with orbit space will benefit the web service company, while a service mechanism without orbit space should be used (see Figures 7-8) when C_1 or C_2 are large. In addition, under the model with orbit space, we can also observe that the net benefit of the web service company in the case of $C_{h1} > C_{h2}$ is higher than that in the case of $C_{h1} < C_{h2}$, when the opportunists may be cooperative (see Figure 10(a)) or noncooperative (see Figure 10(b)). Actually, in real-life situations, C_{h1} is generally greater than C_{h2}, which has been explained in Section II. Figure 10 shows a comparison of net benefits between two models using the three-dimensional graphs. The model with orbit space is more advantageous for the web service company when the green surface is above the yellow surface, while the model without orbit space is better when the yellow surface is above.

- **System improvement and the percentage of profit increase**. Our proposed hybrid service system with orbit only requires adding a control system based on real-time queue-length feedback to the original hybrid service system. However, the need to implement our proposed mechanism must be based on the specific input parameters; for example, when the holding cost in the in-house subsystem is high or the holding cost in the orbit space is relatively low, executing our proposed hybrid service system with orbit can significantly improve the revenue of service providers. According to Figure 11, when $C_0 = 2.6$, adopting a model with a retrial model, the company's revenue increases by approximately 200% in the cooperative case and by nearly 100% in the non-cooperative case. In particular, when $C_0 = 4$, the company's revenue increases by about 250%.

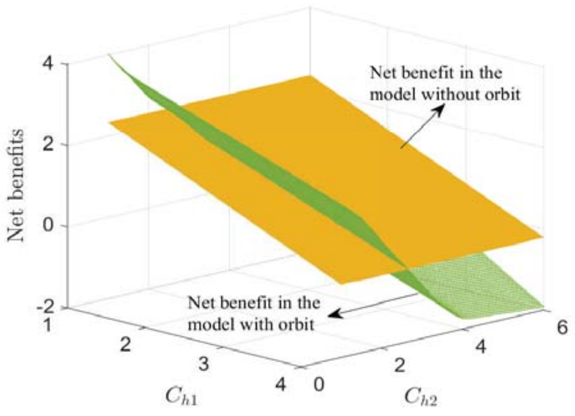
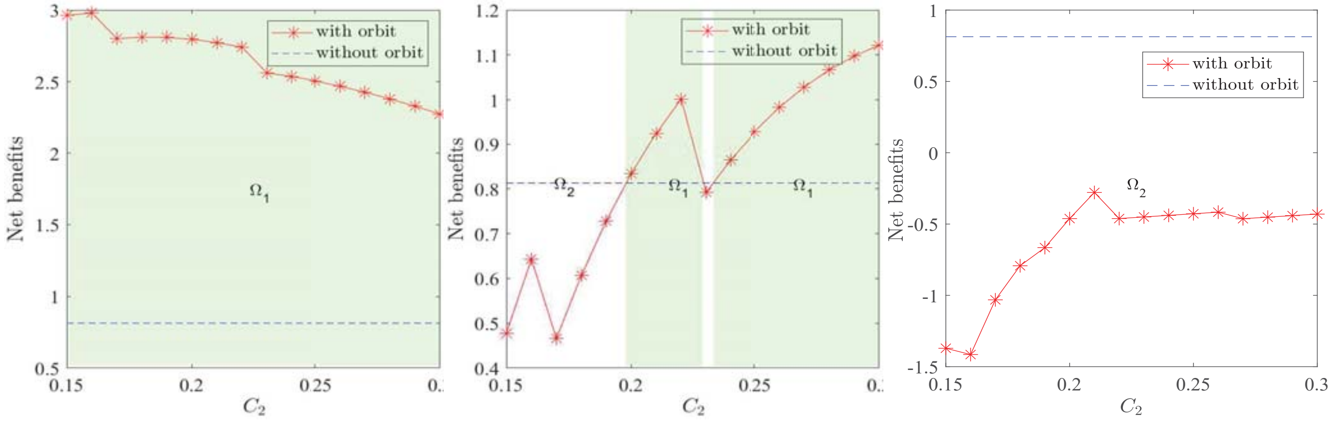


Fig. 10. Net benefits of web service company vs. C_{h1} and C_{h2} in non-cooperative case for $R = 7, P = 4, \lambda = 0.8, \mu = 1, \theta = 2, C_0 = 4, C_1 = 0.4, C_2 = 0.2, \gamma = 0.32$.

VII. Conclusions

In this paper, we considered the optimal decision problems involved in the allocation of jobs in a hybrid service system. We summarized the real problem as a queueing system with two servers and one orbit space, obtaining the optimal queue-length limitation from the viewpoint of the manager. Additionally, we derived the joint optimums of the queue-length limitation and the retrial rate based on the Stackelberg game. We observed an interesting phenomenon: the existence of opportunists in the system can harm greater social interests but may be beneficial to the web service company in certain situations. We also enhanced our previous work in characterizing performance measures by comparing of the hybrid service model with orbit space and the hybrid service system without orbit space to explore whether the orbit space should be set or not and to initially identify specific conditions in applications.

For our future work, we plan to investigate the infrastructure utilization and identify how our model, when applied, affects

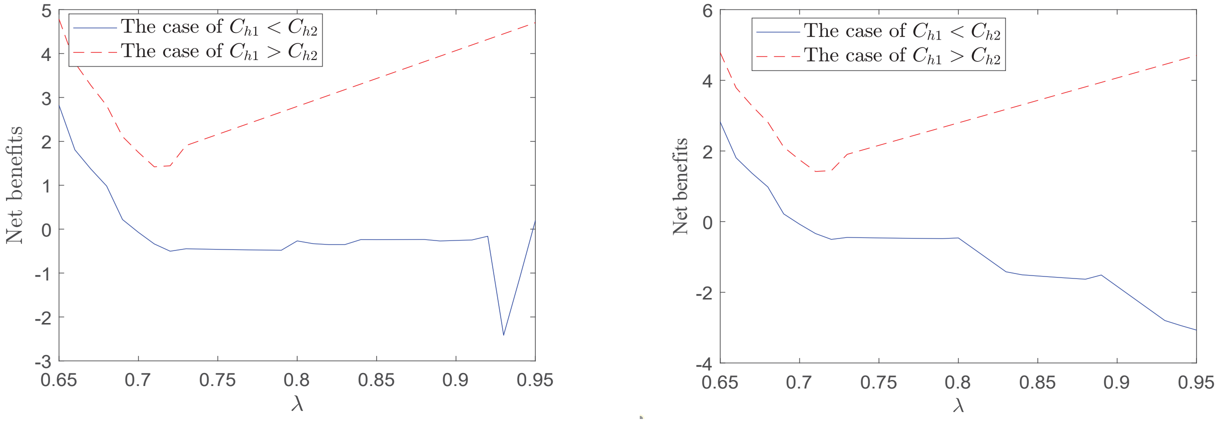


(a) $C_{h2} = 1$

(b) $C_{h2} = 2.5$

(c) $C_{h2} = 6$

Fig. 8. Net benefits of web service company vs. C_0 when the opportunists are non-cooperative for $R = 7, P = 4, \lambda = 0.8, \mu = 1, \theta = 2, C_0 = 4, C_1 = 0.4, C_{h1} = 2.8, \gamma = 0.32$.



(a) The non-cooperative case

(b) The cooperative case

Fig. 9. Net benefits of web service company vs. λ for $R = 7, P = 4, \mu = 1, \theta = 2, C_0 = 4, C_1 = 0.4, C_2 = 0.2, C_{h1} = 2.8, C_{h2} = 2.5$.

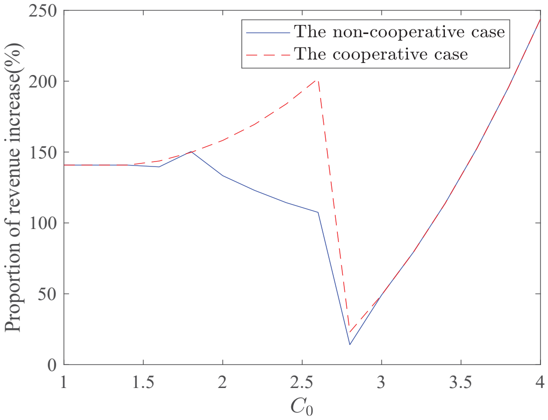


Fig. 11. Proportion of revenue increase vs. C_0 for $R = 7, P = 4, \lambda = 0.8, \mu = 1, C_1 = 0.4, C_2 = 0.2, C_{h1} = 2.8, C_{h2} = 1, \gamma = 0.32$.

specific resources such as links, capacities, latencies. These details are typically obtained only with (at least) a low-level simulator/emulator of the system. Additionally, we aim to extend the model to a more challenging trilateral game model. While in this paper, we only consider the Stackelberg game between customers and the web service company, in reality, the cloud provider interacts with both the web service company and the customers and thus a trilateral game would possibly a more realistic model and is the logical next step for our future research. Furthermore, this paper does not consider the case of heterogeneous customers, another extension of our model that is worth studying.

Acknowledgement

The authors would like to thank the Editor-in-Chief, Dr. Hanan Lutfiyya, the Associate Editor and four anonymous referees for their valuable comments and suggestions which greatly improved the presentation and the quality of this paper.

References

[1] Brender, N., Markov, I., Risk perception and risk management in cloud computing: Results from a case study of Swiss companies. Int. J. Inform. Manag., 33 (2013) 726–733.

[2] Burnetas, A., Economou, A., Equilibrium customer strategies in a single server Markovian queue with setup times. Queueing Systems, 56 (2007) 213–228.

[3] Caldentey, R., Wein, L. M., Analysis of a decentralized production-inventory system. Manufacturing & Service Operations Management, 5 (2003) 1–17.

[4] Economou, A., Manou, A., Strategic behavior in an observable fluid queue with an alternating service process. European J. Operational Research, 254 (2016) 148–160.

[5] Engel, R., Hassin, R. Customer equilibrium in a single-server system with virtual and system queues. Queueing Systems, 87 (2017) 213-228.

[6] Guo, P., Hassin, R., Strategic Behavior and Social Optimization in Markovian Vacation Queues. Operations Research, 41 (2013) 277–284.

[7] Guo, P., Lindsey, R., Zhang, Z.G., On the Downs-Thomson paradox in a self-financing two-tier queuing system. Manufacturing & Service Operations Management, 16 (2014): 315–322.

[8] Hassin, R., Haviv, M., Nash equilibrium and subgame perfection in observable queues. Annals of Operational Research, 113 (2002) 15-26.

[9] Hassin, R., Snitkovsky, R.I. Strategic customer behavior in a queueing system with a loss subsystem. Queueing Systems, 86 (2017) 361–387.

[10] Hua, Z., Chen, W., Zhang, Z. G., Competition and Coordination in Two-Tier Public Service Systems under Government Fiscal Policy. Production and Operations Management, 25 (2016) 1430-1448.

[11] Li, M., et al., A computing offloading game for mobile devices and edge cloud servers. Wireless Communications and Mobile Computing, 2018 (2018) 1-10.

[12] Manou, A., Economou, A., Karaesmen, F., Strategic customers in a transportation station: when is it optimal to wait? Operations Research, 62 (2014) 910-925.

[13] Naor, P., The regulation of queue size by levying tolls. Econometrica, 37 (1969) 15-24.

[14] Shi, Y., Lian Z., optimization and strategic behavior in a passenger-taxi service system. European J. Operational Research, 249 (2016) 1024–1032.

[15] Spencer, J., Sudan, M., Xu, K., Queuing with future information. Annals of Appllied Probability, 24 (2014) 2091–2142.

[16] Tuohy, C. H., Flood, C. M., Stabile, M., How does private finance affect public health care systems? Marshaling the evidence from OECD nations. J. Health Polit. Polic., 29 (2004) 359–396.

[17] Wang, J., Zhang, F., Strategic joining in $M/M/1$ retrial queues. European J. Operational Research, 230 (2013) 76-87.

[18] Xu, K., Necessity of future information in admission control. Operations Research, 63 (2015) 1213-1226.

[19] Xu K, Chan C W. Using future information to reduce waiting times in the emergency department via diversion. Manufacturing & Service Operations Management, 18 (2016) 314-331.

[20] Zhu S., Wang J., Li W. W. Cloud or in-house service? Strategic joining and social optimality in hybrid service systems with retrial orbit. IEEE Systems Journal, 17 (2023) 3810-3821.

Appendix

A.1. Proof of Proposition 3.2 If the VPC equals γ_1, according to Theorem 3.1, the optimal queue-length limitation adopted by the web service company can be written as:

$$L^*(\theta,\gamma_1)=\arg\max_{L\in N}\left\{\Phi_{wsc}(L,\theta,q^e(L,\theta))\middle|\lambda_{cloud}(L,\theta)\leq\gamma_1\right\}. \tag{A.1}$$

Since $\gamma_1 \leq \gamma_2$, we get the inequality as follows:

$$\lambda_{cloud}(L^*(\theta,\gamma_1),\theta)\leq\gamma_1\leq\gamma_2. \tag{A.2}$$

From (3.3), we also obtain

$$L^*(\theta,\gamma_2)=\arg\max_{L\in N}\left\{\Phi_{wsc}(L,\theta,q^e(L,\theta))\middle|\lambda_{cloud}(L,\theta)\leq\gamma_2\right\}.$$

From (A.2), we immediately obtain (3.4) in Proposition 3.2; that is

$$\begin{aligned}&\Phi_{wsc}(L^*(\theta,\gamma_1),\theta,q^e(L^*(\theta,\gamma_1),\theta))\\&\quad\leq\Phi_{wsc}(L^*(\theta,\gamma_2),\theta,q^e(L^*(\theta,\gamma_2),\theta)).\end{aligned} \tag{A.3}$$

This completes the proof.

A.2. Proof of Theorem 4.1 According to the condition of Theorem 4.1, the manager of the web service company is the Stackelberg leader. Assume opportunists are non-cooperative. From Theorem 5.1 in [20], we get the optimal retrial rate $\theta_1^*(L)$ for a given queue-length limitation L. Second, under the condition that $\theta=\theta_1^*(L)$, the optimal queue-length limitation $\hat{L}_{m,1}$ can be obtained from Theorem 3.1. Then, the optimal retrial rate $\hat{\theta}_{m,1}$ can be written as $\theta_1^*(\hat{L}_{m,1})$. Hence, the joint optimum value of the queue-length limitation and the retrial rate is $(\hat{L}_{m,1},\hat{\theta}_{m,1})$ when the manager of the web service company is the Stackelberg leader. If opportunists are cooperative, we can obtain the joint optimum using a similar method.

A.3. Proof of Theorem 4.2 According to the condition of Theorem 4.2, opportunists are the Stackelberg leader. We only prove the case that the opportunists are non-cooperative. A similar analysis can be used for the cooperative case. From Theorem 3.1, we get the optimal length limitation $L^*(\theta,\gamma)$ for the given retrial rate θ and the given VPC γ. Under the condition that $L=L^*(\theta,\gamma)$, the optimal retrial rate $\hat{\theta}_{c,1}$ can be obtained from Theorem 5.1 in [20]. Then, the optimal retrial rate, $\hat{L}_{c,i}$, equals $L^*(\hat{\theta}_{c,1},\gamma)$ in the non-cooperative case. Hence the joint optimum of the queue-length limitation and the retrial rate is $(\hat{L}_{c,1},\hat{\theta}_{c,1})$ when opportunists are the Stackelberg leader and non-cooperative.

A.4. Proof of Theorem 5.1 (1) From (5.7), the expected net benefit per unit time of the web service company in the model without an orbit space is $\lambda P-C_{h1}N_{nr}(L^*_{nos})$. Since the manager is the Stackelberg leader and the opportunists are non-cooperative, from (5.8) we find that the expected net benefit per unit time of the web service company is

$$\begin{aligned}&C_2N_{orbit}(\hat{L}_{m,1},\hat{\theta}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1}))\bar{N}(\hat{L}_{m,1},\hat{\theta}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1}))\\&+\lambda P-C_{h1}N_{slow}(\hat{L}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1}))\\&-C_{h2}N_{orbit}(\hat{L}_{m,1},\hat{\theta}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1})).\end{aligned}$$

Moreover, (5.10) can be rewritten as

$$\begin{aligned}&\lambda P-C_{h1}N_{nr}(L^*_{nos})<\\&C_2N_{orbit}(\hat{L}_{m,1},\hat{\theta}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1}))\bar{N}(\hat{L}_{m,1},\hat{\theta}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1}))\\&+\lambda P-C_{h1}N_{slow}(\hat{L}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1}))\\&-C_{h2}N_{orbit}(\hat{L}_{m,1},\hat{\theta}_{m,1},q^e(\hat{L}_{m,1},\hat{\theta}_{m,1})).\end{aligned} \tag{A.4}$$

The left side of (A.4) is the expected net benefit per unit time of the web service company in the model without an orbit space, and the right side is the expected net benefit per unit time of the web service company in the model with an orbit space. Therefore, we can see that the web service company can obtain more benefits in the model with an orbit space.

(2) A similar method can be used to prove the cooperative case.

A.5. Proof of Theorem 5.2 We ignore the proof of Theorem 5.2, since it is similar to the proof of Theorem 5.1.

Sheng Zhu received the B.Sc. degree from Fuyang Normal University, Fuyang, China, in 2004, the M.Sc. degree from Chongqing University, Chongqing, China, in 2007, and the Ph.D. degree from Beijing Jiaotong University, Beijing, China.

He is an associate professor in the School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, China. He is a member of the Operations Research Society of China (ORSC). His research interests include queueing theory, the applications of game theory and queueing theory in wireless communication and cloud computing, financial mathematics and engineering. He have published more than 10 papers in the proceedings of international conferences and international professional journals such as IEEE Transactions on Vehicular Technology, IEEE Systems Journal, Operations Research Letters, Operational Research, IMA Journal of Management Mathematics, Journal of Industrial and Management Optimization, etc.

Jinting Wang received the B.Sc. degree from Hebei Normal University, Shijiazhuang, China, in 1994, the M.Sc. degree from Hebei University of Technology, Tianjin, China, in 1997, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2000. He is a Distinguished Professor at the School of Management Science and Engineering, Central University of Finance and Economics, Beijing, China. His research interests include issues related to operations research, management science, queueing theory, reliability and the applications of game theory and queueing theory in wireless communication and networking. He has published over 150 peer-reviewed articles in international journals such as Operations Research, Manufacturing & Service Operations Management, IEEE Transactions on Vehicular Technology, IEEE Transactions on Cognitive Communications and Networking, Production and Operations Management, Queueing Systems, European Journal of Operational Research, Journal of Multivariate Analysis, Journal of Network and Computer Applications, etc. He is a member of the Operations Research Society of China (ORSC), and now he serves as the President of Reliability Society affiliated with ORSC. He was the recipient of the Outstanding Research Award for Young Researchers from ORSC in 2004. In 2011, he was honored with the Program for New Century Excellent Talents in University by the Ministry of Education of China. Dr. Wang is currently serving as Associate Editor for several professional journals such as Journal of the Operational Research Society, International Journal of Operations Research, International Journal of Smart Grid and Green Communications and other two Chinese journals.

Wei W. Li (M’99¨CSM’06) received the B.Sc. degree from Shaanxi Normal University, Xi¡¯an, China, in 1982, the M.Sc. degree from the Hebei University of Technology, Tianjin, China, in 1987, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 1994. He was once a tenure track Assistant Professor with the Department of Electrical and Computer Engineering, University of Louisiana at Lafayette, Lafayette, LA, USA, and was a tenured Associate Professor with the Department of Electrical Engineering and Computer Science, The University of Toledo, Toledo, OH, USA. Currently, he is a Professor with the Department of Computer Science and the Director/PI of the NSF-CREST Center for Research on Complex Networks, Texas Southern University, Houston, TX, USA. He is also the author/coauthor of six books and over 160 peer-reviewed articles in professional journals and the proceedings of international conferences. His research interests include dynamic control and optimization of energy-efficient wireless sensor networks, evaluation, complexity, power connectivity, and coverage for wireless sensor networks, adaptation, design, and implementation of dynamic models in wireless multimedia systems, theoretic foundations and advanced analysis in real-time, hybrid, and embedded systems, and performance evaluation of queuing networks and reliability systems. He is serving or has served as a Steering Committee Member, the General Co-Chair, the Technical Program Committee Co-Chair, the Track Chair, or a Technical Program Committee Member, for a number of professional conferences, including INFOCOM, ICDCS, Globecom, ICC, and WCNC, et al. Currently, he is serving as an Editor for several professional journals.