IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 19, 2019, accepted June 14, 2019, date of publication July 3, 2019, date of current version August 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926327

Optimal Overbooking Policy for Cloud Service Providers: Profit and Service Quality

MENGDI YAO[1], DONGLIN CHEN[1], AND JENNIFER SHANG[2]
[1]Department of Economy, Wuhan University of Technology, Wuhan 430070, China
[2]Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA

Corresponding author: Mengdi Yao (mengdiyaowhut@qq.com)

This work was supported by the National Natural Science Foundation of China under Grant 71172043. The work of M. Yao was supported by the China Scholarship Council in 2016.

ABSTRACT A cloud federation is a current paradigm that enables partnered cloud providers to share idle capacities during low demand periods and to purchase spare resources during demand spikes. In this research, we propose an optimal overbooking policy to maximize federation members' profits and enhance cloud users' experiences. The proposed policy overcomes cloud providers' low utilization and increases their profits. Under the market-oriented cloud federation system, we use the number of idle resources in the cloud federation and the operational costs of those resources to help the provider decide on its instance exchange price. Under such a price mechanism, we develop an optimal overbooking model and identify the conditions necessary for optimal solutions. Through implementing the optimal mechanism, we observe that the proposed overbooking policy can improve a federated provider's profits and decrease the probability of service level agreement (SLA) violation. When a provider's capacity is relatively large and the provider adopts the proposed overbooking policy, it could achieve maximum profits and decline its SLA violation when it has unmet customer demands and there are idle resources in the cloud federation. Through establishing the cooperative game model of the cloud federation, we make a reasonable profit distribution based on Shapley value. The cloud provider's profits and the probability of the SLA violation change as the instance price, the distribution of unserved customers, the number of federation members and penalty cost change. Compared with the other overbooking policies and no overbooking mechanisms, our research improves profits and reduces cloud provider's overbooking risks, thereby presenting a win-win situation for both the individual providers and the cloud federation.

INDEX TERMS Cloud federation, overbooking policies, optimization profit distribution, revenue management.

I. INTRODUCTION

Driven by the rapid growth in the Internet of things(IoT), real-time big data, and the adoption of service oriented architectures and Web 2.0 applications, the emergence of cloud computing is rapidly gaining momentum as an alternative to traditional IT [1]. Cloud computing enables convenient on-demand access to a shared pool of configurable computing resources that can be rapidly allocated to users with minimal management effort [2]. According to a Forrester research report, the global cloud computing market will reach $241 billion by the end of 2020 (see Forecaster Research Inc.). With a growing number of cloud-service users in the virtual world, to remain competitive, cloud providers must ensure that their resources are highly utilized and that their costs are reasonably low.

However, the random service requests from various potential users and the often overestimated demand regularly result in low utilization of the data center of cloud providers. A report from Google concludes that only 53% of the available memory is used, whereas CPU utilization is, on average, as low as 40% [3]. The low utilization of data centers can be attributed to two factors. (i) Because customers (especially small and medium enterprises (SMEs)) cannot accurately estimate demand, they prefer to reserve more for contingency, as evidenced by the usage information of Google's data centers. (ii) Resource requirements of multiple jobs on the same machine usually do not jointly reach peak capacity due to

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

aggregate effect. When allocating resources solely based on users' requests, the data center usually experiences low utilization and a high rejection rate to subsequent demand. Thus, it is important to develop strategies to exploit these resources fully. For example, overbooking a cloud provider's resources may be an appealing alternative to enhance utilization and improve efficiency.

The goal of a cloud provider is to maximize utilization and profit, whereas that of cloud customers is to find a high quality service. A service level agreement (SLA) in this research refers to a cloud-computing contract between cloud providers and customers. Such an agreement stipulates the level of service customers expect to receive from the provider and is 100% applicable to our situation. The overbooking policy involves a tradeoff between high utilization/profits and service level. Not providing sufficient capacity for customers who have already reserved the service is a violation of the SLA [3]. If the SLA terms are not fulfilled when the customers use the reserved instances, the cloud provider must pay a penalty to the users. Reference [4] finds that an overbooking policy increases the current revenue of service providers, but it reduces their future revenue. Thus, the benefits derived from an overbooking policy are not sustainable in the long term if the quality of service suffers.

The establishment of a cloud federation enables cloud providers to reduce the possibility of SLA violations by sharing their resources with each other in a resource pool during peak times. Cloud providers can join with others in the cloud federation to offer more services. In recent times, a cloud federation has emerged to allow individual cloud providers to cooperate for the purpose of balancing loads and accommodating spikes in demand. For example, in 2012, cloud providers, including Atos, EMC, and VMware formed an open cloud federation. Similarly, Alibaba, Lenovo, BAIDU and other cloud providers also formed an alliance. The presence of more cloud federations offers two important benefits to cloud providers. First, it enables providers to generate more revenue from computing resources that would otherwise be idle or underutilized. Second, a cloud federation enables cloud providers to expand their geographic footprints and accommodate demand surges without building new points-of-presence [5]. Cloud providers can then access global services without increasing capacity and reduce the chance of violating the SLA and incurring overbooking penalties. Moreover, cloud providers can sell their idle resources to other providers through the cloud federation and generate more revenue [6].

Given the overbooking problem in clouds and the feature of cloud federations, we attempt to address the following questions:

(i) What is the best overbooking policy and the ideal overbooking quantity for cloud providers to adopt?

(ii) How can cloud providers allocate the two separate instances to customers to maximize revenue?

(iii) What is the best cloud federation exchange price for federation members?

(iv) How can cloud federations reduce the probability of SLA violation of the various clouds? To what degree can they reduce the SLA violation of the clouds due to the overbooking policy?

(v) How to realize the fair and equitable profit distribution of the cloud federation members?

To solve the questions regarding the overbooking of cloud resources, we reference and learn from the overbooking strategy of a traditional industry, i.e., the airline alliance [7]. However, unlike the fixed exchange price adopted as part of the airline overbooking strategy [8], to manage the allocation of resources in a cloud federation, we adjust the exchange price dynamically. Through exchanging resources with other cloud members in the cloud federation, the proposed overbooking strategy improves the performance of the cloud providers and maximizes the potential of cloud service to the potential users. Furthermore, due to the disruptive innovations of the information technology industry, improving resource utilization of cloud computing is a key research point. Studying the overbooking strategy to improve the utilization of cloud computing resources extends the research field of overbooking and revenue management.

In this study, we aim to balance between utilization and SLA violations, meanwhile, maximizing profits of cloud providers. We apply two pricing models to allocate capacity and meet customer demand. For example, on-demand instance offers customers service at any time but at a higher price, while the reserved instance provides the resources to the users at a lower contract price. To mitigate the waste of unclaimed reservations, we propose an optimal overbooking model to increase resource utilization. Unlike the current overbooking policy, which mainly considers the single service provider's strategy, through adjusting the exchange price dynamically based on a cloud federation trading mechanism, we propose an optimal overbooking policy to ensure cloud providers exchange resources with cloud federation members. From this resource exchange, cloud providers and other members can maximize their profits/utilization and achieve high customer satisfaction. Through analyzing the corporative game model of cloud federation, the profit can be distributed reasonably based on Shapley value.

This paper is organized as follows. Section 2 reviews the literature on the overbooking policy and the cloud federation. Section 3 describes the trading mechanism and unit exchange price of the cloud federation. Section 4 presents a comprehensive analysis of the overbooking model. Section 5 establish a cooperative game model to distribute profit of cloud federation members based on Shapley value. In Section 6, we conduct numerical studies to examine the impact of the cloud federation on the overbooking policy and on the quantity, profits, and probability of SLA violation associated with cloud providers. Section 6 presents this study's conclusions and identifies future research directions.

II. RELATED WORKS

A. CLOUD FEDERATIONS

A cloud federation is a new prototype that helps cloud providers address resource limitation issues during peak demand by exchanging requests with other federation members. With respect to the structure of a federation, Calheiros *et al.* [2] propose a cloud coordinator to increase performance, reliability, and scalability. Grozev and Buyya [8] propose a cloud federation architecture and application agent mechanism to improve QoS, reliability and cost efficiency. Villegas *et al.* [9] design a cloud federation architecture mediated by a broker and cloud providers based on a layered cloud model. Under a cloud federation, Goiri *et al.* [6] help cloud providers decide when to outsource to cloud providers, when to rent free resources to other providers, and when to turn off unused nodes to achieve revenue maximization. Alternatively, to increase cloud providers' profits, Calheiros textitet al. [2] consider instance prices and spot instance features. Villegas *et al.* [9] define a cloud federation that shares different service layers to increase dynamic scalability and resource utilization. Finally, Yang *et al.* [10] coordinate multiple cloud providers to serve Real-time Online Interactive Applications (ROIA) and improve customer satisfaction, resource usage, and business performance.

In terms of resource allocation in the cloud federation, Giacobbe *et al.* [11] study the cloud federation from the perspective of sustainability and cost savings. Through resource pricing, resource allocation, resource discovery and disaster management, various studies provide data and information to maximize cloud providers' profits. Finally, Celesti *et al.* [12] propose a new strategy that makes use of efficient satellite transmissions to transfer huge amounts of data among federated clouds. In addition, the game theory is applied into the resource allocation and profit distribution in the cloud federation. Toosi *et al.* [5] establish the formation algorithm of cloud federation based on agent, adopt the cooperative game to prove that sharing resource can minimize cost. Hassan *et al.* [13] calculate the optimal solution of resource allocation in the cloud federation based on the cooperative game. Li *et al.* [14] propose a revenue distribution plan according to the core of cooperative game and discuss the stability of the cloud federation's structure.

It is evident that the extant literature on cloud federation focuses on the federation structure, resource management, resource allocation and profit distribution. Our research differs in that we establish a market exchange price to help cloud providers trade their resources and thereby optimize overbooking decisions and distribute profits in a cloud federation environment.

B. OVERBOOKING POLICY

1) OVERBOOKING IN CLOUD COMPUTING

With the growing interest in cloud computing, researchers have begun to focus on enhancing cloud providers' performance and profits. Scholars have studied the problems from different perspectives, e.g., capacity control [5], [15] dynamic pricing [3], [16] [17], service management based on SLA, and client classification [18] to improve profits and resource utilization. In addition, the overbooking policy as a way to improve resource utilization has also been studied. For example, Householder *et al.* [19] confirmed the low resource use in data centers, whereas Wu *et al.* [20] focus on virtual machine oversubscription. Similarly, Breitgand and Epstein [21] discuss bandwidth overbooking, and examine the issue of memory overbooking [22] .

Cloud providers optimize resource utilization through overbooking, but when the methods fail, the results may affect all customers who are using the cloud service and may cause the termination of entire servers in the data center. Compared with the overbooking problems in other industries, cloud providers suffer more damage as the overbooking problem not only angers unserviced customers, but the whole load balance breaks down, which leads to losses for entire data centers.

To manage the problems associated with overbooking, Wo *et al.* [23] propose a traffic-aware strategy that can satisfy the QoS requirements and guarantee performance. Tomas and Tordsson [24] propose a three-level QoS scheme for overbooking to avoid performance degradation and increase utilization. Finally, Breitgand and Epstein [21] advance an algorithmic framework to estimate physical capacity based on the SLA to reduce the risk of overbooking.

2) OVERBOOKING IN GENERAL

As an important strategy of revenue management, overbooking has been applied to various industries, including aviation [7], hotels [25], car rentals [26], and restaurants [27]. Soerag *et al.* [28] propose an effective revenue management model that incorporate customer choice behaviors and cancellations. Similarly, Lopez-pires *et al.* [29] compare the impact of no-shows and cancellations on overbooking under dynamic and static policies and analyze the impact of refunds and denied boarding costs. In healthcare management, Liu and Ziya [30] find that patient show-up probabilities and patient sensitivity to delays are key determinants in the overbooking policies. The studies noted above focus on single service providers. For two service providers, Chen and Hao [7] indicate that an overbooking policy that includes a co-operation agreement increases expected profits and service levels of two airlines. Similarly, Huang *et al.* [4] considered parallel substitutable flights, and through dynamically setting overbooking limits, they propose an effective overbooking policy.

Different from the literature, we propose a federation model to simultaneously address the overbooking and underusage situations. By dynamically adjusting the exchange price, the federation members can exchange resources to enhance utilization. We then derive the optimal overbooking policy to maximize individual and federation members' profits and reduce the probability of SLA violation. Table 1 compares the most relevant literature and positions our research.

TABLE 1. Related literature on overbooking policies.

paper	Research field	Customer behavior	Provider's behavior	Maximize revenue/profit	Resource utilization	Service level	Capacity control
Sierag et al(2015)		√		√			
Liu and Ziya(2014)	Health Care	√		√			
Chua et al.(2016)		√					√
Chen and Hao(2013)	Airline	√	√	√		√	
Huang et al.(2013)	Airline	√	√	√			√
Tomas and Tordsson(2014)	Single Cloud provider	√			√	√	√
López-Pires, Fabio et al.(2018)	Single Cloud Provider	√		√		√	
This paper	Cloud Federation	√	√	√	√	√	

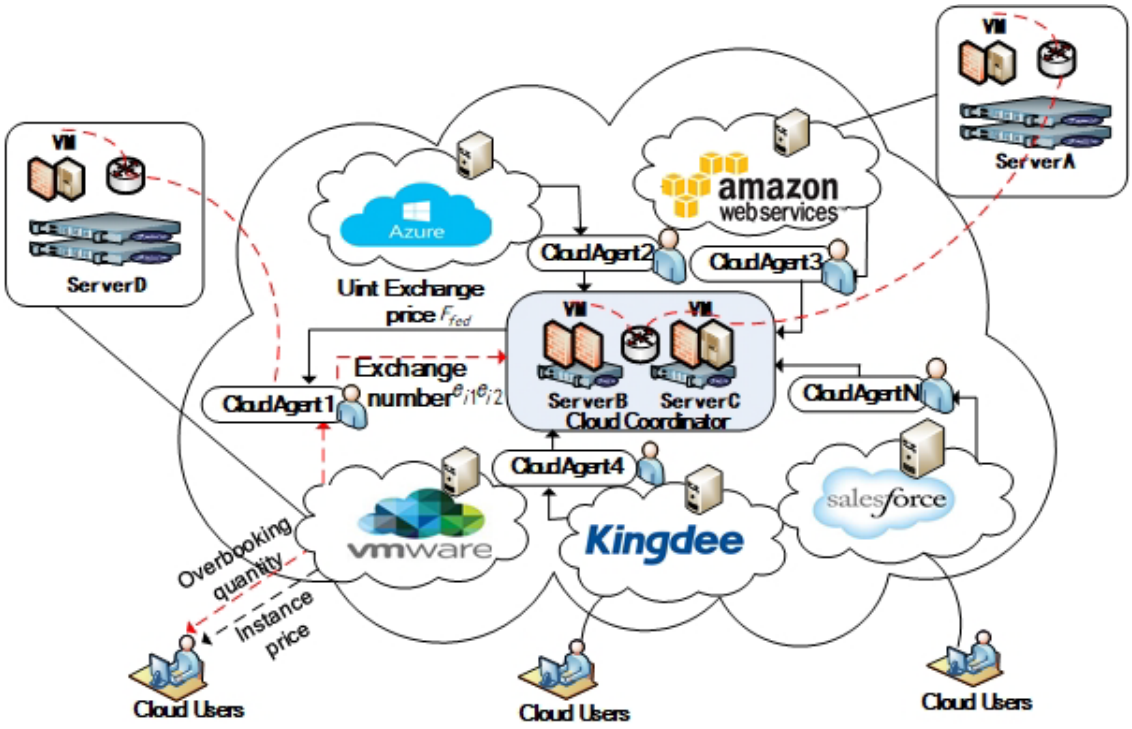


FIGURE 1. Market-oriented cloud federation structure and the process of VM placement.

III. CLOUD FEDERATION PRICE MECHANISM

A cloud federation, which is a union of two or more service providers, distributes and manages various internal and external cloud computing services to meet business needs. Members of the federation share resources by trading idle capacities during low demand periods and acquiring spare resources during spikes. Through cooperation, they sharing the resources to achieve more profit. The computing resource instance in this research is defined as the number of virtual machines (VMs) available. Inspired by the InterCloud project [2], we conduct our research based on the cloud federation structure and the process of VM placement presented in Fig.1.

A cloud federation contains multi-cloud providers, e.g., Amazon, Microsoft Azure, Salesforce, and VMware, and their agents for the purpose of exchanging resources dynamically through a cloud coordinator, i.e., virtual cloud platform operator, who matches surplus capacities of federation members to other members who are in need of extra capacities. Namely, a cloud federation provides a market-oriented structure that helps maximize the utilization of VM resources. Through the exchange of resources among members, cloud providers in the federation can achieve a win-win situation. Such a federation allows cloud providers to overbook VM utilization, while avoiding the violation of an SLA by acquiring idle VM resources at lower prices from other members.

TABLE 2. Notations of parameters and variables.

Notation	Symbol	Definition
Index	M	Number of cloud federation's member
	C_j	Capacity of a cloud provider's VM resource, j=1,2,⋯,m
	I_{ic}	Idle capacity of cloud federation members, j=1,2,⋯,m
	π	Total profit of a cloud provider under overbooking policy
Parameters	p_o	Unit price of on-demand instance
	p_c	Cloud provider's unit operational cost
	β	Decision rate of reserved instance, $\beta \in (0,1)$
	α	Unit upfront reservation fee of reserved instance
	g	Unit penalty cost for cloud members who un-serviced the cloud users
Variables	$f(\bullet)$	Probability density function of reserved but unused instance
	$F(\bullet)$	Cumulative distribution function of reserved but unused instance
	d_o	History demand of on-demand instance
	D_o	Capacity of on-demand instance
	r	Reserved capacity of reserved instance
	o	Overbooking number of reserved instance
	N_j	Number of reserved but unused instance of every cloud provider expect provider 1
	F_{fed}	Unit exchange price of cloud federation
	e_{i1}	Purchase number from cloud federation to cloud providers
	e_{i2}	Selling number of idle resource from cloud providers to cloud federation

Cloud users may include software developers, i.e., individuals or SMEs, social websites, e.g., LinkedIn, E-business websites, e.g., EBay, government websites, e.g., city administration. They can visit cloud providers to obtain computing or storage services (VM resources). Each cloud provider has its own cloud agent who receives information from the cloud provider and communicates with the cloud coordinator. They know that other members are all in the same situation and cooperative with each other to maximize profit of the federation. According to cooperative game model, all of them distribute profit fairly and reasonably. The cloud coordinator serves as a liaison that matches surplus capacity with unmet demands, i.e., shortage of its own resources. The cloud coordinator determines the exchange unit price that the cloud provider must pay to procure the additional resources from the federation.

A pricing model is proposed for federation members to exchange resources. The notations necessary for the development of the model are summarized in Table 2.

Under a cloud federation, cloud providers, e.g., Amazon EC2, may choose on-demand pricing or reserved pricing. Under on-demand pricing, users pay for the actual usage (pay-as-you-go) at a unit price. Amazon EC2 (elastic cloud) offers both one-month and one-year billing cycles. Within the billing cycle, the unit price of an on-demand instance is constant, and customers continue to receive the service until logging out. Under the reserved pricing plan, users pay an upfront reservation fee α to reserve an instance for a period time, and they can request service at will. If there is no available instance for the reserved user, the user can be compensated by cloud providers. In this instance, the unit price of the reserved instance is βp_o, which is less than the on-demand unit price. In Amazon EC2, when using all upfront pricing options for one year with an m4.large reserved instance (m4.large, Lunix, US-east), the upfront fee is \$504/year, which is 43% less than that of the on-demand instance. The provider is required to guarantee resource availability. However, as the reserved capacity is often underutilized, the result is wastage.

A cloud coordinator, e.g., IBM Angle or CloudSim, can determine resource allocation, exchange prices, entry/exit from the federation, and SLAs. Cloud providers with idle resources or insufficient capacity send this information to the cloud coordinator through their agents. The cloud coordinator then examines the idle resources of all cloud federation members and determines the exchange price and SLAs, which apply to all members. In this research, the probability of SLA violation is the percentage of reserved capacity unmet by the service provider. The instant hourly exchange price of a VM in the cloud federation is defined as F_{fed}

$$F_{fed} = \frac{\sum_{j=2}^{m} C_j - \sum_{i=2}^{m} I_i}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c$$

F_{fed} is the unit exchange price of the cloud federation and fluctuates with the number of idle resources in the federation. The value of exchange price F_{fed} is a sequences data, which means the value of is static at a certain time, $\sum_{j=2}^{m} C_j$ is the total capacity of the cloud federation. $\sum_{j=2}^{m} I_j$ is the idle capacity of the cloud federation. p_o is the price of on-demand instance, i.e., the maximum possible unit price, and p_c is the cloud provider's operational cost, e.g., hardware cost, electricity cost, and serves as a proxy for the lowest price . $\frac{\sum_{j=2}^{m} C_j - \sum_{j=2}^{m} I_j}{\sum_{j=2}^{m} C_j}$ is the utilization rate of the overall federation capacity, excluding the specific provider who requires more resources. Thus, F_{fed} must be larger than the unit operational cost, which is the minimal cost to maintain the operation of the VMs.

To improve resource utilization, cloud computing service providers often overbook. Hence, we contend that cloud federation members can benefit by adopting an overbooking policy to increase profits, and therefore, we propose an overbooking policy for a cloud federation.

IV. OVERBOOKING MODEL FOR A CLOUD FEDERATION

A. TOTAL PROFIT OF AN OVERBOOKING MODEL

Suppose the capacity of a cloud provider's VM resources is C. To maximize profits, the cloud provider often allocates its capacity to cloud users under various instances. If the cloud provider allocates r reserved instances, and historically the on-demand is d_o instances, then the capacity allocated to the on-demand market is $D_o = \min(C - r, d_o)$.

To enhance the resource utilization of the federation, the provider will overbook units of reserved instances. Cloud users conventionally reserve more instances than their demand due to uncertainty. We set the total number of reserved but unused instances of every cloud provider as random variables $N_j (j = 1, 2, \cdots, m)$, where the probability density function is $f_j(\cdot)$ and the cumulative distribution function is $f_j(\cdot)$. These random variables are independent of each other. The capacity acquired by a cloud provider i from the federation is e_{i1}, whereas the idle capacity sold to the federation is e_{i2}, and g is the unit penalty cost for each reserved instance whose demand is not satisfied.

To develop an overbooking policy for cloud providers, we assume

(1) The service level agreement (SLA) of all cloud federation members must be at the same level. As the cloud service and its data center are virtual, mobile and secure, cloud users are oblivious to where their tasks are stored or executed. Namely, when there is a resource transfer among cloud members, cloud users will accept the service without any objection.

(2) The instances reserved but unused among cloud providers are independent. Although they could exchange resources through the cloud federation, they are independent and aim to maximize their own performance.

(3) $g > p_o$,$g = (1 + \xi)F_{fed}$, where $0 < \xi < 1$. When instances are reserved but unserved, the cost of such an SLA violation is much higher than the on-demand instance price. The violation of an SLA not only incurs a penalty but also causes low customer satisfaction. In addition, the penalty is higher than the unit exchange price, indicating that the cloud provider prefers to purchase resources to meet reserved but unmet instances.

Let the total number of reserved but unused instances by a member-Provider i be N_i. The units of instances Provider i must purchase from the cloud federation is $(o_i - N_i)^+$. The total instances available in the federation are $\left(N_j^{'} - o_j^{'}\right)$, where $N_j^{'}$ refers o the distribution of the number of reserved but unused instances for the $(m-1)$members. $N_j^{'}$ is $(m-1)$ dimensional random variable $(N_2, N_3, \cdots, N_m)$, which the probability density function is $f_j(n_2, n_3, \cdots, n_m)$ and the cumulative distribution function is $F_j(n_2, n_3, \cdots, n_m)$. $o_j^{'} = \sum_{j=2}^{m} o_j$, which is the total overbooking quality of the (m-1) members in the federation. Thus, the number of instances (e_{i1}) Provider i acquires from the federation is expressed as

$$
\begin{aligned}
e_{i1} &= \min[(o_i - N_i)^+, (N_j' - o_j')^+] \\
&= \begin{cases} o_i - N_i, & (N_j' - o_j') \geq o_i - N \geq 0 \\ (N_j' - o_j'), & o_i - N \geq (N_j' - o_j') \geq 0 \\ 0, & otherwise \end{cases}
\end{aligned}
$$

Conversely, when total idle instances for Provider i is the unit instances the cloud federation must acquire from Provider i is. Thus, the number of instances (e_{i2}) Provider i can transfer to the federation is expressed as

$$
\begin{aligned}
e_{i2} &= \min[(N_i - o_i)^+, (o_j' - N_j')^+] \\
&= \begin{cases} N_i - o_i, & (o_j' - N_j') \geq N_i - o_i \geq 0 \\ (o_j' - N_j'), & N_i - o_i \geq (o_j' - N_j') \geq 0 \\ 0, & otherwise \end{cases}
\end{aligned}
$$

Here, we denote (o_i) as the overbooking policy for the cloud provider when it trades with the $(m-1)$ members of the cloud federation. We formulate the overbooking problem as an optimization problem, that is, the optimal overbooking level (o_i) is the one that maximizes the profit function $\pi_i(o_i)$ is expressed as

$$
\begin{cases}
\max \pi_i(o_i) = \pi_{io} + \pi_{ir} \\
= p_o D_o + \alpha(r_i + o_i) + \beta p_o[r_i - (N_i - o_i)^+ + e_{i1}] \\
-F_{fed} \cdot e_{i1} - g[(o_i - N_i)^+ - e_{i1}] + F_{fed} \cdot e_{i2} \\
s.t.\ o_i \geq 0
\end{cases} \tag{1}
$$

Note that the overbooking quantity o_i is the only decision variable to answer how many VM resources should be overbooked for each cloud provider. Objective (1) aims at maximizing the profit of cloud provider. Constraint $o_i \geq 0$ guarantees the number of overbooking must be greater than zero.

Equation. (1) indicates that Provider i's profits are derived from the on-demand instance and the reserved instance. The first term is the revenue from the on-demand instance. The second term is the upfront fee of the whole reserved instance. The third term is the revenue received from actual use by reserved customers ($r_i-(N_i-o_i)^+$) and from customer demand (e_{i1}) served by the cloud federation. The fourth term is the amount paid for the cloud federation. The fifth term is the penalty cost ($g[(o_i - N_i)^+ - e_{i1}]$) paid for the overbooking of customers who reserved but were unserved. The last term is the revenue from idle resources from the cloud federation.

In addition, the unit price p_o, p_c, F_{fed} are not decision variables and they are assumed to determined somehow in advance. We consider the effects of different prices with numerical study in section VI. As stated by reference [18], the decision of pricing and overbooking quality are at different decision levels in reality. Therefore, this paper focus on examining the impact of overbooking quality o_i of each cloud provider on the profit $\pi_i(o_i)$ under the stochastic properties of the number of reserved but unused instance N_i.

Due to the stochastic nature of the total number of reserved but unused instance by a member provider N_i, the profit maximization problem is defined as a stochastic programming problem. Thus, we use the expect value model to solve the stochastic programming problem, and the problem is searching the optimal overbooking quantity to maximize expected profit. The expected profit of Provider i collaborating with a cloud federation is

$$\begin{cases} \max E[\pi_i(o_i)] = E[p_o D_o + \alpha(r_i + o_i)] + \\ E[\beta p_o (r - (N_i - o_i)^+) - g(o_i - N_i)^+] \\ +E[(\beta p_o + g)e_{i1}] - E[F_{fed} \cdot e_{i1}] + E[F_{fed} \cdot e_{i2}] \\ s.t.\ o_i \geq 0 \end{cases} \tag{2}$$

B. OPTIMAL OVERBOOKING QUANTITY

The optimal overbooking quantity can be derived by applying the FOC, i.e., first-order conditions, derivative of (2) with respect to o_i.

$$\begin{aligned} \frac{\partial E(\pi_i(o_i))}{\partial o_i} &= \alpha + \beta p_o[1 - \Pr(N_i \leq o_i)] - g\Pr(N_i \leq o_i) \\ &\quad + (\beta p_o + g)\Pr((N_j' - o_j') \geq o_i - N_i \geq 0) \\ &\quad - \frac{\partial(F_{fed} \cdot e_{i1})}{\partial o_i} + \frac{\partial(F_{fed} \cdot e_{i2})}{\partial o_i} \end{aligned} \tag{3}$$

The unit exchange price F_{fed} plays an important role in the cloud federation exchange as F_{fed} is influenced by the idle capacity in the cloud federation. The pricing mechanism is discussed based on the relationship between Provider i's needs and the surplus relative to those of the federation.

Case I: Federation surplus > Provider i's needs.

$$\begin{aligned} F_{fed} &= \frac{\sum_{j=2}^{m} C_j - \sum_{j=2}^{m} I_j}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c \\ &= \frac{\sum_{j=2}^{m} C_j - ((N_j' - o_j')^+ - e_{i1})}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c \end{aligned} \tag{4}$$

Case II: Provider i's needs > Federation surplus

$$\begin{aligned} F_{fed} &= \frac{\sum_{j=2}^{m} C_j - \sum_{j=2}^{m} I_j}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c \\ &= \frac{\sum_{j=2}^{m} C_j - ((N_j' - o_j')^+ - e_{i1})}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c = p_o \end{aligned} \tag{5}$$

Case III: Federation's needs > Provider i's surplus

$$\begin{aligned} F_{fed} &= \frac{\sum_{j=2}^{m} C_j - \sum_{j=2}^{m} I_j}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c \\ &= \frac{\sum_{j=2}^{m} C_j - ((o_j' - N_j')^+ - e_{i2})}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c \end{aligned} \tag{6}$$

Case IV: Provider i's surplus > Federation's needs

$$\begin{aligned} F_{fed} &= \frac{\sum_{j=2}^{m} C_j - \sum_{j=2}^{m} I_j}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c \\ &= \frac{\sum_{j=2}^{m} C_j - ((o_j' - N_j')^+ - e_{i2})}{\sum_{j=2}^{m} C_j}(p_o - p_c) + p_c = p_o \end{aligned} \tag{7}$$

We simplify the expression in (3) as follows:

$$\begin{aligned} B(o_i, o_j') &= \Pr((N_j' - o_j') \geq o_i - N_i \geq 0) \\ S(o_i, o_j') &= \Pr((o_j' - N_j') \geq N_i - o_i \geq 0) \end{aligned}$$

The term $B\left(o_i, o_j'\right)$ refers to the probability of transferring cloud users' tasks from Provider i to the federation, indicating Provider i should purchase idle resources and $S\left(o_i, o_j'\right)$ is the probability that Provider i will sell idle resources to the federation.

Substituting F_{fed} in (4) to (7) for that in (3), we have

$$\begin{aligned}
&\frac{\partial E\left(\pi_i(o_i)\right)}{\partial o_i} \\
&= \alpha + \beta p_o - (\beta p_o + g)\Pr(N_i \le o_i) + (\beta p_o + g)\cdot B \\
&\quad - [\frac{2(p_o - p_c)}{\sum_{j=2}^{m} C_j}(o_i - N_i) - \frac{(p_o - p_c)}{\sum_{j=2}^{m} C_j}(N_j' - o_j') + p_o]\cdot B \\
&\quad + [\frac{2(p_o - p_c)}{\sum_{j=2}^{m} C_j}(o_i - N_i) - \frac{(p_o - p_c)}{\sum_{j=2}^{m} C_j}(N_j' - o_j') - p_o]\cdot S
\end{aligned} \tag{8}$$

The SOC, i.e., second-order condition, of Provider i's expected profit with regard to o_i is as follows:

Define $b^1 \equiv \int_0^{o_i} f_j(o_i + (o_j' - N_j'))f_i(N_i)dN_i$, $b^2 \equiv \int_{o_j'}^{+\infty} f_i(o_i)f_j(N_j')dN_j'$, $s^2 \equiv \int_0^{o_j'} f_i(o_i)f_j N_j' dN_j'$, $s^1 \equiv \int_0^{o_j'} f_i(o_i + (o_j' - N_j'))f_j N_j' dN_j'$

$$\begin{aligned}
&\frac{\partial^2 E\left(\pi_i(o_i)\right)}{\partial^2 o_i} \\
&= (\beta p_o + g)f_i(o_i) - (\beta p_o + g - p_o)(b^1 - b^2) \\
&\quad - \frac{2(p_o - p_c)}{\sum_{j=2}^{m} C_j}(B - S) + \frac{2(p_o - p_c)}{\sum_{j=2}^{m} C_j}(o_i - N_i)(b^1 - b^2) \\
&\quad - \frac{(p_o - p_c)}{\sum_{j=2}^{m} C_j}(N_j' - o_j')(b^1 - b^2) \\
&\quad + \frac{2(p_o - p_c)}{\sum_{j=2}^{m} C_j}(o_i - N_i)(s^1 - s^2) \\
&\quad - \frac{(p_o - p_c)}{\sum_{j=2}^{m} C_j}(N_j' - o_j')(s^1 - s^2) - p_o(s^1 - s^2)
\end{aligned} \tag{9}$$

More specifically, we have

$$\frac{\partial^2 E\left(\pi_i(o_i)\right)}{\partial o_i^2} \approx -\frac{2(p_o - p_c)}{\sum_{j=2}^{m} C_j}\cdot(B - S) \tag{10}$$

where $f_i(o_i) = s^2$, $B\left(o_i, o_j'\right)$ and $S\left(o_i, o_j'\right)$ are slowly varying function. Thus, $(b^1 - b^2) << B$, $(s^1 - s^2) << S$.

From (10), we find that when $(B - S) > 0$, that is $N_j' - o_j' \ge o_i - N_i \ge 0$, then the expected profit (3) is a concave function for the given o_i^*. Therefore, the optimal overbooking quantity of the reserved instance can be achieved at the extreme point. Thus, the optimal overbooking quantity (o_i^*) satisfies

$$\begin{aligned}
&\frac{\alpha + \beta p_o}{\beta p_o + g} \\
&= F(o_i^*) - B(o_i^*, o'^*_j) \\
&\quad + \frac{2K(o_i^* - N_i) - K(N_j' - o'^*_j) + p_o}{\beta p_o + g}\cdot B(o_i^*, o'^*_j) \\
&\quad - \frac{2K(o_i^* - N_i) - K(N_j' - o'^*_j) - p_o}{\beta p_o + g}\cdot S(o_i^*, o'^*_j)
\end{aligned} \tag{11}$$

where $K = \frac{p_o - p_c}{\sum_{j=2}^{m} C_j}$.

Therefore, we propose the following.

Proposition 1:Joining with a cloud federation to exchange resources is a good overbooking strategy for a cloud provider (e.g., *Provider i).For Provider i, it will maximize profits* $E(\pi_i(o_i^*))$ *if the overbooking quantity* o_i^* *satisfies (11) when there are unserviced customers in Provider i and idle resources in the cloud federation,*i.e. $N_j' - o_j' \ge o_i - N_i \ge 0$

As SME users often lack exact information regarding their need for cloud services, they tend to overestimate their resource requirements and reserve more capacity as a buffer [15]. For many services, the peak workload exceeds the average workload, and since few users reserve resources for less than the expected peak, resources are idle at non-peak times [31]. Furthermore, as customer demand for cloud services is variable over the job lifecycle, the resources customers reserved are usually not fully utilized, and the unused capacity motivates providers to overbook their resources.

To improve resource utilization, cloud providers tend to overbook as much as possible. However, overbooking may result in unmet customer demand for cloud services. Under the cloud federation, the cloud coordinator will effectively allocate cloud members' resources to decrease the risk of overbooking [2]. Furthermore, the unmet demand in the cloud federation can be reasonably reduced through the help of the idle resources from other cloud federation members. Under such a situation, the cloud provider's profits can be maximized by purchasing idle resources from the cloud federation to meet its cloud customers' needs.

C. THREE KEY ELEMENTS FOR THE OPTIMAL OVERBOOKING MODEL

From the expected profits (2) and the optimal conditions of the overbooking policy (11), we find that the expected profits and optimal overbooking quantities are affected by three key elements: (i) the probability of buying idle resources from the cloud federation; (ii) the probability of selling idle resources; and (iii) the unit exchange price.

(i) Probability of buying idle resources $B\left(o_i, o_j'\right)$

The probability of buying idle resources from the cloud federation is related to the situation when there are unmet demands by the provider and idle resources in the cloud federation. This means that the cloud provider can obtain extra resources and improve its service level, which not only

increases profits but also reduces SLA violations. The cloud provider should pay more attention to more accurately forecasting customer demand to reduce the probability of unmet demands due to overbooking.

(ii) Probability of selling idle resources $S\left(o_i, o_j'\right)$

The probability of selling idle resources to the cloud federation is related to the scenario when there are unused instances and the cloud federation has unmet demands. It is a scenario often observed at non-peak times. As the unit exchange price is generally higher than the operational cost, it is advisable to increase expected profits by selling idle resources to the cloud federation.

(iii) Unit exchange price F_{fed}

The unit exchange price plays an important role in the cloud federation market mechanism. This price is affected by the probability of buying idle resources and the probability of selling idle resources under the cloud federation. Equations. (4) to (7) suggest that when the idle capacity of the cloud federation approximates the capacity of the cloud federation, the unit exchange price is at its minimum, i.e., it equals the operational cost of VM. The quantity of overbooking then increases because the purchase of idle resources from the cloud federation to serve its own cloud users at a lower price is always attractive as it enhances the cloud provider's profits. With the increasing probability of purchasing idle resources, the number of idle resources in the cloud federation is decreased. When the unit of unmet demand approximates the unit of idle resources in the cloud federation, the unit exchange price is at its maximum, i.e., it equals the on-demand instance price p_o. When all surplus resources are depleted, further demand cannot be satisfied, and the expected profit decreases owing to increases in penalty costs.

D. COMPARING OVERBOOKING PERFORMANCE WITH AND WITHOUT A CLOUD FEDERATION

We applied the following metrics to study the differences between a cloud provider's overbooking performances with and without a cloud federation membership.

(i) Profit.

When a provider, e.g., Provider i, does not join a federation, it will have the following expected profit:

$$E[\pi_i'(o_i')] = E[p_o D_o + \alpha(r_i + o_i')] + E[\beta p_o(r_i - (N_i - o_i')^+)] - E[g(o_i' - N_i)^+] \quad (12)$$

Thus, the optimal profit function satisfies $E[\pi_i'(o_i'^*)]$. When Provider i joins the federation, its optimal profit function (2) can be expressed as $E[\pi_i(o_i^*)]$. Defining $E[\pi_i(o_{i'}^*)]$ as any other non-optimal profit performance, we have $E[\pi_i(o_i^*)] \geq E[\pi_i(o_i'^*)]$.In addition,

$$E[\pi_i(o_i'^*)] - E[\pi_i'(o_{i'}^*)] = (\beta p_o + g)E(e_{i1}) - E(F_{fed} \cdot e_{i1}) + E(F_{fed} \cdot e_{i2}) \geq 0$$

where $F_{fed} \cdot e_{i1} \geq F_{fed} \cdot e_{i2}$.

Proposition 2:According to $E[\pi_i(o_i^*)] \geq E[\pi_i(o_i'^*)]$,cloud Provider i can improve its profits when it joins the cloud federation and exchanges resources with its members under the overbooking policy.

(ii) Service Level Agreement (SLA).

In cloud market, the SLA is an important performance metric. The cloud users and providers negotiate SLA agreement to ensure their profit and service quality. Referring to the analyses index—the number of SLA violations when using different overselling policies in Mario and Guitart [18], we define the probability of reserved capacity unsatisfied by the service provider without a federation as the probability of SLA violation and denote it as $\Pr(N_i \leq o_i'^*)$. Recall that $B\left(o_i, o_j'\right)$ is the probability of buying idle resources from the cloud federation. Thus, under the federation, Provider i's probability of not satisfying (rejecting) its customers is $\Pr(N_i \leq o_i'^*) \leq B(o_i^*, o_j^*)$.

Combined with (11), we find that

$$\begin{aligned}\Pr(N_i \leq o_i^*) - B(o_i^*, o_j'^*) &\leq \Pr(N_i \leq o_i^*) - B(o_i^*, o_j'^*) \\ &+ \frac{2K(o_i^* - N_i^*) - K(N_j' - o_j'^*) + p_o}{\beta p_o + g} \cdot B(o_i^*, o_j'^*) \\ &- \frac{2K(o_i^* - N_i^*) - K(N_j' - o_j'^*) - p_o}{\beta p_o + g} \cdot S(o_i^*, o_j'^*) \\ &= \frac{\alpha + \beta p_o}{\beta p_o + g} = \Pr(N_i \leq o_i'^*) \end{aligned} \quad (13)$$

Proposition 3:Cloud federation market could reduce the probability of SLA violation by sharing resources among federation members.

V. THE PROFIT DISTRIBUTION BASED ON SHAPLEY VALUE

When the unit exchange price and the optimal overbooking quality was calculated, the Cloud Provider i can achieve the profit with cloud federation. Therefore, the members should play a game to distribute profit. The game behavior can be simulated base Shapley value. The Shapley value method is a mathematical method proposed by Shapley to solve the problem of interest distribution in n cooperation. It's based on the marginal contribution of cloud providers to allocate the total profit. The only solution of cooperative game can be obtained by using Shapley value. Here we assume that,

(iv) For the provider i, it cooperative with other cloud federation members if and only if the overbooking profit under cloud federation is more than that without a cloud federation membership, that is,$\varphi_i(v) \geq E[\pi_i'(o_i'^*)]$. Here, the $\varphi_i(v)$ refers to the value of profit distribution in the cloud federation.

(v) There exist a reasonable profit distribution to guarantee the establishment of cloud federation. It is obvious that the sum of the profit that each member exchange resource is equal to the total profit they cooperative, that is, $\sum_{j=2}^{m} \varphi_j(v) = v(M)$. Here, $v(M)$ is denoted as the maximize profit of all cloud federations.

TABLE 3. The instance types and unit prices of cloud provider.

Instance type	a1.large	a1.xlarge	a1.2xlarge
On-demand instance price	$p_o = 0.054\$ / hour = 446.76\$ / year$	$p_o = 0.108\$ / hour = 948.08\$ / year$	$p_o = 0.204\$ / hour = 1787.04\$ / year$
Reserved instance price	α=0， $\beta p_o = 0.68 p_o$=303.80$/year	α=0， $\beta p_o = 0.68 p_o$=652.02$/year	α=0， $\beta p_o = 0.68 p_o$=1252.18$/year
	α=154$， $\beta p_o = 0.6 p_o$=268.06$/year	α=308$， $\beta p_o = 0.6 p_o$=568.85$/year	α=617$， $\beta p_o = 0.6 p_o$=1072.22$/year

TABLE 4. The capacity information of each cloud federation member.

	The total number	Reserved instance $D_O \sim U(0,0.4)$	Average reserved but unused instances
Cloud Provider 1 (Amazon EC2 instance)	1250	1000	100
Cloud Provider 2	200	160	16
Cloud Provider 3	250	200	20
Cloud Provider 4	300	240	24

We assume that an operation τ is the array of the cloud federation $M = \{1, 2, \ldots, m\}$, $(\tau \boldsymbol{M}, \boldsymbol{y})$ is the substitution game of $(\boldsymbol{M}, \boldsymbol{v})$.We define $y(\cdot)$ is

$$y(\tau(j_1), \tau(j_2), \cdots, \tau(j_k)) = v(j_1, j_1, \cdots, j_k), \quad \forall k = \{j_1, j_1, \cdots, j_k\} \in 2^M / \Phi \quad (14)$$

It is obvious that, and are same game.

According to the (2), when Provider i joins the federation, its optimal profit function can be expressed as $E[\pi_i(o_i^*)]$. Each member want to maximize its profit through playing game, that is, $v(i, j) = \sum_{j=2}^{m} \pi_j$.

The value of the Shapley formula is

$$\varphi_j(v) = \sum_{s \in L(j)} \frac{(s-1)!(m-s)!}{m!} [v(S) - v(S - \{j\})] \quad (15)$$

In the profit distribution process of the cooperative game, we conclude that each member under same situation can gain the same profit in (14); from (15), it shows that is the linear function of, that is, The new game is the direct addition of the original two games after the combination of two independent games. In other words, the cooperative game model in the cloud federation under overbooking policy super additive and is have nonempty core. The value of Shapley is as the value of profit distribution of the cooperative game. It is based on the marginal contribution of cloud providers to allocate the total profit, which can make the formation and development of the cloud federation.

VI. CASE STUDY

The overbooking strategy of cloud providers in a cloud federation environment is depicted through a case study in this work. The purpose of the case study is to illustrate the process of resource exchange under the optimal overbooking policy between cloud federation members. Amazon EC2, the largest cloud provider, is openly recruiting the cloud providers to establish a cloud federation. Four cloud providers (Amazon EC2, Windows Azure, VMware, and Rackspace) denoted as the cloud federation members. The instance types and prices in the case study are taken from the real cloud Amazon EC2, which is shown in Table 3.Tha capacity of cloud federation members in given in Table 4 are randomly assigned in the case study. Other related data are shown in the numerical study to examine the model.

A. OPTIMAL OVERBOOKING STRATEGY FOR CLOUD PROVIDER 1

In this section, we conduct a numerical study to explore, whether the overbooking strategy with joint cloud federation (Over With) could improve the profits and decrease the SLA violation comparing with the no overbooking (No Over) and overbooking without joint cloud federation (Over Without), how the prices, the distribution of unused customers, the penalty cost, and the number of cloud federation members impact the profits and SLA violation of Cloud provider1.

Based on the data collected from Amazon Elastic Compute Cloud (EC2) and referring to the parameter values in Reference [18], [29], the parameter values are as follows.

(i) *Price of on-demand and reserved instance:* We make use of the price data available in Amazon EC2 to conduct our experiment. For simply, we choose different types instance (Linux, Amazon East), that is, a1.large, a1.xlarge and a1.2xlarge. Take the instance type_a1.xlarge for example, the unit price of an on-demand instance for each cloud provider is $p_o = 0.108\$/\text{hour} = 948.08\$/\text{year}$. As for the reserved instance, this is a discount rateβ with different upfront reservation fee α . When the users choose no upfront reservation fee, $\alpha = 0$, the unit price of a reserved

instance $\beta p_o = 0.68p_o = 0.074\$/\text{hour} = 652.02\$/\text{year}$; if the users choose part upfront reserved fee $\alpha = 308\$/\text{year}$, the unit price of a reserved instance $\beta p_o = 0.6p_o = 0.07\$/\text{hour} = 568.85\$/\text{year}$. The instance types and prices offered are summarized in Table 3. (see https://aws.amazon.com/ec2/pricing/reservedinstances/pricing/. The operational cost of an instance is $p_c = 0.1p_o = 94.808\$/\text{year}$.

(ii) *Capacity of cloud providers:* As we do not know the exact capacity of the Amazon cloud, for illustration purposes, we assume the total number of Amazon's EC2's Instance Type to be 1250. Assuming there are an additional three members who form the federation with Amazon, each of the three members has different capacity and the details are shown in Table 4. We use the number of cloud provider members as the benchmark. Then, the impact of the number of cloud federation members on profit and SLA violation under different overbooking strategies will be analyzed in further.

(iii) *Proportion of on-demand instance:* Reference [23] assumed the correlation among all demand are equal. In order to focus on the impact of reserved but unused instance, we define that each federation member allocates a certain proportion of its capacity for on-demand instances according to a uniform distribution, $D_0 \sim U(0, 0.4)$.

(iv) *The distribution of the reserved but unused instances:* The reserved but unused instances follows a normal distribution $r \sim N(u, \sigma^2)$. Let $u = 0.4 * r_o$. If Amazon's EC2's on-demand average ratio is 0.2 (random number from $D_0 \sim U(0, 0.4)$), then its reserved units is 1,000 (=1,250∗.8) and its average reserved but unused instances are 100 (1000∗0.1), which are available to share with members. Similarly, the reserved units of the other three members are shown in Table 4. The standard deviations of the reserved but unused instance is one-third the sigma of the mean. We define this normal distribution as Case1. Then, the impact of the distribution of the reserved but unused instances on profit and SLA violation under different overbooking strategies will be analyzed in section IV-C.

(v) *The exchange price:* Under the cloud federation trading mechanism, a cloud provider can exchange resources at the exchange price, i.e., F_{fed} (see Section 3.2).

(vi) *The unit penalty cost:* If the SLA terms are not fulfilled when the customers use the reserved instances, the cloud provider must pay a penalty to the users. We assume the penalty rate ξ=[0.2, 0.4, 0.6, 0.8], then the penalty cost $g = (1 + \xi)F_{fed}$.

Using Matlab 7.0 and (11), we achieve the following results.

1) PROFIT GAINS

According to the cooperative game model (M, v), $M = \{1, 2, \ldots, m\}$,there are four cloud members in the cloud federation, which can form 8 sub-federation, i.e.,.{1},{1, 2},{1, 3},{1, 4},{1, 2, 3},{1, 2, 4},{1,3,4},{1, 2, 3, 4}Under the cloud federation {1, 2, 3, 4}, we calculate that the optimal overbooking quantity of the leading cloud provider (Provider 1) under the cloud federation is 104. There are four units of overbooking available from Provider 1 and six (reserved and unused (= 60)-overbooked units (= 54)) units in the cloud federation. The optimal trading strategy for cloud Provider 1 is to purchase four instances from the cloud federation. Accordingly, the federation will achieve the maximum profit. By applying (2), we find the optimal expected profit $v(\{1, 2, 3, 4\})$ is \$888,158.00/year. Similarly, we can get the other 6 optimal expected profit of sub-federation, $v(\{1, 2\}) = \$652, 672.04/\text{year}$, $v(\{1, 3\}) = \$653, 674.08/\text{year}$, $v(\{1, 4\}) = \$649, 988.31/\text{year}$,$v(\{1, 2, 3\}) = \$653, 976.58/\text{year}$, $v(\{1, 2, 4\}) = \$654, 078.12/\text{year}$, $v(\{1, 3, 4\}) = \$651, 432.03/\text{year}$.

Without a federation, the optimal overbooking quantity of the leading cloud provider (Provider 1) is 92, By applying (12), we find that the expected profit of Provider 1$v(\{1\})$ is \$884,097.00/year.

According to the Shapley value, we can calculate the distributed profit for Provider 1is \$890,326.19/year. Compared with the expected profit without a federation, the expected profit of Provider 1 and other members are increased by 0.705%, which are shown in Table 5.

2) THE PROBABILITY OF SLA VIOLATION

With respect to SLA violation, Table 5 presents the probability of SLA violation for Provider 1 under the federation decreases by 8.38% comparing with the overbooking without joint cloud federation strategy.

In addition, we compare our optimal overbooking policy under cloud federation environment (Over With) with overselling policies based on Revenue Maximization (Ovrs^{RM}) proposed by Mario and Guitart [18]. We compare the difference between our approach (Over With) with Over^{RM} from two aspects: SLA violation and profits under diverse factors.

B. CHANGES IN INSTANCE TYPES AND PRICES UNDER DIFFERENT OVERBOOKING STRATEGIES

As stated by Mario and Guitart [18], for cloud provider 1 to obtain higher profit, the price and overbooking quantity are important factors at the different revenue management decision making levels. To examine the impact of different types and prices on profit and SLA violation under different overbooking strategies, we choose three contract types of on-demand instance, and set two kinds of reserved ratio of the reserved instance from Amazon EC2. We analyses three overbooking strategies: No overbooking strategy (No Over), Overbooking without joint cloud federation strategy (Over Without), Overbooking with joint cloud federation strategy (Over With), and Over^{RM} policy.

We find in Fig.2. that all overbooking strategies have a positive impact on profit. Cloud provider 1 receives lower profits when applying No overbooking policy. While it will achieve higher profit with joint cloud federation strategy. Over^{RM} policy and the Overbooking without joint cloud federation stay

TABLE 5. Cloud provider1's profits and sla violation with and without a joint cloud federation.

	Cloud Provider 1's profits ($/year)			Cloud rejection probability		
	(1) Joint cloud federation	(2) Without joint cloud federation	(3)=((1)-(2))/(1)) Profit growth (%)	(1) Joint cloud federation	(2) Without joint cloud federation	(3)=(2)-(1) The decrease of SLA violation (%)
Cloud provider 1	890,326.19	884,097.00	0.705%	31.62	40	8.38

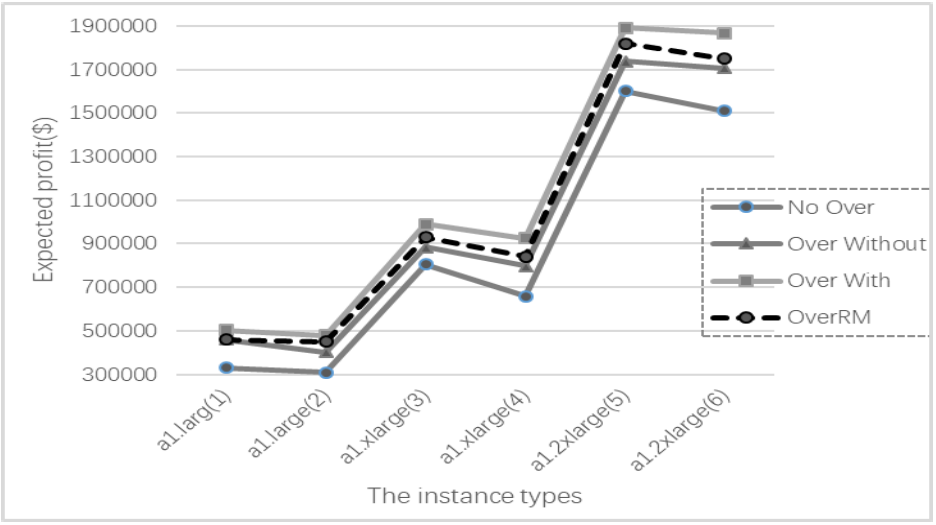


FIGURE 2. Impact of instance types and unit prices.

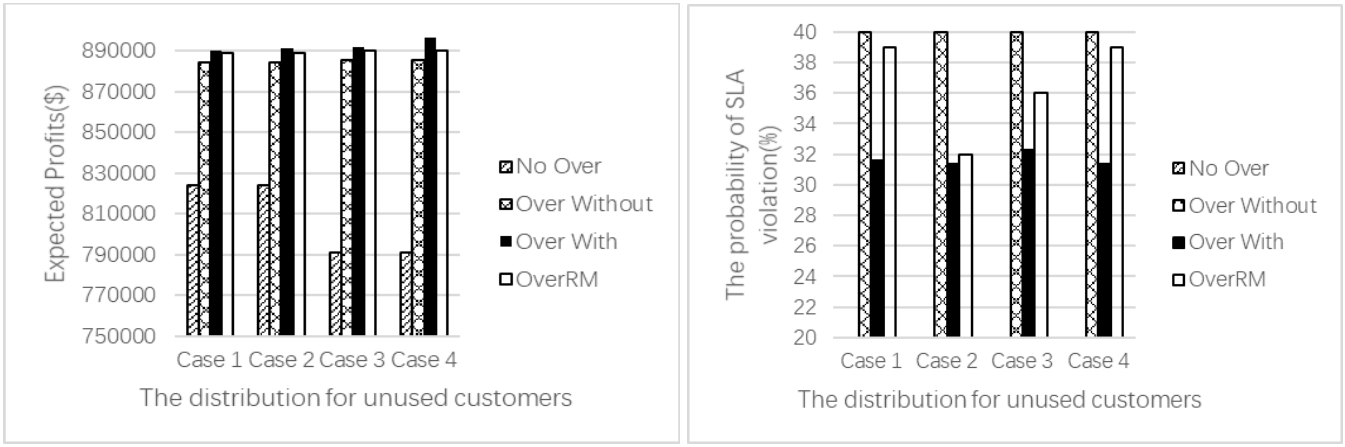
in the middle of both.This is because the increase of profit caused by the cooperative game decrease the risk and penalty cost of the SLA violation. Comparing with different instance performance, such as a1.large and a1.xlarge, under the same users 'demand, the greater unit instance price, the more profit the cloud providers will receive. In addition, the upfront reserved fee do not significantly affect the profit. This is because, the greater upfront reservation fee will decrease unit price of reserved instance, thus decrease the profit. However, the more upfront reservation fee, the lower of the probability of reserved but unused.

C. CHANGES IN THE DISTRIBUTIONS OF RESERVED CUSTOMERS

In Section VI-A, we assume that the overbooking quantity and the profits are optimal when the probability of reserved but unused customers is the normal distribution, and. We define this scenario as Case 1. To conduct a fair comparison, when the probability of reserved but unused customers follows a different normal distribution, we derive the optimal profit and the probability of SLA violation of cloud Provider 1under different overbooking strategies (No Over, Over Without, Over With, Over^{RM}).The optimal overbooking qualities are achieved under the biggest cloud federation {1,2,3,4}.Four cases are examined in Fig. 3. The other three distributions $= [\mu = 0.1r_o, \sigma^2 = 1/2\mu; \mu = 0.15r_o, \sigma^2 = 1/3\mu; \mu = 0.15r_o, \sigma^2 = 1/2\mu]$.

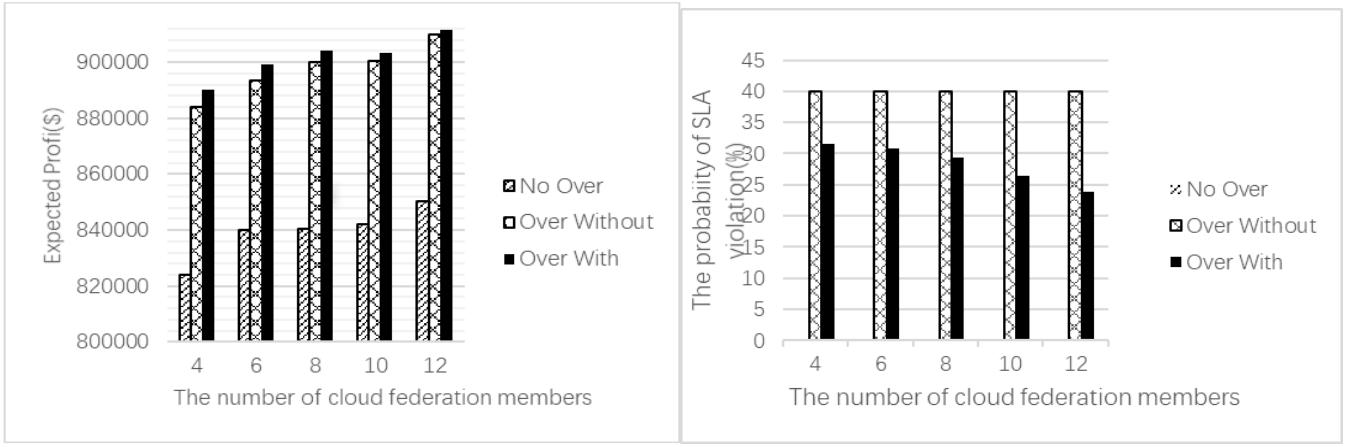
As presented in Fig 3(a), regardless of the different features of the normal distribution in the four cases, the overbooking with joint cloud federation strategy reveals positive improvements comparing with the other strategies. Given the same average reserved but unused customer numbers, the greater the uncertainty of the reserved but unused customers, the greater the improvements in profits However, when the probability of reserved but unused customers remains stable, the profit increases as the number of reserved but unused customers increases.

As for the probability of SLA violation, we can find in Fig.3 (b) that, the overbooking with joint cloud federation strategy decrease the probability of SLA violation under a higher profit when comparing with Over^{RM} and the overbooking without joint cloud federation strategy. Due to the associated error to the predict component, Over^{RM} policy brings about higher probability of SLA violation. When the probability of reserved but unused customers remains stable, the probability of SLA violation remains stable. Under the No overbooking strategy, the reserved but unused instance will lead to idle resource which makes the probability of SLA violation equals zero.



(a) Expected Profit of Cloud provider 1 (b) The probability of SLA violation

FIGURE 3. Impact of the distribution features for unused customers.



(a) Expected Profits of Cloud provider 1 (b) The probability of SLA violation

FIGURE 4. Impact of the number of cloud federation members.

D. IMPACT OF CLOUD FEDERATION SIZE UNDER DIFFERENT OVERBOOKING STRATEGIES

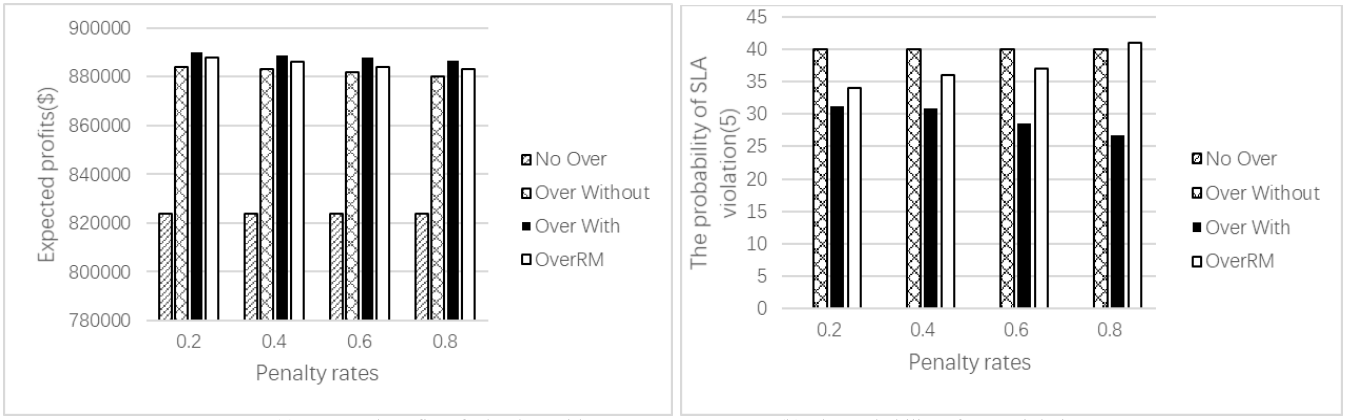
In the beginning of the formation of a cloud federation, the number of members is small, as evidenced by the current situation of Amazon and Alibaba, who aim to establish large cloud ecosystems. Fig.4 presents the impact of federation size on Provider 1's, e.g., Amazon, expected profits and SLA violation under different overbooking strategies.

We find in Fig 4 (a) that the Provider 1's profit increase as the federation size increase under three overbooking strategies. In addition, the increase of the profit's improvement under the No-overbooking strategy is less than Overbooking with joint cloud federation strategy. Due to the super-additive of the cooperative game model, the profit of Provider 1 under the biggest federation that all cloud member's joint is more than the sub-federation. It means that the biggest federation is the optimal federation. Then, Provider 1's profit improvements decrease as the federation size increases. The larger the federation size, the greater the surplus resources. When the federation's surplus resources grow faster than that of Provider 1's overbooked quantity, the unit exchange price decreases due to the increase in idle resources in the federation.

However, as shown in Fig 4 (b), Provider 1's SLA violation decrease as the size of the federation increases because more resources are available to serve Provider 1's customers. Comparing with the Over With policy, the SLA violation probability is stable and equal to 40% under the Over Without policy. Due to there is no resource exchange to reduce risk, if cloud providers do not join the cloud federation.

E. CHANGES IN THE PENALTY COSTS

Under cloud market, not providing sufficient capacity for reserved customers is a violation of SLA, so the cloud provider must pay a penalty to the users. To examine whether the optimal overbooking with joint cloud federation strategy remains valid under different penalty rates, we vary the penalty rate $\xi = [0.2, 0.4, 0.6, 0.8]$. Fig.5 (a) shows that the expected profit would decrease due to the increase in



(a) Expected Profits of Cloud provider 1

(b) The probability of SLA violation

FIGURE 5. **Impact of the penalty rates.**

TABLE 6. **6 ∗ 4 ∗ 4 ∗ 5 = 480 sensitivity analysis and results.**

	Level for sensitivity factors				Profit improvement（$）			The probability of SLA violation（%）		
ID	Prices of on-demand and reserved instance	Distribution of unused customers	Penalty rates(ξ)	Numbers of federation members	No overbooking	Overbooking Without joint cloud federation	Overbooking with Joint cloud federation	No overbooking	Overbooking Without joint cloud federation	Overbooking with Joint cloud federation
1	a1.large(α=0,β=0.68)	μ=0.1r_0 $\sigma^2=1/3\mu$	0.2	4	411919	439048	443163	0	40%	31.4
2	a1.large(α=154USD,β=0.6)	$u=0.1r_o$ $\sigma^2=1/3\mu$	0.2	4	399561	425876	429868	0	40%	31.4
…	……	……	……	……	……	……	……	……	……	……
96	a1.xlarge(α=0,β=0.68)	$u=0.1r_o$ $\sigma^2=1/2\mu$	0.4	6	823838	882982	888924	0	40%	30.81
…	……	……	……	……	……	……	……	……	……	……
240	a1.xlarge(α=308USD,β=0.6)	$u=0.15r_o$ $\sigma^2=1/3\mu$	0.6	8	800219	867521	872129	0	40%	32.3
…	……	……	……	……	……	……	……	……	……	……
480	a1.2xlarge(α=617USD,β=0.6)	$u=0.15r_o$ $\sigma^2=1/2\mu$	0.8	12	1425892	1630211	164872	0	40%	24.1

penalty cost under the overbooking strategies, while it remain stable under the No overbooking strategy. When the unit penalty cost g decreases, both optimal overbooking quantity tend to increase because joint cloud federation to exchange resource becomes a more attractive choice to deal with the unserved users. However, the risk of SLA violation will be decrease with the increase of the penalty rate, which is shown in Fig. 5 (b).In addition, the probability of SLA violation under the Over^{RM} policy is higher than the Over With joint cloud federation because the exists different clients classification imply higher penalties.To understand the overbooking strategy's overall quality in withstanding uncertainties in diverse environments, we next conduct a multifactor sensitivity analysis.

F. MULTIFACTOR ROBUSTNESS STUDY

The optimal overbooking with joint cloud federation strategy is deemed robust if it can cope with significant uncertainty in different scenarios. Using a 6∗4∗4∗5 sensitivity analysis, we conduct 480 experiments for each overbooking strategy in Table 6 to test 4 factors: (i) instance types and unit prices, (ii) the distribution of reserved customers, (iii) the cloud federation size, (iv) the penalty cost. Each experiment corresponds to a unique combination of these factors. The expected

profits and probability of SLA violation under different overbooking strategies are shown in the six rightmost columns. The sensitivity analysis suggest that the overbooking model is robust under various scenarios. As far as the expected profit and SLA violation is concerned, we find that when facing higher average reserved but unused customer numbers, greater federation size, and higher upfront reserved fee, the overbooking strategies performs significantly better, whereas the penalty rates is inconsequential.

VII. CONCLUSION

In this report, we present an optimal overbooking policy for cloud providers to enhance their profits and resource utilization. We first develop a market-oriented cloud federation trading structure (framework). Based on the proposed structure, we determine a cloud federation exchange price. Unlike the previous literature on fixed exchange prices in other industries, we propose the dynamic exchange price, which changes based on the resource utilization of the cloud federation and the operational costs of VMs. Based on the dynamic exchange price, we establish the optimal overbooking strategy for a cloud federation environment, derive the optimal overbooking quantity and compare this strategy with that of a no federation environment. Through establishing the cooperative game model of the cloud federation, we prove the super-additive and nonempty cores of the model and make a reasonable profit distribution based Shapley value. The results indicate that the expected profits can be improved and the probability of SLA violation can be declined under a federation.

Using the price data available from Amazon EC2 and the cloud ecosystem that it aims to establish, we determine that joining a cloud federation is beneficial for cloud providers, as it improves their overbooking performance with respect to both profits and customer service. We further observe that under diverse instance types and unit prices, different normal distributions of the probability of reserved but unused users, and different penalty rates, the profits can be improved and the probability of SLA violation can be decreased under the overbooking with joint cloud federation strategy comparing to the over without joint cloud federation policy and No-overbooking strategy. However, given the penalty costs, the greater the penalty rate is, the lower the SLA violation and the expected profit is. In addition, we examine the impact of the size of the cloud federation on the overbooking policy when considering the development of the cloud federation and find that the growth in providers' profits decreases, whereas decrease in the SLA violation, as the size of the federation increases.

There are limitations associated with our study. First, to focus on a single cloud provider's overbooking policy, we assume all the cloud federation's members are the same. Therefore, taking into account the competencies of different members in a cloud federation is a direction for future research. Second, we set a fixed instance unit price for one time.In fact, cloud providers offers diverse instance types prices at the same time to meet customers' demand. Future research should address such issues. It is also of interest to examine how to overbook resources to meet the heterogeneous consumers' demands through a cloud federation.

REFERENCES

[1] K. J. Preacher and A. F. Hayes, "Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models," *Behav. Res. Methods*, vol. 40, no. 3, pp. 879–891, Aug. 2008.
[2] R. N. Calheiros, A. N. Toosi, C. Vecchiola, and R. Buyya, "A coordinator for scaling elastic applications across multiple clouds," *Future Gener. Comput. Syst.*, vol. 28, no. 8, pp. 1350–1362, Oct. 2012.
[3] M. Al-Roomi, S. Al-Ebrahim, S. Buqrais, and I. Ahmad, "Cloud computing pricing models: A survey," *Int. J. Grid Distrib. Comput.*, vol. 6, no. 5, pp. 93–106, Oct. 2013.
[4] Y. Huang, Y. Ge, X. Zhang, and Y. Xu, "Overbooking for parallel flights with transference," *Int. J. Prod. Econ.*, vol. 144, no. 2, pp. 582–589, Aug. 2013.
[5] A. N. Toosi, K. Vanmechelen, K. Ramamohanarao, and R. Buyya, "Revenue maximization with optimal capacity control in infrastructure as a service cloud markets," *IEEE Trans. Cloud Comput.*, vol. 3, no. 3, pp. 261–274, Sep. 2015.
[6] Í. Goiri, J. Guitart, and J. Torres, "Economic model of a cloud provider operating in a federated cloud," *Inf. Syst. Frontiers*, vol. 14, no. 4, pp. 827–843, Sep. 2012.
[7] X. Chen and G. Hao, "Co-opetition alliance models of parallel flights for determining optimal overbooking policies," *Math. Comput. Model.*, vol. 57, nos. 5–6, pp. 1101–1111, Mar. 2013.
[8] M. Grozev and R. Buyya, "Inter-Cloud architectures and application brokering: Taxonomy and survey," *Softw., Pract. Exper.*, vol. 44, no. 3, pp. 369–390, Mar. 2014.
[9] D. Villegas, N. Bobroff, I. Rodero, J. Delgado, Y. Liu, A. Devarakonda, L. Fong, S. M. Sadjadi, and M. Parashar, "Cloud federation in a layered service model," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1330–1344, 2012.
[10] X. Yang, B. Nasser, M. Surridge, and S. Middleton, "A business-oriented cloud federation model for real-time applications," *Future Generat. Comput. Syst.*, vol. 28, no. 8, pp. 1158–1167, 2012.
[11] M. Giacobbe, A. Celesti, M. Fazio, M. Villari, and A. Puliafito, "Towards energy management in cloud federation: A survey in the perspective of future sustainable and cost-saving strategies," *Comput. Netw.*, vol. 91, no. 14, pp. 438–452, Nov. 2015.
[12] A. Celesti, M. Fazio, M. Villari, and A. Puliafito, "Virtual machine provisioning through satellite communications in federated Cloud environments," *Future Gener. Comput. Syst.*, vol. 28, no. 1, pp. 85–93, Jan. 2012.
[13] M. M. Hassan, M. S. Hossain, A. M. J. Sarkar, and E.-N. Huh, "Cooperative game-based distributed resource allocation in horizontal dynamic cloud federation platform," *Inf. Syst. Frontiers*, vol. 16, no. 4, pp. 523–542, Sep. 2014.
[14] B. Li, Z. Zhao, Y. Guan, N. Ai, X. Dong, and B. Wu, "Task placement across multiple public clouds with deadline constraints for smart factory," *IEEE Access*, vol. 6, pp. 1560–1564, 2018. doi: 10.1109/ACCESS.2017.2779462.
[15] T. Püschel, G. Schryen, D. Hristova, "Revenue management for Cloud computing providers: Decision models for service admission control under non-probabilistic uncertainty," *Eur. J. Oper. Res.*, vol. 244, no. 2, pp. 637–647, Jul. 2015.
[16] S.-H. Chun and B.-S. Choi, "Service models and pricing schemes for cloud computing," *Cluster Comput.*, vol. 17, no. 2, pp. 529–535, Jun. 2014.
[17] B. Javadi, R. K. Thulasiram, and R. Buyya, "Characterizing spot price dynamics in public cloud environments," *Future Gener. Comput. Syst.*, vol. 29, no. 4, pp. 988–999, Jun. 2013.
[18] M. Macías, J. Guitart, "SLA negotiation and enforcement policies for revenue maximization and client classification in cloud providers," *Future Gener. Comput. Syst.*, vol. 41, pp. 19–31, Dec. 2014.
[19] R. Householder, S. Arnold, and R. Green, "Simulating the effects of cloud-based oversubscription on datacenter revenues and performance in single and multi-class service levels," in *Proc. IEEE 7th Int. Conf. Cloud Comput.* Anchorage, AK, USA, Jun./Jul. 2014, pp. 562–569. doi: 10.1109/CLOUD.2014.81.

[20] H. Wu, S. Ren, G. Garzoglio, S. Timm, G. Bernabeu, K. Chadwick, and S.-Y. Noh, "A reference model for virtual machine launching overhead," *IEEE Trans. Cloud Comput.*, vol. 4, no. 3, pp. 250–264, Jul./Sep. 2016. doi: 10.1109/TCC.2014.2369439.

[21] D. Breitgand and A. Epstein, "Improving consolidation of virtual machines with risk-aware bandwidth oversubscription in compute clouds," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 2861–2865.

[22] J. Heo, X. Zhu, P. Padala, and Z. Wang, "Memory overbooking and dynamic control of Xen virtual machines in consolidated environments," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage.*, Long Island, NY, USA, Jun. 2009, pp. 630–637. doi: 10.1109/INM.2009.5188871.

[23] T. Wo, Q. Sun, B. Li, and C. Hu, "Overbooking-based resource allocation in virtualized data center," in *Proc. IEEE 15th Int. Symp. Object/Compon./Service-Oriented Real-Time Distrib. Comput. Workshops*, Apr. 2012, pp. 142–149.

[24] L. Tomás and J. Tordsson, "An autonomic approach to risk-aware data center overbooking," *IEEE Trans. Cloud Comput.*, vol. 2, no. 3, pp. 292–305, 2014.

[25] S. H. Ivanov, "Optimal quality of service overbooking limits for a hotel with three room types and with upgrade and downgrade constraints," *Tourism Econ.*, vol. 21, no. 1, pp. 223–240, Feb. 2015.

[26] B. B. Oliveira, M. A. Carravilla, and J. F. Oliveira, "Fleet and revenue management in car rental companies: A literature review and an integrated conceptual framework," *Omega*, vol. 71, pp. 11–26, Sep. 2016.

[27] F. Guerriero, G. Miglionico, F. Olivito, "Strategic and operational decisions in restaurant revenue management," *Eur. J. Oper. Res.*, vol. 237, no. 3, pp. 1119–1132, Sep. 2014.

[28] D. G. Sierag, G. M. Koole, R. D. van der Mei, J. I. van der Rest, and B. Zwart, "Revenue management under customer choice behaviour with cancellations and overbooking," *Eur. J. Oper. Res.*, vol. 246, no. 1, pp. 170–185, Oct. 2015.

[29] F. Lopez-pires, B. Baran, L. Benítez, S. Zalimben, and A. Amarilla, "Virtual machine placement for elastic infrastructures in overbooked cloud computing datacenters under uncertainty," *Future Gener. Comput. Syst.*, vol. 79, no. 3, pp. 830–848, Sep. 2017.

[30] N. Liu and S. Ziya, "Panel size and overbooking decisions for appointment-based services under patient no-shows," *Prod. Oper. Manage.*, vol. 23, no. 12, pp. 2209–2223, 2014.

[31] M. Armbrust, M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Int. J. Comput. Technol.*, vol. 4, no. 2b1, pp. 50–58, 2013.

[32] Z. Ye, S. Mistry, A. Bouguettaya, and H. Dong, "Long-term QoS-aware cloud service composition using multivariate time series analysis," *IEEE Trans. Services Comput.*, vol. 9, no. 3, pp. 382–393, May/Jun. 2016.

[33] Y. Wei, L. Pan, S. Liu, L. Wu, and X. Meng, "DRL-scheduling: An intelligent QoS-aware job scheduling framework for applications in clouds," *IEEE Access*, vol. 6, pp. 55112–55125, 2018. doi: 10.1109/ACCESS.2018.2872674.

[34] G. Baranwal and D. P. Vidyarthi, "Admission control in cloud computing using game theory," *J. Supercomput.*, vol. 72, no. 1, pp. 317–346, Jan. 2016.

[35] Y. Jia, Z. Mi, Y. Yu, Z. Song, and C. Sun, "A bilevel model for optimal bidding and offering of flexible load aggregator in day-ahead energy and reserve markets," *IEEE Access*, vol. 6, pp. 67799–67808, 2018. doi: 10.1109/ACCESS.2018.28790.

[36] M. Aloqaily, B. Kantarci, and H. T. Mouftah, "Multiagent/multiobjective interaction game system for service provisioning in vehicular cloud," *IEEE Access*, vol. 4, pp. 3153–3168, 2016. doi: 10.1109/ACCESS.2016.2575038.

MENGDI YAO received the M.E. degree from the Wuhan University of Technology, in 2015, where she is currently pursuing the Ph.D. degree in electricity ecommerce. Her research interests include cloud computing and pricing.

DONGLIN CHEN received the M.E. degree from the Wuhan University of Science and Technology, in 1995, and the Ph.D. degree in management science and engineering from the Huazhong University of Science and Technology, in 2003. He is currently a Professor of economics with the Wuhan University of Technology. His research interests include cloud computing and sharing economy.

JENNIFER SHANG has published more than 40 research articles in various professional journals. Her research interests include manufacturing and service operations management, revenue management, quantitative methods, operations research, and information systems.