Advance Reservation Games

ERAN SIMHON and DAVID STAROBINSKI, Boston University

Advance reservation (AR) services form a pillar of several branches of the economy, including transportation, lodging, dining, and more recently, cloud computing. In this work, we use game theory to analyze a slotted AR system in which customers differ in their lead times. For each given time slot, the number of customers requesting service is a random variable following a general probability distribution. Based on statistical information, the customers decide whether or not making an advance reservation of server resources in future slots for a fee. We prove that only two types of equilibria are possible: either none of the customers makes AR or only customers with lead time greater than some threshold make AR. Our analysis further shows that the fee that maximizes the provider's profit may lead to other equilibria, one of which yielding zero profit. In order to prevent ending up with no profit, the provider can elect to advertise a lower fee yielding a guaranteed, but smaller profit. We refer to the ratio of the maximum possible profit to the maximum guaranteed profit as the price of conservatism. When the number of customers is a Poisson random variable, we prove that the price of conservatism is one in the single server case, but can be arbitrarily high in a many-server system.

Categories and Subject Descriptors: []

1. INTRODUCTION

Advance reservation services play a key role in several branches of the economy, such as transportation, lodging, dining, and health care. There has also been a growing interest in applying AR in cloud computing systems [Sotomayor 2009]. For instance, Moab Workload Manager[1] and IBM Platform Computing Solutions[2] support AR. In both of these packages, an administrator can decide whether or not to enable AR and define an AR pricing scheme. In most systems supporting AR, customers can choose whether making AR or not. Since the payoff of each customer is affected by decisions of other customers, it is natural to analyze the behavior of such systems as strategic games.

In this paper, we study a set of strategic non-cooperative games, referred to as *advance reservation games*, where players (customers) can reserve future resources in advance for a fee. Charging a reservation fee is a common practice in different venues and is also used in cloud computing. For example, Amazon EC2 cloud offers *reserved instance* service, in which customers pay a fee that allow them to use resources later on for a lower cost.

Typically, advance reservations are offered when customers are sensitive to the service starting point and will not agree to wait if service is not available when needed. Accordingly, we consider a loss system where customers leave the system if they cannot get service at their desired time slots. In cloud computing, this typically occurs when parallel computing is needed [Sotomayor 2009]. We assume that the service time is slotted which is a common assumption in the literature of AR [Charbonneau and Vokkarane 2012] due to the complexity of analyzing a continuous time queue that supports AR.

In loss systems, customers have a clear incentive of making AR to increase their chances to get service. Likewise, the provider is also motivated to offer AR, since she can increase her overall profit by charging an appropriate reservation fee. Upon deciding whether making AR or not, customers face two uncertainties. The first uncertainty lies in the number of customers competing for the same set of

[1]See http://docs.adaptivecomputing.com/mwm/7-0/mwmAdminGuide-7.0.pdf.
[2]See http://www.redbooks.ibm.com/redbooks/pdfs/sg248073.pdf.

resources in a given slot. We refer to this number as the *demand*, which is a random variable following a general probability distribution.

The second uncertainty lies in the time at which other customers consider making AR. In the models introduced in this paper, customers differ by their lead times, where the lead time of a customer is the time elapsing between the point when he realizes that he will need service at a given time slot and the starting time of that slot. The lead times of customers are continuous i.i.d. random variables that follow a general distribution, but remain private (i.e., customers know the statistics but do not know the realizations of the demand and the lead times of other customers).

To illustrate the model, consider the following example: a system with several servers has a slot duration that lasts for one day, starting at 12:00 AM. A customer realizes on Monday 6:00 PM that he will need service on Wednesday. Thus, his lead time is 30 hours. Upon realizing that service will be needed on Wednesday, the customer can either reserve that slot in advance or avoid AR. The customer does not know how many other customers wish to be served on that day. However, statistical information is available to him.

The model assumes a monopolistic setting. Thus, the provider can choose any AR price and AR mechanism that maximizes her expected profit. The profit has two sources: AR fees and service fees. Since customers only decide whether making AR or not, their decisions have no impact on the number of customers getting service in a given slot (their decisions only impact *who* get service). Hence, the provider can ignore the profit obtained from service when choosing the AR price and AR mechanism. For brevity, henceforth, profit means AR profit. We note that reservation fees are a major source of revenue in several booking systems, including cloud computing [Aazam and Huh 2015].

In this paper, we evaluate different types of AR games and derive their Nash equilibria (we only consider symmetric equilibria, a common assumption made in the literature of queueing games [Hassin and Haviv 2003]. For the games under consideration, we prove that only two types of equilibria are possible: either none of the customers makes AR or only customers with lead time greater than some threshold make AR.[3] Furthermore, we show that, at equilibrium, informing customers that free servers are available does not impact the provider's expected profit. However, charging an AR fee from all customers attempting AR (i.e., not only those granted service) can only decrease the expected profit.

Once a mechanism is chosen, another question arises: what is the AR fee that maximizes the provider's expected profit? The answer to this question turns out to be more complicated. We show that there exists a range of fees, such that choosing a fee within this range leads to multiple equilibria with one of them yielding zero profit. Therefore, in order to properly set the AR fee, the provider should consider both the fee yielding the maximum *possible* profit and the fee yielding the maximum *guaranteed* profit. For this purpose, we introduce the concept of *price of conservatism* (PoC), which corresponds to the ratio of the maximum possible profit to the maximum guaranteed profit. We assume that the demand follows a Poisson distribution and derive the price of conservatism in different settings. First, we analyze the case of a single server, where we prove that $PoC = 1$ (i.e., no loss). Next, we conduct the analysis of a many-server system and prove that the price of conservatism can be arbitrarily high. This situation occurs when the system is slightly overloaded.

We note that since AR games are zero-sum games, the social welfare (i.e., the total payoff of all players in the game, including the provider) is not affected by the decisions of the provider and customers. Therefore, the price of anarchy [Koutsoupias and Papadimitriou 1999] in such games always equals one. In contrast, the price of conservatism measures the loss of profit from the viewpoint of the provider.

[3]Although the former equilibrium, in which none of the customers make AR, could be viewed as a special case of the latter equilibrium, distinguishing between the two equilibria is needed for the game analysis.

The rest of this paper is organized as follows. In the next section, we cover related work. In Section 3, we describe the different models. In Section 4, we analyze each model and find the equilibria structure. In Section 5, we compare between the expected profits of the different models. In section 6, we define the Price of Conservatism and compute it for different system sizes. Finally, in Section 7, we conclude and suggest directions for future research.

2. RELATED WORK

Queueing systems and communication networks that support advance reservations have extensively been researched for the past two decades. Most of the research focuses on performance evaluation and algorithmic aspects of AR systems. For example, [Smith et al. 2000] propose a scheduling model that supports AR and evaluate several performance metrics. [Kaushik et al. 2006] suggest an AR model with flexible time window and show that this model has a lower blocking probability and a higher utilization than a model without window. [Guérin and Orda 2000] analyze the effect of AR on the complexity of path selection. [Virtamo 1992] evaluates the impact of advance reservation on server utilization. [Buyya et al. 2009] report a simulation-based comparison between different payment mechanisms. [Cohen et al. 2009] propose algorithms for network routing that support advance channel reservations. For a survey on the field, see [Charbonneau and Vokkarane 2012].

Research on advance reservation can be also found in the literature on revenue management. For example, [Liberman and Yechiali 1978] analyze a hotel reservation system where overbooking is allowed and the goal is to find the optimal overbooking level. [Reiman and Wang 2008] propose an admission control strategy for reservation system with different classes of customers. [Bertsimas and Shioda 2003] propose a policy for accepting/rejecting restaurant reservations.

The application of game theory to analyze customers' behavior in queues (shortly, *queueing games*) is pioneered by [Naor 1969]. In that paper, the author considers an $M/M/1$ queue where customers observe the queue length and then decide whether to join or bulk. Follow-up work analyzed the behavior of customers in other queueing models. [Edelson and Hilderbrand 1975] analyze an unobservable $M/M/1$ queue, where customers decide whether to join or bulk without knowing the queue state. [Altman and Shimkin 1998] analyze an observable processor sharing system, where customers decide whether or not to join after observing the number of users in the system. [Balachandran 1972] analyzes an observable $M/M/1$ queue with priorities, where customers decide on a payment and accordingly priorities are assigned. [Haviv et al. 2010] analyze an unobservable $M/M/N/N$ system that is initially empty and customers decide whether to join or bulk based on their arrival time. [Jain et al. 2011] introduce concert queueing games, where customers, interested in early service with minimal wait, choose their arriving time into a system with a specific opening time. [Haviv and Roughgarden 2007] analyze an unobservable system with non-identical servers, where customers wish to minimize their waiting time and select a server accordingly.

Several models of queueing games focus on revenue management. In those games, a provider designs the system such that her revenue will be maximized. For example, [Masuda and Whang 2006] assume that the pricing scheme is controlled by the provider and the goal is to find a pricing scheme that maximizes the provider's profit. Such types of games are a special case of Stackelberg games. In a Stackelberg game, a leader (a provider, for instance) acts first, and followers (customers, for instance) respond to the leader's move. In our work, we follow a similar approach, where the objective of the provider is to maximize her revenue from AR fees. For a review on the field of queueing games, see [Hassin and Haviv 2003].

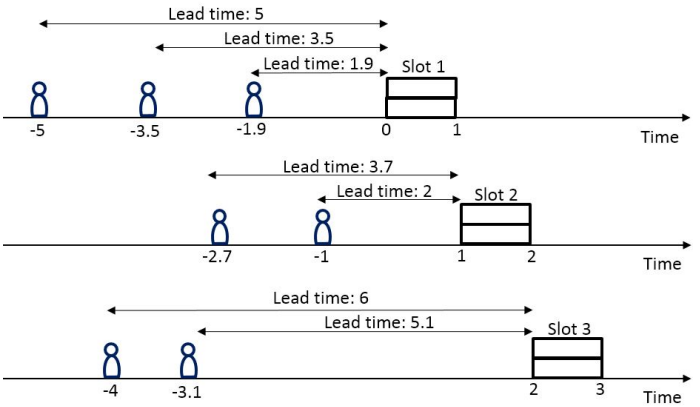

Fig. 1. Slotted system with two servers. The demand for each slot is independent. The demand for slots 1, 2 and 3, is respectively 3, 2 and 2.

3. THE MODELS

First we describe the assumptions that are common to all three models:

(1) The system consists of N servers.

(2) The service time axis is slotted. That is, in each slot, customers are served from the beginning till the end of the slot.

(3) The *demand*, which represents the number of customers that request service in a specific slot (each customer requests one server) is a random variable. This formulation is common for slotted queues (e.g., [Kang and Tan 1993] and [Blanchet et al. 2009]). The demand for slot i is denoted by D_i.

(4) Customers that do not get service in a given slot do not make another trial (a common assumption in the literature of loss systems [Ross 1995]). Thus, the demand in each slot is independent of the history and follows a general distribution supported in $\mathbb{N}$.

(5) The customers of each slot differ by the time elapsing between considering making AR (i.e, realizing that service will be needed in that future slot) and the slot starting time. We refer to this time interval as the *lead time* of a customer. The lead times of all customers are independent and identically distributed random variables, supported in $\mathbb{R}^+$, with cumulative distribution function denoted $F(\cdot)$.

(6) Each customer chooses one of two actions: make AR or not make AR, denoted AR and AR' respectively.

(7) If the demand for a slot is larger than N, the servers are allocated to the first N customers that made AR. If fewer than N customers made AR, the remaining servers are arbitrarily allocated between the customers that did not make AR.

(8) The customers and the provider know the number of servers N and statistical information on the system (i.e., the distribution of the demand and the lead times).

(9) The provider charges a fixed reservation fee denoted C. All the customers have the same utility U from the service. Without loss of generality, we set $U = 1$.

Figure 1 illustrates the model. The models analyzed in this paper differ in their reservation mechanisms as follows:

Table I. Payoff summary

-	Make AR		Not make AR	
Model	Served	Not served	Served	Not served
1 and 3	$1-C$	0	1	0
2	$1-C$	$-C$	1	0

(1) **Unobservable model 1**: customers have no information regarding the availability of servers at the time of reservation. If a customer makes an AR request, he is then informed whether a server will be allocated at the requested slot or not. In the first case, a reservation fee is charged. In the second case, the customer leaves the system with no gain or cost.
(2) **Unobservable model 2**: as in Model 1, customers have no information regarding the availability of servers at the time of reservation. In this model, however, a reservation fee is charged from each customer that makes an AR request.
(3) **Observable model**: customers are informed, prior to their decision, if a server is available at their requested time slot. A customer that has been informed that there is no free server leaves the system.

We note that, in the observable model, customers leave the system if they see that no server is available. Thus, a situation where a customer pays the AR fee but does not get service does not exist in that model.

The possible payoffs of the three models are summarized in Table I.

4. EQUILIBRIA ANALYSIS

4.1 Classification of the equilibria

We analyze the three models as non-cooperative games where each player (customer) aims to maximize his payoff. Since the demand for each slot is an i.i.d random variable, the analysis of a single slot is sufficient for analyzing the game. Since we only consider one slot, we simply denote the demand by D.

Any fee greater or equal to one has a trivial result where none of the customers makes AR. No fee or a negative fee have the trivial result of all customers making AR. Hence, in our analysis, we consider only fees between zero and one (i.e., $0 < C < 1$).

We note that the demand seen by a customer may be different from that seen by an external observer (the provider, for instance). Indeed, the fact that a customer seeks service in a given time slot affects his estimation of the number of other customers seeking service in that time slot. On the one hand, a customer is more likely to fall in a slot with large demand than in one with small demand. On the other hand, he must exclude himself. This phenomenon is known as the discrete case of the *waiting time paradox* (or *residual life paradox*). We define $\tilde{D}$ as the number of customers seen by a customer beside himself. The probability distribution function (PDF) of $\tilde{D}$ is known to be [Avineri 2004]:

$$\mathbb{P}\left(\tilde{D}=j\right)=\mathbb{P}\left(D=j+1\right)\frac{(j+1)}{\mathbb{E}[D]}. \tag{1}$$

The following lemma states that each customer makes his decision upon realizing that service is required.

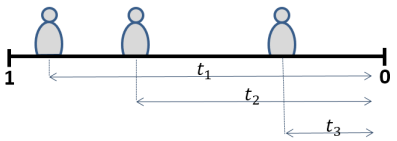

Fig. 2. Example of a realization of the demand in a given slot. The service starts at (normalized) lead time 0. As users have greater (normalized) lead times, they have the opportunity to reserve a server earlier.

LEMMA 4.1. *For all customers, making AR when realizing that service is required yields at least the same payoff as making AR later on.*

PROOF. In all three models, if a customer makes AR when all servers are already reserved, his payoff will be the same as if making AR later on. If a customer makes AR when there is at least one free server, his payoff will be $1-C$. If he makes a reservation at a later point, his payoff will be the same if there is still at least one free server. If there is no more server available, his payoff will be zero in the first and third models and negative in the second model. □

Given a lead time γ, we set $t=F(\gamma)$ and refer to it as the *normalized lead time*. Due to the probability integral transformation theorem [Dodge 2006, p. 320], we know that t is a random variable, uniformly distributed in $[0,1]$. Note that $F(\gamma)$ is also the average fraction of customers with lead time smaller than γ. Fig. 2 illustrates the notion of the normalized lead time.

For each game, we define a strategy function $\sigma\colon t\to[0,1]$, which represents the probability that a tagged customer with normalized lead time $t\in[0,1]$ makes AR. Since we consider only symmetric equilibria, all customers follow the same strategy function. Through conditioning, given that there are k other customers with normalized lead times $t_1,\ldots,t_k$ that follow strategy σ, the tagged customer can find his probability of getting service (we denote that event by S) for each action he chooses. His probability to get service, when choosing action $\alpha\in\{AR,AR'\}$ is

$$\mathbb{P}(S|t,\alpha,\sigma)=\mathbb{P}\left(\tilde{D}<N\right)+\sum_{k=N}^{\infty}\mathbb{P}\left(\tilde{D}=k\right)\int_{t_1=0}^{1}\cdots\int_{t_k=0}^{1}\mathbb{P}(S|t,\alpha,k,t_1,\ldots,t_k,\sigma)\,\mathrm{d}t_1\cdots\mathrm{d}t_k. \tag{2}$$

The first term in (2) is the probability that the number of customers (beside the tagged customer) is smaller than N. In this case, all customers get service, regardless of their decisions. The second term is the weighted sum of the probabilities of getting service when the number of customers (beside the tagged customer) is at least N. In this case, the probability that the tagged customer gets service depends on his action and on the strategy followed by the other customers and their lead times (note that the PDF of the random variable t_j equals 1, for each $j\in\{1,..,k\}$). As shown in the sequel, deriving an explicit expression for $\mathbb{P}(S|t,\alpha,k,t_1,\ldots,t_k,\sigma)$ is not required for the equilibria analysis.

Given the model and strategy function followed by all other customers, one can express the expected payoff, denoted $U_\sigma(t,\alpha)$, for each action α by multiplying $\mathbb{P}(S|t_i,\alpha,\sigma_{-i})$ and $1-\mathbb{P}(S|t_i,\alpha,\sigma_{-i})$ with the relevant payoffs, as summarized in Table I. For example, for the second model:

$$U_\sigma(t,AR)=\mathbb{P}(S|t,AR,\sigma)(1-C)+(1-\mathbb{P}(S|t,AR,\sigma))(-C) \tag{3}$$

and

$$U_\sigma\left(t,AR'\right)=\mathbb{P}\left(S|t,AR',\sigma\right)\cdot 1+\left(1-\mathbb{P}\left(S|t,AR',\sigma\right)\right)\cdot 0. \tag{4}$$

At equilibrium, each customer chooses an action that maximizes his expected payoff. Thus, we define an equilibrium strategy (i.e., a strategy that leads to equilibrium) as follows.

Definition 4.2. Strategy σ is an equilibrium strategy if the following holds for any normalized lead time $t\in[0,1]$:

(1) If $\sigma(t) = 0$ then $U_\sigma(t, AR) \leq U_\sigma(t, AR')$.
(2) If $0 < \sigma(t) < 1$ then $U_\sigma(t, AR) = U_\sigma(t, AR')$.
(3) If $\sigma(t) = 1$ then $U_\sigma(t, AR) \geq U_\sigma(t, AR')$.

That is, at equilibrium, a customer chooses the action AR', only if he is (weakly) better off not making AR; he randomizes his action, only if he is indifferent between the two outcomes; and he chooses the action AR, only if he is (weakly) better off making AR.

Next we show that at equilibrium all customers follow the same *threshold strategy*, defined below.

Definition 4.3. A threshold strategy has the following structure:

$$\sigma(t) = \begin{cases} 1 & \text{if } t > \tau \\ 0 & \text{if } t \leq \tau. \end{cases}$$

where τ is a threshold value in the interval $(0, 1]$.

LEMMA 4.4. *In the unobservable models, at equilibrium, all customers follow a threshold strategy.*

PROOF. Consider a tagged customer with normalized lead time t and assume that the rest of the customers follow a strategy function σ. All customers that do not make AR have the same probability to get service. Thus, $U_\sigma(t, AR')$ does not depend on t. For brevity, we denote this value by β. From Lemma 4.1, we know that the expected payoff when making AR $U_\sigma(t, AR)$ is a non-decreasing function of t. Hence, the two expected payoffs can intersect at most once.

If $U_\sigma(t, AR) < \beta$ for all $t \in [0, 1]$, then σ is an equilibrium strategy only if none of the customers makes AR (i.e., it is a threshold strategy with $\tau = 1$). If $U_\sigma(t, AR) > \beta$ for all $t \in [0, 1]$, then σ is an equilibrium strategy only if all customers make AR (i.e., it is a threshold strategy with $\tau = 0$). However, if all customers make AR, a customer with normalized lead time 0^+ has the same probability to get service with and without AR. Thus, he is better off not making AR. Therefore, an equilibrium where all customers make AR cannot exist.

Finally, if the two expected payoff functions intersect, they can either intersect at a single point t_0 or along an interval $[t_2, t_1]$. In the first case, $U_\sigma(t, AR) < \beta$ for all $t < t_0$ and $U_\sigma(t, AR) > \beta$ for all $t > t_0$. Thus, in this case, σ is an equilibrium strategy only if it is a threshold strategy with $\tau = t_0$.

In the second case, $U_\sigma(t, AR)$ has the same value for all $t \in [t_2, t_1]$, which can only happen if $\sigma(t) = 0, \forall t \in [t_2, t_1]$ (we ignore the case of $\sigma(t) \neq 0$ over a measure zero subset of $[t_2, t_1]$, since the probability that a customer will have a normalized lead time within this subset is zero). Since none of the customers make AR in the interval $[t_2, t_1]$, and since $U_\sigma(t, AR) < \beta$ for all $t < t_2$ and $U_\sigma(t, AR) > \beta$ for all $t > t_1$, we conclude that σ is an equilibrium strategy only if it is a threshold strategy with $\tau = t_1$. □

LEMMA 4.5. *In the observable model, at equilibrium, all customers follow a threshold strategy.*

PROOF. Suppose a customer is informed that a server is available. For this case, we show that the expected payoff of not making AR is a non-increasing function of the normalized lead time while the payoff of making AR is fixed to $1 - C$. Using similar arguments as in the unobservable case, one can then show that the only possible equilibrium is a threshold strategy.

We define $\tilde{D}_{AR}(t)$ as the number of reservations made before the normalized lead time t. We need to show that for any $t_1 > t_2$, regardless of the strategy σ followed by the rest of the customers, the following holds:

$$\mathbb{P}\left(S|t_1, \tilde{D}_{AR}(t_1) < N, AR'\right) \leq \mathbb{P}\left(S|t_2, \tilde{D}_{AR}(t_2) < N, AR'\right), \quad (5)$$

where the left (right) hand side is the probability of a customer with normalized lead time t_1 (t_2) to get service, given that the number of reservations made earlier (i.e., by customers with greater lead times) is smaller than N and the chosen action is AR'. Using conditional probability, Eq. (5) can be rewritten as

$$\frac{\mathbb{P}\left(S, \tilde{D}_{AR}(t_1) < N | t_1, AR'\right)}{\mathbb{P}\left(\tilde{D}_{AR}(t_1) < N | t_1, AR'\right)} \leq \frac{\mathbb{P}\left(S, \tilde{D}_{AR}(t_2) < N | t_2, AR'\right)}{\mathbb{P}\left(\tilde{D}_{AR}(t_2) < N | t_2, AR'\right)}. \tag{6}$$

The event $\{S\}$ is contained in the event $\{\tilde{D}_{AR}(\cdot) < N\}$. Moreover, under action AR' the probability to get service does not depend on the lead time. We deduce that the numerators on both sides of the equation above are equal.

Since $\tilde{D}_{AR}(t_2)$ is stochastically larger or equal to $\tilde{D}_{AR}(t_1)$ when $t_2 < t_1$, we deduce that the denominator of the right hand side of Eq. (6) is smaller or equal to the denominator of the left hand side of Eq. (6). Thus, we have shown that Eq. (5) holds. □

After showing that a threshold strategy is the only possible equilibrium strategy, we distinguish between two types of equilibria.

Definition 4.6. None-make-AR is an equilibrium in which all customers follow a threshold strategy with threshold $\tau_e = 1$.

Definition 4.7. Some-make-AR is an equilibrium in which all customers follow a threshold strategy with threshold $\tau_e \in (0, 1)$.

Using the results obtained so far, we find next the equilibria structure for each model separately.

4.2 Equilibria structure

In this section, we show that different ranges of fees lead to different equilibria. The following theorem summarizes the main results.

THEOREM 4.8. *For each model $i = 1, 2, 3$, there exist quantities $\underline{C}$ and $\overline{C}_i \geq \underline{C}$, such that:*

—If $0 < C < \underline{C}$, there is at least one some-make-AR equilibrium.
—If $\underline{C} < C < \overline{C}_i$, there is a none-make-AR equilibrium and at least two some-make-AR equilibria.
—If $C > \overline{C}_i$, none-make-AR is the unique equilibrium.

For simplicity, we do not consider the boundary cases $C = \underline{C}$ and $C = \overline{C}_i$ in our discussion.

4.2.1 ***Unobservable model 1***. We consider the first unobservable model. For each type of equilibria, we determine the range of fees in which they may occur.

***Some-make-AR* equilibria.** If all customers follow a strategy with threshold τ_e, that strategy is an equilibrium strategy if and only if a customer with normalized lead time τ_e (referred to as a *threshold customer*) is indifferent between the actions AR and AR'. We denote $\pi_{AR}(\tau_e)$ the probability that a threshold customer gets service upon chosen action AR, and $\pi_{AR'}(\tau_e)$ the probability that a threshold customer gets service upon chosen action AR'. Hence, a strategy with threshold τ_e is an equilibrium if and only if

$$(1 - C)\,\pi_{AR}(\tau_e) = \pi_{AR'}(\tau_e), \tag{7}$$

where the left hand side of Eq. (7) is the expected payoff of AR and the right hand side is the expected payoff of AR'. Using Eq. (7), we express the fee as a function of the threshold

$$C_1(\tau_e) \triangleq 1 - \frac{\pi_{AR'}(\tau_e)}{\pi_{AR}(\tau_e)}. \tag{8}$$

Next, we develop the expressions $\pi_{AR}(\tau_e)$ and $\pi_{AR'}(\tau_e)$. The former expression corresponds to the probability that either the demand is at most N or the demand exceeds N but fewer than N customers make AR. The number of customers making AR, given $\tilde{D} = j$ with $j \geq N$, is a random variable that follows a binomial distribution. The number of trials is j and the success probability is $1 - \tau_e$. The probability that the threshold customer gets service is equal to the probability that the number of successes is at most $N - 1$. By summing this probability over all possible values of j we get:

$$\pi_{AR}(\tau_e) = \mathbb{P}\left(\tilde{D} < N\right) + \sum_{j=N}^{\infty} \sum_{i=0}^{N-1} \mathbb{P}\left(\tilde{D} = j\right) (1-\tau_e)^i \tau_e^{j-i} \binom{j}{i}. \tag{9}$$

Likewise, we have

$$\pi_{AR'}(\tau_e) = \mathbb{P}\left(\tilde{D} < N\right) + \sum_{j=N}^{\infty} \sum_{i=0}^{N-1} \mathbb{P}\left(\tilde{D} = j\right) (1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} \frac{N-i}{j+1-i}. \tag{10}$$

In that case, if the demand exceeds N but fewer than N customers make AR, service is not guaranteed. Given a demand j and a number of reservations i, the probability to get service without AR is the ratio of the number of unreserved servers $N - i$ to the number of customers that did not make AR, $j + 1 - i$.

Next, we prove that these two functions are continuous.

LEMMA 4.9. *The functions $\pi_{AR'}$ and π_{AR} are continuous functions of τ_e in the range $[0, 1]$.*

PROOF. Starting with Eq. (9) and ignoring the first term of the function which does not depend on τ_e, we need to show that the second term is continuous. The inner sum of the second term is continuous, since it is a finite sum of polynomial functions. To prove that the outer sum is continuous, we use Cauchy's uniform convergence criterion [Trench 2003]. We shall show that for any $\epsilon > 0$ there exists an integer M such that

$$\sup_{0 \leq \tau_e \leq 1} \sum_{j=n}^{m} \sum_{i=0}^{N-1} \mathbb{P}\left(\tilde{D} = j\right) (1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} < \epsilon \quad \forall n, m \geq M. \tag{11}$$

The above expression is upper bounded by $\mathbb{P}\left(n \leq \tilde{D} \leq m\right)$ which in turn is upper bounded by $\mathbb{P}\left(\tilde{D} \geq n\right)$. For any discrete distribution and $\epsilon > 0$ there exists M such that $\mathbb{P}\left(\tilde{D} \geq n\right) < \epsilon$ for any $n > M$. Thus, we have shown that Eq. (11) holds true. Since, for any $\tau_e \in [0, 1]$, $\pi_{AR}(\tau_e) \geq \pi_{AR'}(\tau_e)$, the proof is also valid for $\pi_{AR'}$.

□

Since both $\pi_{AR'}$ and π_{AR} are continuous and positive in the range $\tau_e \in [0, 1]$, we deduce that $C_1(\tau_e)$ is a continuous function in this range. Next, we observe that if all customers make AR, then the probability of service of a customer with lead time zero (i.e., the last arriving customer) does not depend on his decision. Hence, $C_1(0) = 0$. In any other case, the probability to get service is greater when making AR. Hence, $C_1(\tau_e) > 0$ for any $0 < \tau_e \leq 1$. We denote the supremum value of $C_1(\tau_e)$ as

$$\overline{C}_1 \triangleq \sup_{0 < \tau_e < 1} C_1(\tau_e). \tag{12}$$

Since the equation $C_1(\tau_e) = C$ has a solution if and only if C is smaller that the supremum value of $C_1(\tau_e)$, we conclude that a *some-make-AR* equilibrium exists if $C < \overline{C}_1$ and does not exist if $C > \overline{C}_1$.

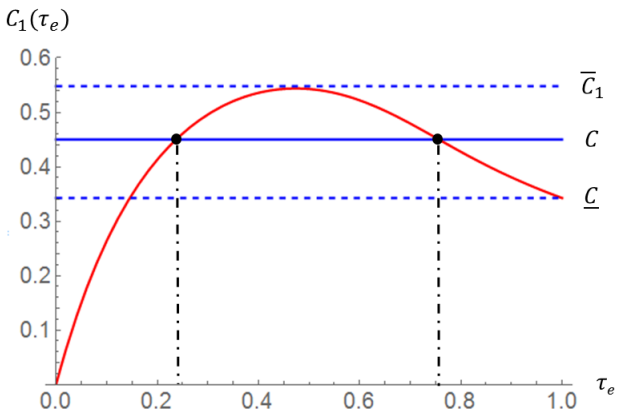

Fig. 3. An example with $N = 10$ servers, Poisson distributed demand with mean $\lambda = 15$ and AR fee $C = 0.45$. The line C and the function $C(\tau_e)$ intersect twice. Hence, there are two *some-make-AR* equilibria. Since $C(1) < C$, there is also a *none-make-AR* equilibrium.

***None-make-AR* equilibrium**. If none of the customers makes AR, all have the same expected payoff $\pi_{AR'}(1)$. A customer that deviates gets service with probability $\pi_{AR}(1) = 1$ and his payoff is $1 - C$. Thus, if the provider chooses a fee such that $1 - C < \pi_{AR'}(1)$, then none of the customers will have an incentive to deviate. On the other hand, if $1-C > \pi_{AR'}(1)$, then all the customers will have an incentive to deviate. By defining $\underline{C} \triangleq C_1(1)$, we conclude that if $C > \underline{C}$, then a *none-make-AR* equilibrium exists. If $C < \underline{C}$, then a *none-make-AR* equilibrium does not exist.

By definition $\overline{C}_1 \geq \underline{C}$. Therefore, we have shown that for any value of $0 < C < 1$, at least one equilibrium exists. Furthermore, if the interval $I = (\underline{C}, \overline{C}_1)$ is not empty (i.e., the supremum point is not reached at $\tau = 1$), then for any $C \in I$, the equation $C = C_1(\tau_e)$ must have at least two solutions due to the continuity of the function. Therefore, any fee $C \in I$ has at least two different *some-make-AR* equilibria (the exact number of *some-make-AR* equilibria depends on the number of maximal points of the function $C(\tau_e)$). See Figure 3 for an illustration.

4.2.2 ***Unobservable model 2***. In this section, we show that the second game has the same equilibria structure as the first one, but with different ranges.

***Some-make-AR* equilibria.** If all the customers follow a strategy with threshold τ_e, the probability to get service with or without making AR is calculated in the same way as in the previous model. Thus, the functions π_{AR} and $\pi_{AR'}$ can be also used in the analysis of this model. As in the first game, at a *some-make-AR* equilibrium, the threshold customer is indifferent between the two actions AR and AR'. Thus,

$$\pi_{AR}(\tau_e) - C = \pi_{AR'}(\tau_e), \tag{13}$$

where the left hand side of the equation is the expected payoff of AR, while the right hand side is the expected payoff AR'. In this model, the fee as a function of the threshold is:

$$C_2(\tau_e) \triangleq \pi_{AR}(\tau_e) - \pi_{AR'}(\tau_e). \tag{14}$$

We define

$$\overline{C}_2 \triangleq \sup_{0<\tau_e<1} C_2(\tau_e). \tag{15}$$

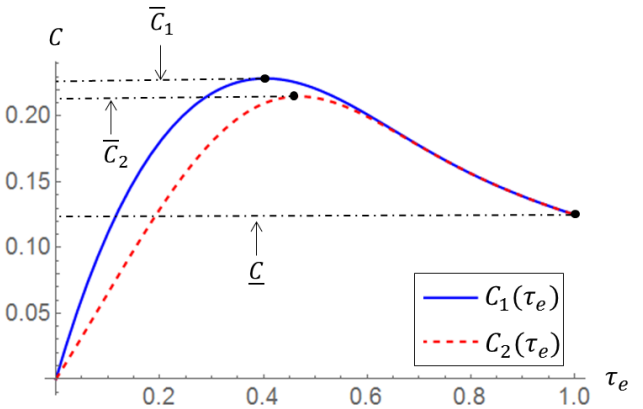

Fig. 4. The fee function in the two unobservable models with $N = 10$. The demand is a Poisson random variable with mean $\lambda = 10$.

As is the first model, $C_2(0) = 0$ and $C_2(\tau_e) > 0$ for any $\tau_e > 0$. Thus, a *some-make-AR* equilibrium exists if $C < \overline{C}_2$ and does not exist if $C > \overline{C}_2$.

***None-make-AR* equilibrium.** If none of the customers makes AR, then the expected payoffs of not making AR and the expected payoff of deviating are the same as in the first model. Therefore, the range of fees that have a *none-make-AR* equilibrium is the same as in the first model.

In conclusion, the difference between the analyses of the two games is that $\overline{C}_1$ may be different from $\overline{C}_2$. From Eq. (8) and (14), we obtain

$$\frac{C_1(\tau_e)}{C_2(\tau_e)} = \frac{1}{\pi_{AR}(\tau_e)}, \quad \forall \tau_e \in (0,1). \tag{16}$$

In any *some-make-AR* equilibrium, the probability to get service is smaller than one. Hence, for any $\tau_e \in (0,1)$, $\overline{C}_1 > \overline{C}_2$ (as illustrated in Figure 4). This result indicates that if the provider is aiming to achieve a certain fraction of customers making AR, she will have to advertise a lower fee if using the second model.

4.2.3 ***Observable model***. In this model, customers make decisions not only based on statistical information but also based on the knowledge that a server is currently available at the desired slot. Next, we show that this additional information has no effect on decisions of customers and each fee leads to the same set of equilibria as in the first unobservable model.

***Some-make-AR* equilibrium**. Consider the first model and a *some-make-AR* equilibrium with threshold τ_e, but assume that the threshold customer is being informed that a server is available, namely $\tilde{D}_{AR}(\tau_e) < N$. If making AR, his payoff is $1 - C$. The expected payoff of not making AR is $\mathbb{P}\left(S|\tau_e, \tilde{D}_{AR}(\tau_e) < N, AR'\right)$, which is the probability of the threshold customer to get service, given that all customers follow a strategy with threshold τ_e, there is at least one free server and the decision AR'. Next, we show that

$$\mathbb{P}\left(S|\tau_e, \tilde{D}_{AR}(\tau_e) < N, AR'\right) = \frac{\pi_{AR'}(\tau_e)}{\pi_{AR}(\tau_e)}. \tag{17}$$

By conditioning on the event $\{\tilde{D}_{AR}(\tau_e) < N\}$ we get

$$\mathbb{P}\left(S|\tau_e, \tilde{D}_{AR}(\tau_e) < N, AR'\right) = \frac{\mathbb{P}\left(S, \tilde{D}_{AR}(\tau_e) < N|\tau_e, AR'\right)}{\mathbb{P}\left(\tilde{D}_{AR}(\tau_e) < N|\tau_e, AR'\right)}. \tag{18}$$

Since a customer cannot get service when observing no free servers, the numerator $\mathbb{P}\left(S, D_{AR}(\tau_e) < N|\tau_e, AR'\right)$ is equal to $\mathbb{P}\left(S|\tau_e, AR'\right)$ which is equal by definition to $\pi_{AR'}\left(\tau_e\right)$.

The denominator $\mathbb{P}\left(\tilde{D}_{AR}(\tau_e) < N|\tau_e, AR'\right)$ is the probability that the threshold customer will see the event $\{\tilde{D}_{AR}(\tau_e) < N\}$ (the fact that he does not make AR is irrelevant). This in turn can be rephrased as the probability to get service when making AR exactly at the threshold point without knowing if there are free servers, which is the definition of $\pi_{AR}\left(\tau_e\right)$. Thus, $\mathbb{P}\left(\tilde{D}_{AR}(\tau_e) < N|\tau_e, AR'\right) = \pi_{AR}\left(\tau_e\right)$.

We have shown that Eq. (17) holds true for any τ_e. Using Eq. (7), we deduce that the threshold customer stays indifferent between the two actions after being informed that a server is available. Hence, we conclude that if a threshold strategy is an equilibrium strategy in the first model, it is also an equilibrium strategy in the third model.

***None-make-AR* equilibrium**. If none of the customers makes AR, the expected payoffs of not making AR and the expected payoff of deviating are the same as in the first and second models. Therefore, the range of fees that have a *none-make-AR* equilibrium is the same as in the other two models.

By noticing that the profit of the provider is defined in the same way in both models, that is, the number of customers that make AR and being served multiplied by the fee C, we obtain the following:

THEOREM 4.10. *In AR games, if AR fees are charged only from served customers, then informing customers that servers are available or hiding this information lead to the same equilibria.*

5. PROFIT MAXIMIZATION

In this section, we compare between the maximum possible expected profits in the two unobservable models. We define the maximum possible profit for model $i = 1, 2$ as

$$R_i^* = \sup_{0<\tau_e<1} R_i\left(\tau_e\right). \tag{19}$$

Under *some-make-AR* equilibrium with threshold τ_e, the number of reservations is $D_{AR}\left(\tau_e\right)$ or simply D_{AR} from now and on. In the first model, the expected profit per server is the expected number of reserved servers, multiplied by the fee and normalized by the number of servers N:

$$R_1\left(\tau_e\right) = \frac{\mathbb{E}[\min\left(D_{AR}, N\right)]C_1\left(\tau_e\right)}{N}. \tag{20}$$

In the second model, it is the expected number of reservations, multiplied by the fee and normalized by N:

$$R_2\left(\tau_e\right) = \frac{\mathbb{E}[D]\left(1-\tau_e\right)C_2\left(\tau_e\right)}{N}. \tag{21}$$

By comparing the two expressions, we state the following result:

THEOREM 5.1. *In AR games, the maximum possible profit, at equilibrium, is greater when charging the AR fee only from customers that get service and not from all customers that make AR requests.*

PROOF. We prove the theorem by showing that for any given threshold, the first model yields greater profit than the second. Namely, we show that for any value of $\tau_e \in (0, 1)$ the following holds:

$$R_1\left(\tau_e\right) > R_2\left(\tau_e\right). \tag{22}$$

From Eqs. (20), (21) and (16) we obtain that showing that $R_1(\tau_e) > R_2(\tau_e)$ is equivalent to showing that

$$\mathbb{E}[\min(D_{AR}, N)] - \mathbb{E}[D](1-\tau_e)\pi_{AR}(\tau_e) > 0. \tag{23}$$

First, we expand the first term of (23):

$$\mathbb{E}[\min(D_{AR}, N)] = \sum_{i=0}^{N} \mathbb{P}(D_{AR} = i)\, i + \sum_{i=N+1}^{\infty} \mathbb{P}(D_{AR} = i)\, N. \tag{24}$$

The PDF of D_{AR} is

$$\mathbb{P}(D_{AR} = i) = \sum_{j=i}^{\infty} \mathbb{P}(D = j)(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i}. \tag{25}$$

Combining Eq. (24) and (25) we get

$$\mathbb{E}[\min(D_{AR}, N)] = \sum_{i=1}^{N}\sum_{j=i}^{\infty} \mathbb{P}(D = j)(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} i + \sum_{i=N+1}^{\infty}\sum_{j=i}^{\infty} \mathbb{P}(D = j)(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} N. \tag{26}$$

Next, we expand $\pi_{AR}(\tau_e)$:

$$\begin{aligned}
\pi_{AR}(\tau_e) &= \mathbb{P}\left(\tilde{D} < N\right) + \sum_{i=0}^{N-1}\sum_{j=N}^{\infty} \mathbb{P}\left(\tilde{D} = j\right)(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} \\
&= \sum_{i=0}^{N-1}\sum_{j=i}^{\infty} \mathbb{P}(D = j+1)\frac{(j+1)}{\mathbb{E}[D]}(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} \\
&= \sum_{i=0}^{N-1}\sum_{j=i+1}^{\infty} \mathbb{P}(D = j)\frac{j}{\mathbb{E}[D]}(1-\tau_e)^i \tau_e^{j-i-1} \binom{j-1}{i} \\
&= \sum_{i=1}^{N}\sum_{j=i}^{\infty} \mathbb{P}(D = j)\frac{j}{\mathbb{E}[D]}(1-\tau_e)^{i-1} \tau_e^{j-i} \binom{j-1}{i-1} \\
&= \sum_{i=1}^{N}\sum_{j=i}^{\infty} \mathbb{P}(D = j)\frac{j}{\mathbb{E}[D]}(1-\tau_e)^{i-1} \tau_e^{j-i} \binom{j}{i}\frac{i}{j}.
\end{aligned} \tag{27}$$

The explanation for Eq. (27) is as follows. We start from Eq. (9). We merge the two terms in Eq. (9) and substitute $P\left(\tilde{D} = j\right)$ by the right hand side of Eq. (1). Next, we replace j by $j-1$ and start the sum at $j = i$ instead of $j = i+1$. Next, we do a similar change with the variable i. Finally, we multiple and divide the expression by $\binom{j}{i}$.

In the next step, we multiply both sides of Eq. (27) by $\mathbb{E}[D](1-\tau_e)$:

$$\mathbb{E}[D](1-\tau_e)\pi_{AR}(\tau_e) = \sum_{i=1}^{N}\sum_{j=i}^{\infty} \mathbb{P}(D = j)(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} i. \tag{28}$$

Finally, we substitute the first term of the left hand side of Eq. (23) with the right hand side of Eq. (26) and the second term of the left hand side of Eq. (23) with the right hand side of Eq. (28). We then get

$$\mathbb{E}[\min(D_{AR}, N)] - \mathbb{E}[D](1-\tau_e)\pi_{AR}(\tau_e) = \sum_{i=N+1}^{\infty}\sum_{j=i}^{\infty} \mathbb{P}(D = j)(1-\tau_e)^i \tau_e^{j-i} \binom{j}{i} N > 0, \tag{29}$$

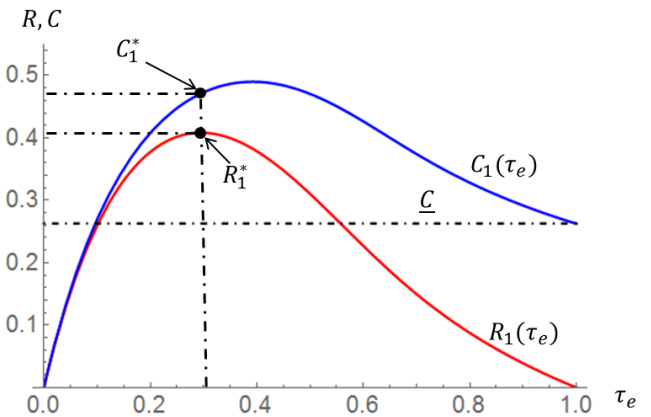

Fig. 5. An example with $N = 15$ servers and Poisson distributed demand with mean $\lambda = 20$. To obtain the maximum possible profit, the fee C must be greater than $\underline{C}$.

which completes the proof. □

The result of Theorem 5.1 is reassuring, since a mechanism that charges reservation fees only from customers getting service (first model) appears as more fair than a mechanism that charges reservation fees from all customers making AR requests (second model). Further, one may argue that the demand for service would decrease when under the latter charging scheme. The theorem proves that the latter scheme is detrimental for the provider even if the demand for service were not reduced.

6. PRICE OF CONSERVATISM

In the previous sections, we showed that in order to maximize the profit, the provider should choose the first or the third model, which yield the same profit for any given fee. In this section, we assume that the first model is chosen and investigate different fees and their impact on the profit.

By means of example, we next show that the fee that maximizes the profit may yield more than one equilibrium, where one of them yields zero profit.

Example 6.1. Consider a system with 15 servers and a Poisson distributed demand with parameter (mean) 20. In this case, the maximum profit per resource is $R_1^* = 0.41$ and it is achieved with fee $C_1^* = 0.47$. Since $\underline{C} = 0.26$, if charging C_1^*, then *none-make-AR* is also an equilibrium. Hence, charging C_1^* may yield the maximum possible profit but may also yield zero profit. The profit and fee functions are illustrated in Figure 5.

If the fee that yields the maximum possible profit is not unique, the provider may prefer a fee with smaller but guaranteed profit. In order to weigh the different options, we propose the metric of *price of conservatism* (PoC). In the rest of this section we formally define the term PoC and derive it for different settings. Since we only deal with the first model, the model index is removed in this section.

In order to have a positive guaranteed profit, the provider must choose a fee smaller than $\underline{C}$. Furthermore, if that fee has more than one equilibrium, then the guaranteed profit is defined as the minimum between the profits of the different *some-make-AR* equilibria. We define Z_C as the set of *some-make-AR* equilibria of the fee C, namely, $Z_C = \{\tau_e : C(\tau_e) = C, 0 < \tau_e < 1\}$. The maximum expected guaranteed profit is defined as follows:

$$R_g^* = \sup_{0<C<\underline{C}} \left(\inf_{\tau_e \in Z_C} R(\tau_e) \right). \tag{30}$$

The following definition captures the potential profit loss resulting from a conservative pricing decision.

Definition 6.2. The price of conservatism (PoC) is the ratio of the expected maximum possible profit R^* to the expected maximum guaranteed profit R_g^*.

Next, we evaluate the provider's profit and PoC under the assumption that the demand D is a Poisson random variable with parameter λ. We denote the number of customers not making AR by $D_{AR'}$. Due to the properties of Poisson games [Myerson 1998], D_{AR} and $D_{AR'}$ are independent Poisson random variables with parameter $\lambda(1-\tau_e)$ and $\lambda\tau_e$, respectively. Furthermore, the total number of customers and the number of customers making each action, as seen by a customer if not counting himself, has the same distributions as D, D_{AR} and $D_{AR'}$ respectively.

6.1 Single-server Case

We start with the special case $N=1$. If all customers follow a strategy with threshold τ_e, the probability that the threshold customer will get service is:

(1) If making AR:

$$\pi_{AR}(\tau_e) = e^{-\lambda(1-\tau_e)}, \tag{31}$$

which is the probability that beside the customer with lead time equals to the threshold, no one makes AR (i.e., the lead times of all other customers are smaller than the threshold).

(2) If not making AR:

$$\pi_{AR'}(\tau_e) = e^{-\lambda(1-\tau_e)} \sum_{i=0}^{\infty} \frac{e^{-\lambda\tau_e}(\lambda\tau_e)^i}{i!} \frac{1}{i+1} = \frac{e^{-\lambda}\left(-1+e^{\lambda\tau_e}\right)}{\lambda\tau_e}, \tag{32}$$

which is the probability that none of the customers makes AR, multiplied by the probability to get service given that none of the customers makes AR.

By substituting Eqs. (31) and (32) in Eq. (8) we get

$$C(\tau_e) = \frac{e^{-\lambda\tau_e} + \lambda\tau_e - 1}{\lambda\tau_e}. \tag{33}$$

LEMMA 6.3. *For the case $N=1$, $C(\tau_e)$ is a monotonic increasing function in the interval $\tau_e \in (0,1)$.*

PROOF. The derivative of $C(\tau_e)$ is:

$$\frac{\partial C}{\partial \tau_e} = \frac{e^{-\lambda\tau_e}\left(e^{\lambda\tau_e} - 1 - \lambda\tau_e\right)}{\lambda\tau_e^2}. \tag{34}$$

Since $\lambda\tau_e^2 \geq 0$, $e^{-\lambda\tau_e} \geq 0$ and $e^{\lambda\tau_e} - 1 - \lambda\tau_e > 0$ for any $\lambda > 0$ and $\tau_e \in (0,1)$, we conclude that $\frac{\partial C}{\partial \tau_e} > 0$. □

From the lemma, we infer that $\overline{C} = C(1)$. Thus, by definition,

$$\overline{C} = \underline{C} = \frac{e^{-\lambda} + \lambda - 1}{\lambda}. \tag{35}$$

Therefore, for any fee smaller than $\overline{C}$ there is no *none-make-AR* equilibrium. Furthermore, for any value of C between zero and $\overline{C}$, the equation $C = C(\tau_e)$ has a single solution and therefore the *some-make-AR* equilibrium is unique. The result is stated in the following theorem.

THEOREM 6.4. *In a single server system the equilibrium is unique and its type is:*

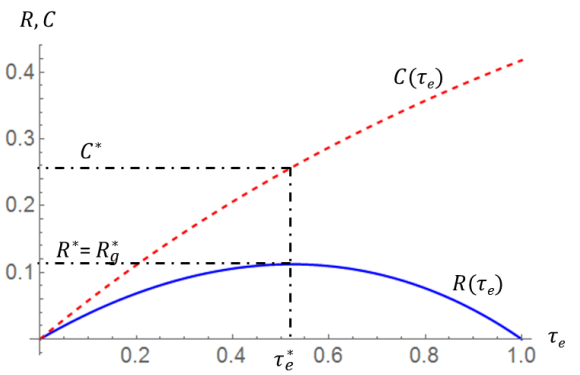

Fig. 6. An example with $N = 1$ server and mean demand $\lambda = 1.2$. The reservation fee $C(\tau_e)$ is a monotonic increasing function of the threshold τ_e. The profit function $R(\tau_e)$ is concave with maximum value achieved at $\tau_e^* = 0.523$.

— *Some-make-AR equilibrium if* $0 < C < \frac{e^{-\lambda}+\lambda-1}{\lambda}$.

—*None-make-AR equilibrium if* $C > \frac{e^{-\lambda}+\lambda-1}{\lambda}$.

The expected profit $R(\tau_e)$ in the case $N = 1$ is equal to the probability that at least one customer makes AR multiplied by the fee:

$$R(\tau_e) = \left(1 - e^{-\lambda(1-\tau_e)}\right)\left(\frac{e^{-\lambda\tau_e} + \lambda\tau_e - 1}{\lambda\tau_e}\right). \tag{36}$$

Since the equilibrium is unique, the provider will maximize her expected profit by choosing $C^* = C(\tau_e^*)$ where:

$$\tau_e^* = \arg\max_{0<\tau_e<1} R(\tau_e). \tag{37}$$

Due to the uniqueness of the equilibrium, $R^* = R_g^* = R(\tau_e^*)$. Hence:

COROLLARY 6.5. *In a single server system, the price of conservatism is 1.*

Example 6.6. We consider a system with $N = 1$ server and average demand $\lambda = 1.2$. For this system, the maximum fee that leads to a *some-make-AR* equilibrium is $\overline{C} = 0.417$. The optimal fee is obtained when $\tau_e^* = 0.523$ (i.e., when on average 47.7% of the customers make AR). This threshold is achieved when the provider sets a fee $C^* = 0.257$. The provider's expected profit in this case is $R(\tau_e^*) = 0.112$. Figure 6 shows the fee and profit as functions of the threshold τ_e.

6.2 Many-server Case

In this section we study the behavior of the system when the number of servers goes to infinity. We distinguish between overloaded and underloaded systems.

6.2.1 *Overloaded system.* We start with an overloaded system and we show that the PoC is a function of the ratio between the average demand and the number of servers.

THEOREM 6.7. *In an overloaded many-server system, where* $\lambda = \alpha N$ *and* $\alpha > 1$*, the following holds:*

$$\lim_{N\to\infty} R^* = 1. \tag{38}$$

$$\lim_{N\to\infty} R_g^* = 1 - \frac{1}{\alpha}. \tag{39}$$

Hence,

$$PoC = \frac{\alpha}{\alpha - 1}. \tag{40}$$

PROOF. In order to prove Eq. (38), we show that if the fee approaches one from below, there is a *some-make-AR* equilibrium where almost all servers are reserved.

Let $\tau_e = 1 - 1/\alpha$, hence D_{AR} is Poisson distributed with parameter N. The probability that the threshold customer gets service is equivalent to the probability that D_{AR} will be smaller than N, which in turn is equal to

$$\lim_{N\to\infty} \mathbb{P}\left(D_{AR} < N\right) = \frac{1}{2}. \tag{41}$$

Next, we show that when $\tau_e = 1 - 1/\alpha$, the probability to get service if not making AR tends to zero as $N \to \infty$. First recall Chebyshev's inequality which states that for any random variable X and real positive number Q

$$\mathbb{P}\left(|X - \mathbb{E}X|\right) \geq Q) \leq \frac{\mathrm{Var}X}{Q^2}. \tag{42}$$

Setting $Q = \delta\sqrt{N}$ where δ is a positive real number, we get from Eq. (42)

$$\mathbb{P}\left(|D_{AR} - N|) \geq \delta\sqrt{N}\right) \leq \frac{1}{\delta^2}. \tag{43}$$

In the same way, setting $Q = \epsilon\sqrt{(\alpha-1)\,N}$ where ϵ is a positive real number, we get

$$\mathbb{P}\left(|D_{AR'} - (\alpha-1)\,N|) \geq \epsilon\sqrt{(\alpha-1)\,N}\right) \leq \frac{1}{\epsilon^2}. \tag{44}$$

Hence,

$$\mathbb{P}\left(D_{AR'} \leq (\alpha-1)\,N - \epsilon\sqrt{(\alpha-1)\,N}\right) \leq \frac{1}{\epsilon^2}. \tag{45}$$

From Eqs. (43) and (45), we deduce that with probability one the number of free servers $D_{AR} - N$ is $O\left(\sqrt{N}\right)$ while $D_{AR'}$ is $(\alpha-1)\,N + O\left(\sqrt{N}\right)$. Hence, for any $\alpha > 1$, as $N \to \infty$, a customer that does not make AR will get service with probability zero.

We showed that when $N \to \infty$ and τ_e is such that on average N customers make AR, the expected payoff of the threshold customer tends to $0.5(1 - C)$ if making AR and to zero if not making AR. Thus, a strategy with threshold τ_e is an equilibrium only if C tends to one. Therefore, we conclude that there exists a value of τ_e such that on average N customers make AR while the fee is almost one. Therefore, in an overloaded system:

$$\lim_{N\to\infty} R^* = 1. \tag{46}$$

Next, we show that Eq. (39) holds. If none of the customers makes AR, they all have the same probability to get service. Again, due to Chebyshev's inequality, for any $\delta > 1$, the following holds:

$$\mathbb{P}\left(\alpha N - \delta\sqrt{\alpha N} \leq D \leq \alpha N + \delta\sqrt{\alpha N}\right) \leq \frac{1}{\delta^2}. \tag{47}$$

In other words, with probability one the demand is $\alpha N + O(\sqrt{N})$. In this case, as $N \to \infty$ the fraction of customers getting service converges to $1/\alpha$. Hence,

$$\lim_{N\to\infty} \pi_{AR'}(1) = \frac{1}{\alpha}. \tag{48}$$

On the other hand, when deviating from the none-make-AR strategy, the probability to get service is $\pi_{AR}(1) = 1$. Thus,

$$\lim_{N\to\infty} \underline{C} = 1 - \frac{1}{\alpha}. \tag{49}$$

Next, we show that in the overloaded system, if $C < \underline{C}$, then in any *some-make-AR* equilibrium almost all servers are reserved. By contradiction, we assume that there exists a *some-make-AR* equilibrium with threshold τ_e such that, $C(\tau_e) < 1-1/\alpha$ and $\mathbb{E}[D_{AR}] = \delta N$ where $0 < \delta < 1$. In this case, the probability of the threshold customer to get service if making AR converges to one as $N \to \infty$. Thus, his expected payoff is greater than $1/\alpha$. If not making AR, his probability to get service is smaller than $1/\alpha$ (which is the probability to get service if none makes AR). Therefore, the expected payoff of the threshold customer is greater if making AR than if not making AR, which contradicts the definition of a *some-make-AR* equilibrium. Hence, we have shown that the assumption cannot hold true. Thus, with probability one, the number of reservation will be at least $N + o(N)$. Therefore, with probability one, the ratio between the number of free servers and the number of servers is zero. The provider will maximize her guaranteed profit by advertising a fee just below $\underline{C}$ and we finally obtain

$$\lim_{N\to\infty} R_g^* = 1 - \frac{1}{\alpha}. \tag{50}$$

□

The results indicate that if α is almost one and none of the customers makes AR then, in order to persuade customers to deviate, the provider will have to advertise a fee close to zero. Such fee will yield almost zero profit per resource. In other words, although there is an equilibrium that yields a profit per resource of almost one, if initially none of the customers makes AR, any fee the provider will advertise will not significantly increase her profit.

6.2.2 *Underloaded system.* In an underloaded many-server system we show that any fee leads to an asymptotically zero profit.

THEOREM 6.8. *In an underloaded many-server system, where $\lambda = \alpha N$ and $\alpha < 1$, the following holds:*

$$\lim_{N\to\infty} R^* = R_g^* = 0. \tag{51}$$

PROOF. Given any $\alpha < 1$ and any $\epsilon > 0$, we can find a large enough N such that

$$\mathbb{P}(D > N) \leq \epsilon. \tag{52}$$

In other words, for large enough N, the probability that the demand will exceed the number of servers tends to zero. In this case, the dominant strategy of all customers, regardless of their lead time, is not to make AR. Hence,

$$\lim_{N\to\infty} \overline{C} = 0. \tag{53}$$

□

7. CONCLUSIONS AND FUTURE WORK

In this paper we introduce advance reservation games: games where customers are asked to pay a fee if they wish to reserve a future resource in advance. First, we show that, at equilibrium, either all customers with lead times greater than some threshold make AR or none of them makes AR. Next, we prove the existence of at least one Nash equilibrium and find the range of fees that determine

each equilibrium. Furthermore, we show that a fee may yield more than one equilibrium, with one of them bringing zero profit to the provider. Next, we show that providing information to the customers about the availability of servers has no impact on the game outcome. However, charging a fee from all customers attempting to reserve a server can only reduce the provider's profits.

In order of a provider to decide on a proper AR fee, we propose the concept of *Price of Conservatism (PoC)* which corresponds to the ratio of the maximum possible expected profit to the maximum guaranteed expected profit. A greater PoC indicates greater potential profit loss if the provider opts to be conservative. We focus on the models where charges are collected only from the customers getting service and assume that the demand is Poisson distributed. First, we show that in a single server system the equilibrium is unique. Thus, $PoC = 1$ and the provider experiences no loss. Next, we show that in an overloaded many-server system where the average demand is $\lambda = \alpha N$ with $\alpha > 1$, the maximum possible expected profit tends to one, while, the maximum guaranteed expected profit tends to $1 - 1/\alpha$ as $N \to \infty$. Hence $PoC = \alpha/(\alpha - 1)$, which increases in an unbounded fashion as α approaches 1 from above. Finally, we show that in an underloaded many-server system, the provider cannot make profit.

The extensions of advance reservation games to more complex settings (e.g., with users differing in their utilities) and analysis of PoC in other systems represent interesting directions for future work.

Acknowledgment

This research was supported in part by the US National Science Foundation under grant CNS-1117160.

REFERENCES

AAZAM, M. AND HUH, E.-N. 2015. Cloud broker service-oriented resource management model. *Transactions on Emerging Telecommunications Technologies*.

ALTMAN, E. AND SHIMKIN, N. 1998. Individual equilibrium and learning in processor sharing systems. *Operations Research 46,* 6, 776–784.

AVINERI, E. 2004. A cumulative prospect theory approach to passengers behavior modeling: waiting time paradox revisited. In *Intelligent Transportation Systems*. Vol. 8. Taylor & Francis, 195–204.

BALACHANDRAN, K. 1972. Purchasing priorities in queues. *Management Science 18,* 5-Part-1, 319–326.

BERTSIMAS, D. AND SHIODA, R. 2003. Restaurant revenue management. *Operations Research 51,* 3, 472–486.

BLANCHET, J., GLYNN, P., AND LAM, H. 2009. Rare event simulation for a slotted time M/G/s model. *Queueing Systems 63,* 1-4, 33–57.

BUYYA, R., YEO, C. S., VENUGOPAL, S., BROBERG, J., AND BRANDIC, I. 2009. Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems 25,* 6, 599–616.

CHARBONNEAU, N. AND VOKKARANE, V. M. 2012. A survey of advance reservation routing and wavelength assignment in wavelength-routed wdm networks. *Communications Surveys & Tutorials, IEEE 14,* 4, 1037–1064.

COHEN, R., FAZLOLLAHI, N., AND STAROBINSKI, D. 2009. Path switching and grading algorithms for advance channel reservation architectures. *Networking, IEEE/ACM Transactions on 17,* 5, 1684–1695.

DODGE, Y. 2006. *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

EDELSON, N. M. AND HILDERBRAND, D. K. 1975. Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society*, 81–92.

GUÉRIN, R. A. AND ORDA, A. 2000. Networks with advance reservations: The routing perspective. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. Vol. 1. IEEE, 118–127.

HASSIN, R. J. AND HAVIV, M. 2003. *To Queue or Not to Queue: Equilibrium Behaviour in Queueing Systems*. Vol. 59. Kluwer Academic Pub.

HAVIV, M., KELLA, O., AND KERNER, Y. 2010. Equilibrium strategies in queues based on time or index of arrival. *Probability in the Engineering and Informational Sciences 24,* 1, 13.

HAVIV, M. AND ROUGHGARDEN, T. 2007. The price of anarchy in an exponential multi-server. *Operations Research Letters 35,* 4, 421–426.

JAIN, R., JUNEJA, S., AND SHIMKIN, N. 2011. The concert queueing game: to wait or to be late. *Discrete Event Dynamic Systems 21,* 1, 103–138.

KANG, C. G. AND TAN, H. H. 1993. Queueing analysis of explicit priority assignment partial buffer sharing schemes for ATM networks. In *INFOCOM'93. Proceedings. Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies. Networking: Foundation for the Future, IEEE*. IEEE, 810–819.

KAUSHIK, N. R., FIGUEIRA, S. M., AND CHIAPPARI, S. A. 2006. Flexible time-windows for advance reservation scheduling. In *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006. MASCOTS 2006. 14th IEEE International Symposium on*. IEEE, 218–225.

KOUTSOUPIAS, E. AND PAPADIMITRIOU, C. 1999. Worst-case equilibria. In *STACS 99*. Springer, 404–413.

LIBERMAN, V. AND YECHIALI, U. 1978. On the hotel overbooking problem-an inventory system with stochastic cancellations. *Management Science 24,* 11, 1117–1126.

MASUDA, Y. AND WHANG, S. 2006. On the optimality of fixed-up-to tariff for telecommunications service. *Information Systems Research 17,* 3, 247–253.

MYERSON, R. B. 1998. Population uncertainty and poisson games. *International Journal of Game Theory 27,* 3, 375–392.

NAOR, P. 1969. The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, 15–24.

REIMAN, M. I. AND WANG, Q. 2008. An asymptotically optimal policy for a quantity-based network revenue management problem. *Mathematics of Operations Research 33,* 2, 257–282.

ROSS, K. 1995. *Multiservice loss networks for broadband telecommunications networks*. Springer-Verlag NY.

SMITH, W., FOSTER, I., AND TAYLOR, V. 2000. Scheduling with advanced reservations. In *Parallel and Distributed Processing Symposium, 2000. IPDPS 2000. Proceedings. 14th International*. IEEE, 127–132.

SOTOMAYOR, B. 2009. Haizea and private clouds, blog.dsa-research.org blog http://blog.dsa-research.org/?p=138.

TRENCH, W. F. 2003. *Introduction to real analysis*. Prentice Hall/Pearson Education Upper Saddle River, NJ.

VIRTAMO, J. T. 1992. A model of reservation systems. *Communications, IEEE Transactions on 40,* 1, 109–118.