Received: 11 July 2020 | Accepted: 21 February 2022

DOI: 10.1111/poms.13705

POMS

ORIGINAL ARTICLE

Pay to activate service in vacation queues

Zhongbin Wang[1] | **Yunan Liu**[2] | **Lei Fang**[3]

[1]College of Management and Economics, Tianjin University, Tianjin, China

[2]Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina, USA

[3]Business School, Nankai University, Tianjin, China

Correspondence
Yunan Liu, Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695, USA.
Email: yliu48@ncsu.edu

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 72001118, 72132007

Handling Editor: Michael Pinedo

Abstract

We study a vacation queueing model where an arriving customer, upon finding the server to be on vacation, is offered an opportunity to pay a fee to instantaneously end the server's vacation, which is referred to as *pay-to-activate-service* (PTAS). If no one utilizes PTAS, the service will automatically resume when the system's workload reaches a critical level. We investigate customers' equilibrium strategies: (i) joining or balking and (ii) if joining, accepting PTAS or rejecting PTAS, in response to such a mechanism; we show that customers' equilibrium strategies exhibit both *avoid-the-crowd* (ATC) and *follow-the-crowd* (FTC) types of behavior. Our results indicate that the adoption of PTAS is efficient in improving the system performance (e.g., revenue and throughput) when the demand volume is intermediate. We also discover that, upon selecting the appropriate queue-length information disclosure policy, the service provider has to trade off between collecting a higher revenue through PTAS and improving the system throughput, because revealing the queue-length information will impact the aforementioned two performance metrics in opposing directions. Finally, we compare our new setting to other common mechanisms including regular vacation queues and pay-for-priority queues.

KEYWORDS
equilibrium analysis, *N*-policy, pay to activate service, strategic customers, vacation queue

1 | INTRODUCTION

In service systems, allowing servers to take vacations when the congestion level is low can help reduce the system's operating cost and arouse servers' enthusiasm, see Tian and Zhang (2006). An efficient vacation mechanism in practice is to let the server resume service once the system's workload reaches a critical level. For example, in make-to-order production systems with high setup costs, the production line usually begins to operate only when the number of tasks reaches a critical level, see Guo and Hassin (2011) and Li et al. (2016) for more detailed discussions of make-to-order production systems.

This type of vacation termination rule is referred to as the *N*-policy (Yadin & Naor, 1963), that is, the vacation continues as long as the total number of waiting customers is below a threshold N and it terminates otherwise. In a vacation queue with a relatively low threshold N, customers' waiting times have little variations because it only takes a few more arrivals to activate the server (if it is not already active). For instance, a special case of $N = 1$ is equivalent to the conventional queueing model without vacation. On the other hand, when the vacation termination threshold is high, customers' delays can exhibit significant fluctuations because an arrival at the beginning of the server's vacation time has to passively wait for $N - 1$ additional arrivals before the service eventually resumes, while the last arrival will immediately activate the service. This may give rise to an issue on service unfairness (variance of customers' waiting times has been proven a useful metric for the service fairness, see for example Cao et al. (2021) and the references therein); also see Liu and Whitt (2014) and Aras et al. (2018) for analysis on variance of waiting times.

In this paper, we consider a new mechanism in a vacation queueing system, where each arriving customer, upon finding the server to be on vacation, is offered an opportunity to pay a fee to instantaneously end the server's vacation; we

Accepted by Michael Pinedo, after three revisions.

refer to this option as *pay-to-activate-service* (PTAS). Ideas similar to PTAS have already been implemented in several practices. One relevant application is manufacturing systems of high-end products, such as new energy vehicle (NEV). Since the Chinese government has terminated the NEV subsidy policy in 2020,[1] NEV productions have experienced a significant slowdown. Manufacturers intend to adopt more cautious production plans; they begin producing new cars only when the number of pending orders reaches a certain level, unless consumers are willing to pay a premium; see Li and Xu (2020). Another example is the production line of labor-intensive products (e.g., fashion clothing and designer handbags), where the products are made by one designer exclusively. Due to the labor-intensive nature, the designer often resumes to work when there is a sufficient number of orders or a considerable premium is paid by some highly delay-sensitive consumers; see Guo and Hassin (2011) and Li et al. (2016). Other applications close to the PTAS mechanism include online group buying of beauty products (Hu et al., 2021) and car-pooling in ride-sharing platforms (Jacob & Roet-Green, 2021) in those the service may be initiated either by having a sufficient number of requests or fees paid by impatient consumers.

PTAS enables customers to gain proactive control of their own service experience because, if they deem PTAS to be worthy, customers no longer need to wait for other (future) customers to help advance the service process. In some sense, PTAS can help address the fairness issue from the customers' perspective, it turns the control of the server's state from passive to active. In addition, the impact of PTAS is beyond the scope of an individual customer. A customer adopting PTAS may help improve the service experience of other customers, including (i) *old customers* already buffered in queue awaiting the server to return to work and (ii) *new customers* arriving in the near future before the server takes another vacation.

In our vacation queueing model endowed with PTAS, customers are delay-sensitive and strategic. They make the following one-time decisions immediately upon their arrivals: (i) to join or to balk and (ii) if joining, to pay (for PTAS) or not to pay, in anticipation of their expected individual welfare. If no one utilizes PTAS and the server is on vacation, the service will automatically resume when the queue length reaches some designated threshold N. At a glance, the idea of using PTAS to reduce delay can be in some sense similar to paying for a service priority (e.g., the FastPass service in Disney World for reduced waiting times and the Amazon Prime service for quick deliveries). Nevertheless, we draw a distinction: In conventional priority service models, purchasing a higher priority will only reduce the delay cost for that particular customer, while PTAS proposed here will help advance the service process for the entire waiting line, which benefits all customers present in the system. We study customers' equilibrium joining-and-purchasing strategies in response to such a new mechanism under two main information policies: observable and unobservable queue length. From the perspective of the service provider, we aim to answer the following questions: *Does the implementation of PTAS help generate a higher revenue and system throughput? If yes, when is the improvement most significant?What is the optimal information disclosure policy when PTAS is in effect?*

1.1 | Literature review

Our analysis has points of contact to four extant streams of research: (i) strategic behavior in vacation queues, (ii) strategic behavior in priority queues, (iii) two-dimensional customer strategies, and (iv) information provision policies.

1.1.1 | Strategic behavior in vacation queues

The research on strategic customers in queues was pioneered by Naor (1969), where arriving customers decide on whether to join an M/M/1 queue based on the available queue length. The case of unobservable queue for an M/M/1 model was developed by Edelson and Hilderbrand (1975). Following Naor (1969), strategic customer behavior in queueing systems has been widely studied in the literature, see Hassin and Haviv (2003), Stidham Jr (2009), and Hassin (2016) for comprehensive reviews. We hereby focus on reviewing works on vacation queues. The first work on vacation queues with N-threshold policy dates back to Yadin and Naor (1963) where the service is resumed whenever N or more customers are present in the waiting line. Following Yadin and Naor (1963), various vacation queue models have been studied during the past decades. Interested readers can refer to the comprehensive monograph of Tian and Zhang (2006) and the references therein. In particular, Economou and Kanta (2008) studied an observable queue with server vacations due to breakdowns and developed the equilibrium joining strategy for customers. Guo and Hassin (2011) was the first work that studied customer equilibrium behavior in an N-policy vacation queue, and they discovered that customers may prefer to join a longer queue, in anticipation that the service will start sooner. Such an effect is referred to as *positive externalities* in the queueing game literature. The case of heterogeneous customers was investigated by Guo and Hassin (2012).

1.1.2 | Strategic behavior in priority queues

Allowing customers to pay for priority has been proven an efficient way to increase service profit and social welfare in queueing systems. Adiri and Yechiali (1974) were the first to develop the pure equilibrium priority purchasing strategy in an observable queueing model. Their results were later extended by Hassin and Haviv (1997) to allow for mixed strategies in the same settings. Gavirneni and Kulkarni (2016) studied the equilibrium strategy in unobservable queues having heterogeneous customers. Wang et al. (2019) conducted a comparison analysis for the equilibrium performance of a priority queue under different information structures. The partial priority scheme is proposed by Yang et al. (2022) in Covid-19 testing queues. Offering PTAS to customers looks in a way similar to allowing them to purchase a higher

priority over others. Nevertheless, we draw a major distinction: In priority queues, the purchasing behavior is shown to be of pure follow-the-crowd (FTC) type, that is, a customer is more inclined to purchase priority when others do so as well, see, for example, Hassin and Haviv (1997, 2003). However, in our vacation queue model endowed with PTAS, as we will show later, FTC and avoid-the-crowd (ATC) often coexist when the threshold N is neither too large nor too small. Such a distinction is due to the fact that a PTAS-purchasing customer not only will reduce her own delay cost but also will benefit all other customers present in the system.

1.1.3 | Two-dimensional customer strategies

In contrast to previous queueing game literature that focuses on either the customers' joining strategy or purchasing strategy (e.g., in priority queues), our framework allows customers' strategy to be a combination of both. At the heart of our equilibrium analysis is to establish the two-dimensional joining-and-purchasing strategy. To our best knowledge, only a few papers in the queueing game literature have investigated this type of two-dimensional equilibrium strategy. Hassin and Roet-Green (2017) studied a queueing model where customers first decide on whether to inspect the queue length and then make a second join-or-balk decision. This work was later extended by Hassin and Roet-Green (2018) to a two-server queue with inspection costs. Wang et al. (2019) studied the joining and priority purchasing strategy in a priority queueing model. Besides the joining decisions, Cui et al. (2020) and Yang et al. (2021) considered queueing models where customers can choose to pay to improve their queueing positions. Other two-dimensional settings can be found in referral priority models (Yang & Debo, 2019), online retailing queueing models (Wang et al., 2021a), multichannel service models with product exchange (Sun et al., 2022a), and restaurant models allowing orders to be placed ahead and then picked up later (Sun et al., 2022b). Motivated by cloud services, Dierks and Seuken (2021) solved the service provider's profit optimization problem and established a multidimensional equilibria. Abhishek et al. (2012) considered two pricing schemes for selling cloud services to two user classes. Gao et al. (2019) investigated a service system with two competing firms offering services under two different pricing and service rules, in which arriving customers need to decide on (i) whether to receive service; (ii) if yes, from which firm; and (iii) if choosing the bid-based firm, the amount of her bid. We emphasize that the consideration of a two-dimensional customer strategy adds significant complexity to the equilibrium analysis.

1.1.4 | Information revelation policies

There is a stream of queueing literature that studies the impact of information provision on queueing outcomes. By studying social welfare under both full and no queue information, Hassin (1986) discovered that the revelation of real-time queue length improves the social welfare because such information helps better match service capacity with customer demand. Chen and Frank (2004) investigated the system throughput under the two aforementioned information provision policies and discovered that delayed information may have both positive and negative effects on the system throughput. Simhon et al. (2016) considered the optimal information disclosure problem in an M/M/1 queue, and concluded that the commonly adopted threshold policy is never optimal. Hassin and Koshman (2017) proposed a new profit-maximizing mechanism in that customers will be notified whether the queue length is below a certain threshold. Hu et al. (2018) found that throughput and social welfare can be unimodal in the fraction of informed customers; their findings infer that creating the "right" amount of information heterogeneity among customers may lead to improved outcomes. Similar results can be found in the retrial and priority queueing models, see Wang and Wang (2019) and Wang and Fang (2022). Anunrojwong et al. (2020) studied the effective design of information policies with the objective of reducing congestion in social services. Recently, Lingenbrink and Iyer (2019) solved a long-standing open problem on the optimal signaling mechanism in unobservable queues; their illuminating findings suggested that such a signaling mechanism can be effective in achieving the optimal revenue in settings where state-dependent pricing is infeasible.

1.2 | Contributions and organization

In summary, we make the following contributions.

- **Benefit of PTAS**. To the best of our knowledge, the present work is the first to study a vacation queueing model endowed with PTAS, which can be viewed as an extension of regular vacation queues operated under the N-policy. The ingenuity of PTAS lies in its ability to allow the server (when on vacation) to be activated immediately by arriving customers, giving them more controls over their service experiences. We study customers' equilibrium joining-and-purchasing strategies and the corresponding system performance measures. Our results show that, from the service provider's perspective, the model with PTAS can achieve a higher system-level performance than that without PTAS; and from the customers' perspective, they also benefit from receiving additional welfare through the utilization of PTAS.
- **Information provision policies**. We study two base information policies. In case the queue length is unobservable, we discover that the equilibrium PTAS purchasing behavior shifts from FTC to ATC as the threshold N increases. Specifically, the equilibrium is FTC when N is small (e.g., $N = 2$) and is ATC when N is large (e.g., $N \geq 5$). And interestingly, when N is intermediate (e.g., $n = 3, 4$), the equilibrium exhibits both FTC and ATC behavior. When the queue length is observable, we establish a

subgame-perfect equilibrium (SPE) strategy, which, on an equilibrium path, reduces to a parameter-dependent threshold policy. We also consider the third information policy, called the "no-information" case, where neither the queue length nor the server's state is available.

- **Nonmonotonic performance functions**. Under all information policies, we show that the equilibrium system throughput is not necessarily increasing in the service reward, and that the PTAS revenue is in fact decreasing in the service reward. These seem to counter the general intuitions. Another interesting result is that the PTAS revenue is a unimodal function in the demand volume (it is close to 0 when the demand volume is either too small or too large). The performance under different information policies is studied and compared. To gain understanding of these results, we conduct numerical experiments and provide in-depth discussions.

1.3 | Organization of the paper

The rest of the paper is structured as follows. The model description is given in Section 2. In Section 3, we study an unobservable vacation queue model endowed with PTAS. We conduct equilibrium analysis in three steps: We first report the equilibrium joining strategy under a fixed PTAS purchasing probability (Propositions 2 and 3); we next develop the equilibrium PTAS purchasing strategy with exogenous arrival rates (Proposition 4); and finally, we integrate results in the previous two steps to establish the joint joining-and-purchasing strategy (Theorem 1). In Section 4, we study an observable vacation queue endowed with PTAS and characterize the SPE strategies (Theorems 2 and 3); we also provide the system performance in equilibrium. In Section 5 we compare the system performance under the two base information policies, investigate the revenue/pricing implications, and contrast our PTAS setting to other common mechanisms. We develop some extensions of our base models in the Supporting Information and draw concluding remarks in Section 6. All proofs are given in the Supporting Information.

2 | MODEL DESCRIPTION

We consider a production system with N-policy, where an arriving customer, upon finding the server to be on vacation, is offered a one-time opportunity to pay a fee to instantaneously end the server's vacation. Such a mechanism is called PTAS. Specifically, we study an M/M/1 queue having arrivals according to a Poisson process with rate Λ, and *independent and identically distributed* (i.i.d.) service times that are exponentially distributed with rate μ.[2] We denote by $\rho \equiv \Lambda/\mu$ the system's workload. The server alternates between two states: *active* (i.e., at service) and *inactive* (i.e., on vacation). When the server is active, waiting customers are being served under the *first-come first-served* (FCFS) discipline; when no customer is present, the server becomes inactive (takes a vacation). The server's vacation will end either (i) when the total number of waiting customers reaches a critical level N or (ii) an arriving customer adopts PTAS. The fee of PTAS is $P > 0$.

Customers are homogeneous and delay-sensitive. They incur a delay cost at rate C during their total sojourn time and receive a reward R upon completion of their services. All arriving customers are informed of the state of the server (because it may appear to be unreasonable to ask customers to pay for PTAS when the server is already active). Knowing the server's state, each customer needs to decide whether to join the queue or to balk; if the server is on vacation (inactive), a joining customer also has to decide whether to accept PTAS or reject PTAS (in the latter case she relies on future customers to activate the service either by increasing the queue length to N or by adopting PTAS). In summary, arrivals finding the server to be on vacation have three pure strategies: (i) *balking*; (ii) *joining and accepting PTAS*; (iii) *joining and rejecting PTAS*. We assume that customers are risk-neutral, and they aim to maximize their expected utilities conditional on the system state observed upon arrival.[3]

We first study two main information policies: (1) unobservable queue (so customers' behavior will rely on their anticipation of the expected mean delay) and (2) observable queue (so that customers can make strategic decisions using the real-time queue length).[4] We conduct equilibrium analysis in both cases and study which one provides more benefits from the service provider's perspective. To model the system dynamics as a *continuous-time Markov chain* (CTMC), we track the two-dimensional process $\{(B(t), X(t)), t \geq 0\}$ where $B(t) = 1$ ($B(t) = 0$) if the server is active (inactive) at time t, and $X(t)$ is the total number of customers in the system at t. Under the N-policy, the state space of this CTMC is

$$\mathcal{S} = \{(0, n) : 0 \leq n \leq N - 1\} \cup \{(1, n) : n \geq 1\}. \quad (1)$$

3 | UNOBSERVABLE QUEUE

In this section, we establish customers' equilibrium strategy when the queue length is unobservable. In Section 3.1 we first study the steady-state system performance under an arbitrary (mixed) strategy. In Sections 3.2–3.4, we fully describe the joint joining-and-purchasing equilibrium strategy.

3.1 | Preliminaries

Because the server's state is observable, we let λ_0 and λ_1 be the *effective* customer arrival rates when the server is inactive and active, respectively, and let p be the probability that a customer finding an inactive server accepts PTAS. Therefore, customers' strategy can be described by the triplet $\pi \equiv (p, q_0, q_1)$, where $q_0 = \lambda_0/\Lambda$ and $q_1 = \lambda_1/\Lambda$. We also define $\rho_0 \equiv \lambda_0/\mu$ and $\rho_1 \equiv \lambda_1/\mu$ as the effective traffic intensities in the two cases.

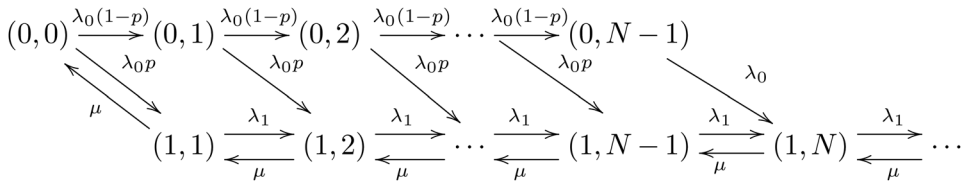


FIGURE 1 State transition diagram under strategy (p, q_0, q_1)

3.1.1 | Steady-state performance

The state transition diagram of the CTMC $\{(B(t), X(t)), t \geq 0\}$ is depicted in Figure 1. To understand Figure 1, for example, consider the state (1,2), one of two events may occur next: a service completion with rate μ (in which case state (1,1) follows) because the server is active, or a new customer arrival with rate λ_1 (in which case (1, 3) follows). On the other hand, since an external customer arrival seeing $B(t) = 0$ can activate the server with probability p, the CTMC will move from states (0,1), (1, 1), and (1,3) to state (1,2) with rate $\lambda_0 p$, λ_1, and μ, accordingly. Transitions regarding other states are similar.

Let $\pi_{i,n}$ denote the steady-state probabilities of (B, X), which satisfy the following balance equations

$$\pi_{0,0}\lambda_0 = \pi_{1,1}\mu, \tag{2}$$

$$\pi_{0,i}\lambda_0 = \pi_{0,i-1}\lambda_0(1-p), \quad i \in \{1, 2, \ldots, N-1\}, \tag{3}$$

$$\pi_{1,i}(\lambda_1 + \mu) = \pi_{0,i-1}\lambda_0 p + \pi_{1,i-1}\lambda_1 + \pi_{1,i+1}\mu, \quad i \in \{1, 2, \ldots, N-1\}, \tag{4}$$

$$\pi_{1,N}(\lambda_1 + \mu) = \pi_{0,N-1}\lambda_0 + \pi_{1,N-1}\lambda_1 + \pi_{1,N+1}\mu, \tag{5}$$

$$\pi_{1,i}(\lambda_1 + \mu) = \pi_{1,i-1}\lambda_1 + \pi_{1,i+1}\mu, \quad i \in \{N+1, N+2, \ldots\}. \tag{6}$$

For any given strategy (p, q_0, q_1), the steady-state probabilities and expected queue length along with other system performance measures are given below.

Proposition 1 (Steady-state system performance under a given strategy (p, q_0, q_1)). *Consider an unobservable M/M/1 vacation queue with PTAS. Assume that all customers follow strategy (p, q_0, q_1).*

(*i*) *The steady-state probabilities are*

$$\pi_{0,i} = C_0 p(1-p)^i, \quad 0 \leq i \leq N-1;$$

$$\pi_{1,i} = C_0 \frac{\rho_0 p\left[(1-p)^{i\wedge N} - \rho_1^{i\wedge N}\right]\rho_1^{(i-N)^+}}{1-p-\rho_1}, \quad i \geq 1; \tag{7}$$

where $C_0 \equiv \frac{1-\rho_1}{[1-(1-p)^N](1+\rho_0-\rho_1)}$, $x \wedge y \equiv \min(x, y)$, *and* $x^+ \equiv \max\{x, 0\}$.

(*ii*) *The expected queue length is*

$$\mathbb{E}[X] = Q(p, q_0, q_1) = \bar{Q}(q_0, q_1) + Q_N(p), \tag{8}$$

where $\bar{Q}(q_0, q_1) \equiv \frac{\rho_0}{(1-\rho_1)(1+\rho_0-\rho_1)}$, $Q_N(p) \equiv \frac{(1-p)[1+(N-1)(1-p)^N - N(1-p)^{N-1}]}{p[1-(1-p)^N]}$.

(*iii*) *The conditional expected waiting times of an arriving customer seeing an inactive server and an active server are*

$$w_0(p, q_0) = \frac{1}{\mu} + \frac{Q_N(p)}{\mu}\left[1 + \frac{1}{\rho_0}\right] \quad \text{and}$$

$$w_1(p, q_1) = \frac{1}{\mu}\left[\frac{1}{1-\rho_1} + 1 + Q_N(p)\right]. \tag{9}$$

(*iv*) *The system throughput is*

$$\lambda^u(q_0, q_1) = \frac{\lambda_0}{1+\rho_0-\rho_1}. \tag{10}$$

(*v*) *The service provider's revenue collected by selling PTAS is*

$$\Pi^u(p, q_0, q_1) = \frac{(1-\rho_1)p\lambda_0 P}{1+\rho_0-\rho_1}. \tag{11}$$

Remark 1 (Decomposition of steady-state queue length). According to (8), the mean steady-state queue length $Q(p, q_0, q_1)$ can be separated to two parts: $\bar{Q}(q_0, q_1)$ and $Q_N(p)$. The first term $\bar{Q}(q_0, q_1)$, which is independent of threshold N and purchasing probability p, is the mean queue length in a conventional M/M/1 queue having state-dependent arrivals (with arrival rates λ_0 and λ_1 when the server is inactive and active, respectively); the second term $Q_N(p)$, which is independent of λ_0 and λ_1, can be interpreted as the extra queue size incremented due to the vacation mechanism. This term will become smaller when p increases (it will vanish if everyone adopts PTAS). Because $Q_N(p)$ will play an important role in characterizing customers' equilibrium strategies, we next establish the structural properties for $Q_N(p)$.

Lemma 1. *$Q_N(p)$ is decreasing in p, with $Q_N(0) = (N-1)/2$ and $Q_N(1) = 0$.*

Remark 2 (Important special cases). To make contact with existing results in the literature, we advocate that our model is general and covers several previously studied queueing models. If no one adopts PTAS (i.e., $p = 0$), we have

$$w_0(0, q_0) = \frac{1}{\mu} + \frac{N-1}{2\mu}\left[1 + \frac{1}{\rho_0}\right] \quad \text{and}$$
$$w_1(0, q_1) = \frac{1}{\mu}\left[\frac{1}{1-\rho_1} + \frac{N+1}{2}\right], \tag{12}$$

which coincide with the waiting time formulas for the N-policy vacation queue model (Guo & Li, 2013). On the other hand, if everyone accepts PTAS (i.e., $p = 1$), we have

$$w_0(1, q_0) = \frac{1}{\mu} \quad \text{and} \quad w_1(1, q_1) = \frac{1}{\mu} + \frac{1}{\mu - \lambda_1}, \tag{13}$$

which are the conditional expected delays of customers finding the server to be inactive and that of those finding the server to be active in a standard M/M/1 queue, respectively. Finally, if we take $N \to \infty$, we have

$$\lim_{N\to\infty} w_0(p, q_0) = \frac{1}{\mu} + \frac{1-p}{p\mu}\left[1 + \frac{1}{\rho_0}\right] \quad \text{and}$$
$$\lim_{N\to\infty} w_1(p, q_1) = \frac{1}{\mu}\left[\frac{1}{1-\rho_1} + \frac{1}{p}\right], \tag{14}$$

which degenerate to delays in a vacation queue with Bernoulli schedule (Gao & Liu, 2013), where the system can be activated by each arriving customer with a certain probability p.

Using results derived so far, we can investigate the equilibrium joining-and-purchasing strategies. When all customers adopt strategy (p, q_0, q_1), the expected utilities of an arriving customer seeing an inactive server and an active server are

$$\widehat{U}_0(p, \rho_0) = R - \frac{C}{\mu}\left(1 + Q_N(p)\left[1 + \frac{1}{\rho_0}\right]\right) - pP \quad \text{and}$$
$$\widehat{U}_1(p, \rho_1) = R - \frac{C}{\mu}\left[\frac{1}{1-\rho_1} + 1 + Q_N(p)\right], \tag{15}$$

which are increasing and decreasing in λ_0 and λ_1, respectively. Then the ex ante expected utility is given by

$$\widehat{U}(p, q_0, q_1) = \frac{1-\rho_1}{1+\rho_0-\rho_1} \cdot \widehat{U}_0(p, \rho_0) + \frac{\rho_0}{1+\rho_0-\rho_1} \cdot \widehat{U}_1(p, \rho_1), \tag{16}$$

where $(1-\rho_1)/(1+\rho_0-\rho_1)$ and $\rho_0/(1+\rho_0-\rho_1)$ are the steady-state probabilities that the server is inactive and active, respectively.

Although customer arrivals arise from a homogeneous Poisson stream, we assign one of two "labels" to all arrivals immediately upon their arrivals; we do so according to the specific server state (busy working or on vacation) they observe. In particular, customers seeing an active server (i.e., $B(t) = 1$), thus assigned with a label "B_1," need to determine the joining probability q_1, while those finding the server on vacation ($B(t) = 0$), thus assigned with a label "B_0," will first determine their joining probability q_0 and next the PTAS purchasing probability p. To characterize customers' best response functions, we let $U_0((p, q_0); (p', q_0', q_1'))$ be the expected utility of a tagged B_0-customer who adopts (p, q_0), while assuming that all other B_0-customers adopt (p', q_0') and all B_1-customers adopt q_1'. Similarly, let $U_1(q_1; (p', q_0', q_1'))$ be the expected utility of an individual B_1-customer who adopts q_1, while assuming that all B_0-customers adopt (p', q_0') and all other B_1-customers adopt q_1'. Below we carefully define a symmetric Nash equilibrium.

Definition 1 (Symmetric Nash equilibrium). A strategy profile $(\boldsymbol{\alpha}^e, q_1^e)$ with $\boldsymbol{\alpha}^e \equiv (p^e, q_0^e)$ is a symmetric Nash equilibrium strategy if and only if

$$\boldsymbol{\alpha}^e \in \underset{\boldsymbol{\alpha}\in[0,1]\times[0,1]}{\arg\max}\ U_0(\boldsymbol{\alpha}; (\boldsymbol{\alpha}^e, q_1^e)) \quad \text{and}$$
$$q_1^e \in \underset{q_1\in[0,1]}{\arg\max}\ U_1(q_1; (\boldsymbol{\alpha}^e, q_1^e)). \tag{17}$$

Throughout the paper, we restrict our attention to symmetric Nash equilibrium. Similar definitions of state-dependent symmetric equilibria can be found in (3.3) and (3.4) of Wang and Wang (2019).

As will soon become clear in subsequent analysis, there often exist multiple equilibria. To identify those that are most relevant, we resort to notion of utility dominance.

Definition 2 (Pareto-dominant equilibrium). Given two equilibrium strategies $(\boldsymbol{\alpha}, q_1)$ and $(\boldsymbol{\alpha}', q_1')$, we say $(\boldsymbol{\alpha}, q_1)$ strictly dominates $(\boldsymbol{\alpha}', q_1')$ if $\widehat{U}(\boldsymbol{\alpha}, q_1) > \widehat{U}(\boldsymbol{\alpha}', q_1')$. An equilibrium $(\boldsymbol{\alpha}^*, q_1^*)$ is a Pareto-dominant equilibrium strategy if no other equilibrium strictly dominates it.

In what follows, we will adopt the notion of Pareto dominance to identify the most efficient equilibrium strategy among multiple equilibria (whenever exist) that maximizes the ex ante expected utility of customers. We will first characterize the equilibrium strategy for B_1-customers while assuming that the B_0-customer strategy is held fixed.

3.1.2 | Equilibrium strategy for B_1-customers

The following lemma guarantees the uniqueness of the equilibrium for B_1-customers for any given strategy of B_0-customers.

Proposition 2. *For any given strategy of B_0-customers $\boldsymbol{\alpha} = (p, q_0)$, the unique equilibrium strategy of B_1-customers is given by*

(1) *If $Q_N(p) > \mu R/C - 2$, then $q_1^e(p) = 0$.*
(2) *If $Q_N(p) \le \mu R/C - 2$, then*

$$q_1^e(p) = \begin{cases} 1, & \text{if } \rho \le \rho_l(p); \\ \dfrac{1}{\rho} - \dfrac{1}{\rho[\mu R/C - Q_N(p) - 1]}, & \text{if } \rho > \rho_l(p), \end{cases} \tag{18}$$

where $\rho_l(p) = 1 - \frac{1}{\mu R/C - Q_N(p) - 1}$ and $q_1^e(p)$ is independent of q_0.

Proposition 2 indicates that B_1-customers' equilibrium joining probability is independent of q_0 (only dependent on the PTAS purchasing probability). To see this, note that if the server is already active, there will be no future arrival of any B_0-customers as long as the queue has a positive content, so that B_0-customers' strategy has no bearing whatsoever on B_1-customers' joining behavior. In the subsequent subsections, we first develop the equilibrium strategy for B_0-customers with that of B_1-customers held fixed.

3.2 | Equilibrium strategy for B_0-customers

Because B_0-customers make two decisions, the characterization of their equilibrium strategy is less straightforward. We describe our roadmap in three steps. First, we derive the equilibrium joining probability q_0^e with any given p (Section 3.2.1); Next, we obtain the equilibrium PTAS purchasing strategy with all other strategies held fixed (Section 3.2.2); Last, we fully characterize the joint equilibrium strategy building on results in Section 3.1 and Section 3.2 (Section 3.3).

It should be noted that in the subsequent analysis, the Pareto-dominance criteria are no longer helpful in distinguishing two mixed equilibria because they can both induce a zero expected utility. In these cases, we will turn to the so-called *evolutionarily stable strategy* (ESS) (Hassin & Haviv, 2003) in face of multiple mixed equilibria.

Definition 3 (ESS). A two-dimensional equilibrium strategy $\boldsymbol{\alpha}$ is said to be an ESS if $U_0(\boldsymbol{\alpha}; \boldsymbol{\beta}) > U_0(\boldsymbol{\beta}; \boldsymbol{\beta})$ for all $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$.

ESS is useful in excluding the unstable mixed equilibria: If an equilibrium is stable, the system dynamics, when facing a small perturbation in customer behavior, is guaranteed to return to that equilibrium point. But this is not true for an unstable equilibrium. Following the steps to establish Proposition 2, we first obtain the best response of B_0-customers with a given $q_1 \in [0, 1]$, and we investigate it stability.

3.2.1 | Joining strategy

Proposition 3 (Equilibrium joining strategy with a fixed p). *Consider an unobservable M/M/1 vacation queue with PTAS. For a given PTAS purchasing probability $p \in [0, 1]$, when $\rho \le \rho_s(p)$, $q_0^e(p) = 0$; when $\rho > \rho_s(p)$, both $q_0^e(p) = 1$ and $q_0^e(p) = \frac{Q_N(p)}{\rho[\mu(R-pP)/C - Q_N(p) - 1]}$ are the equilibria, where $\rho_s(p) = \frac{Q_N(p)}{\mu(R-pP)/C - Q_N(p) - 1}$ and $q_0^e(p) = 1$ is the ESS.*

Remark 3 (Monotonicity in ρ and R).

(i) According to Propositions 2 and 3, q_0^e and q_1^e, the equilibrium joining probabilities for both B_0 and B_1 customers exhibit opposite monotonicity in ρ. We provide some intuitive explanations: When the server is inactive, more frequent arrivals reduce the server's vacation times to mitigate the queueing congestion, which encourages more customers to join the queue. On the other hand, if the server is already active, increasing the system's congestion level will incur a bigger waiting cost, leading to an increased number of customer balking.
(ii) For a given p, a bigger service reward R attracts more customers to join the queue, hence both $q_0^e(p)$ and $q_1^e(p)$ are increasing in R.

3.2.2 | PTAS purchasing strategy

In this subsection, we develop B_0-customers' equilibrium purchasing strategy with arrival rates λ_0 and λ_1 (or equivalently q_0 and q_1) held fixed. As will soon become clear in the next section, integration of results in Sections 3.2.1 and 3.2.2 will establish the joint join-and-purchase strategy for B_0-customers.

Note that when $N = 1$, $p = 0$ is a dominant strategy because it is never optimal for an arriving customer to accept PTAS. Hence, we hereby focus on the case $N \ge 2$. When all other customers adopt the strategy (p, q_0, q_1), consider a "tagged" customer who arrives and finds the system in state $(0, i)$ with $0 \le i < N$. Suppose she rejects PTAS, then let N_i be the number of additional future arrivals until the server becomes active. It is obvious that N_i is a geometric random variable with parameter p truncated at $N - i$, so

$$\mathbb{E}[N_i] = \sum_{j=1}^{N-i-1} \mathbb{P}[N_i \ge j] = \sum_{j=1}^{N-i-1} (1-p)^{j-1} = \frac{1-(1-p)^{N-i-1}}{p}. \tag{19}$$

If the tagged customer adopts PTAS with probability p', her expected delay is

$$W_i(p'; p) = (1-p') \cdot \left(\frac{1}{\lambda_0} \cdot \mathbb{E}[N_i] + \frac{i+1}{\mu} \right) + p' \cdot \frac{i+1}{\mu}. \tag{20}$$

When all other customers adopt strategy (p, q_0, q_1), by *Poisson arrivals see time averages* (PASTA), the system probabilities as observed by an arriving customer are identical to the steady-state probabilities, which are given in Proposition 1. By switching from rejecting PTAS to accepting PTAS, the tagged customer can reduce her expected delay cost by

$$\Delta W_N(p) \equiv C \sum_{i=0}^{N-2} \big[W_i(0;p) - W_i(1;p)\big]\pi_{0,i} = \frac{C\big[1-[1+p(N-1)](1-p)^{N-1}\big](1-\rho_1)}{\lambda_0 p[1-(1-p)^N](1+\rho_0-\rho_1)}. \quad (21)$$

Therefore, the best response of the tagged customer is to purchase PTAS if and only if $P \le \Delta W_N(p)$. Since the equilibrium strategies largely depend on the structural properties of $\Delta W_N(p)$, we next provide a careful analysis of $\Delta W_N(p)$.

Lemma 2 (Structural properties of $\Delta W_N(p)$).

(*i*) *For any given* $p \in [0,1]$, $\Delta W_N(p)$ *is increasing in* N *for* $N \ge 2$.

(*ii*) *When* $N = 2$, $\Delta W_N(p)$ *is increasing in* $p \in [0,1]$*; when* $N = 3, 4$, $\Delta W_N(p)$ *is unimodal in* $p \in [0,1]$*; when* $N \ge 5$, $\Delta W_N(p)$ *is decreasing in* $p \in [0,1]$.

Remark 4 (On structural properties of $\Delta W_N(p)$). Part (*i*) of Lemma 2 is intuitive because a higher threshold N is more difficult to reach, which makes PTAS more effective in activating the server and thus reducing the delay cost. Using Part (i) of Lemma 2, we can show that $\Delta W_N(p) > 0$ for any $N \ge 2$ and $p \in [0,1]$ when the system is stable (i.e., $\rho_1 \in [0,1)$), because $\Delta W_N(p) > \Delta W_2(p) = \frac{1-\rho_1}{\lambda_0(2-p)(1+\rho_0-\rho_1)} > 0$.

Part (*ii*) of Lemma 2 characterizes the impact of p on the customers' best responses. At a first glance, offering PTAS is similar to offering a "higher priority" to customers who are willing to pay. Nevertheless, unlike priority queueing models where customers' equilibrium strategy always exhibits FTC behavior (Hassin and Haviv 1997), Part (*ii*) of Lemma 2 infers the coexistence of both FTC and ATC behavior. We provide some intuitions in the following three cases:

- When N is small (e.g., $N = 2$), PTAS benefits only when a customer arrival finds the system in state (0,0) (because if the state is (0,1), the server is automatically activated). When more customers adopt PTAS (i.e., p is bigger), the server's vacation time is reduced so it is more likely for the tagged customer to find an empty system (i.e., $\pi_{0,0}$ increases). In this situation, the adoption of PTAS gives a bigger delay reduction for the tagged customer. Therefore, the equilibrium exhibits FTC behavior.
- When N is large (e.g., $N \ge 5$), if all other customers choose PTAS with a higher probability, there will be a bigger chance for the server to be activated by future customer arrivals, making it unnecessary for the tagged customer to pay for PTAS. Therefore, $\Delta W_N(p)$ decreases in p, indicating ATC behavior.
- When N is intermediate (e.g, $N = 3, 4$), the equilibrium exhibits both FTC and ATC behavior. If p is small, adopting PTAS is an efficient way to reduce the delay as the probability $\pi_{0,i}$ $(0 \le i \le N-2)$ increases in p; but if p is large, PTAS is already adopted by many other customers, it becomes less effective for the tagged customer to pay for PTAS.

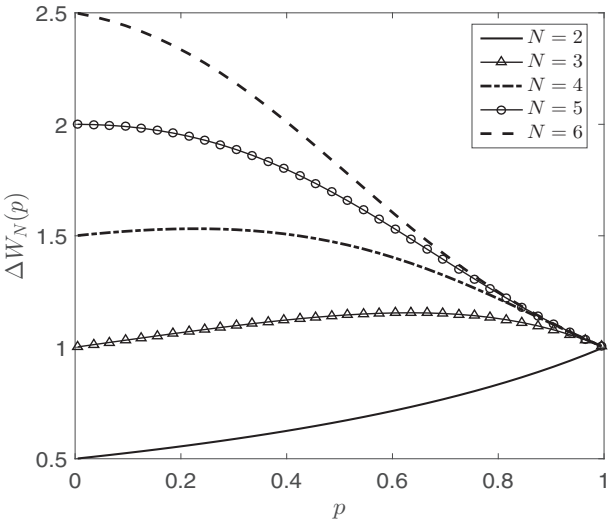


FIGURE 2 The function $\Delta W_N(p)$ with $0 \le p \le 1$, $2 \le N \le 6$, $\lambda_0 = \lambda_1 = 0.5$, and $C = \mu = 1$

See Figure 2 for a graphical illustration of these three cases.

Proposition 4 (Equilibrium purchasing strategy with exogenous arrival rates). *Consider an unobservable M/M/*1 *vacation queue with PTAS. For a given joining probability* (q_0, q_1)*, the equilibrium purchasing strategy* $p^e(q_0, q_1)$ *is given as follows:*

(*i*) *If* $N = 2$,

$$p^e(q_0,q_1) = \begin{cases} 0, & \text{if } P > \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}; \\ \underline{0}, \dfrac{2P\lambda_0(1+\rho_0-\rho_1)-C(1-\rho_1)}{P\lambda_0(1+\rho_0-\rho_1)}, \underline{1}, & \text{if } \dfrac{C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)} < P \le \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}; \\ 1, & \text{if } P \le \dfrac{C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}. \end{cases} \quad (22)$$

(*ii*) *If* $N = 3$,

$$p^e(q_0, q_1) = \begin{cases} 0, & \text{if } P > \dfrac{2\sqrt{3}C(1-\rho_1)}{3\lambda_0(1+\rho_0-\rho_1)}; \\ \underline{0}, p_3^{(1)}, \underline{p_3^{(2)}}, & \text{if } \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)} < P \le \dfrac{2\sqrt{3}C(1-\rho_1)}{3\lambda_0(1+\rho_0-\rho_1)}; \\ 1, & \text{if } P \le \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}. \end{cases} \tag{23}$$

(*iii*) *If* $N = 4$,

$$p^e(q_0, q_1) = \begin{cases} 0, & \text{if } P > \Delta W_N(\hat{p}_4); \\ \underline{0}, p_4^{(1)}, \underline{p_4^{(2)}}, & \text{if } \dfrac{3C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)} < P \le \Delta W_N(\hat{p}_4); \\ p_4, & \text{if } \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)} < P \le \dfrac{3C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}; \\ 1, & \text{if } P < \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}. \end{cases} \tag{24}$$

(*iv*) *If* $N \ge 5$,

$$p^e(q_0, q_1) = \begin{cases} 0, & \text{if } P > \dfrac{C(N-1)(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}; \\ p_N, & \text{if } \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)} < P \le \dfrac{C(N-1)(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}; \\ 1, & \text{if } P \le \dfrac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}, \end{cases} \tag{25}$$

where $p_3^{(1)} = \frac{3}{2} - \frac{(1-\rho_1)C + \sqrt{[(1-\rho_1)C]^2 - \frac{3(\lambda_0(1+\rho_0-\rho_1)P)^2}{4}}}{\lambda_0(1+\rho_0-\rho_1)P}$, $p_3^{(2)} = \frac{3}{2} - \frac{(1-\rho_1)C - \sqrt{[(1-\rho_1)C]^2 - \frac{3(\lambda_0(1+\rho_0-\rho_1)P)^2}{4}}}{\lambda_0(1+\rho_0-\rho_1)P}$, $p_4^{(2)}$ *is the solution of* $\Delta W_N(p) = P$ *in* $p \in (\hat{p}_4, 1)$, $\hat{p}_4 \in (0,1)$ *uniquely solves* $4 - 24p + 32p^2 - 16p^3 + 3p^4 = 0$, *and* p_N *uniquely solves* $\Delta W_N(p) = P$ *when* $\frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)} < P \le \frac{C(N-1)(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}$ *for* $N \ge 4$. *In the presence of multiple equilibria, the ESS is underlined.*

Remark 5 (PTAS vs. priority: mixed ESS). First, it is the FTC behavior that gives rise to multiple equilibria (as in Cases (*i*)–(*iii*)). In this sense, the structure appears to be somewhat similar to the priority-purchasing strategy in priority queueing models. Nevertheless, a major distinction here is that a mixed strategy, which is never an ESS in priority queues (Hassin & Haviv, 1997), can in fact be an ESS in the present PTAS model (Cases (*ii*) and (*iii*) in Proposition 4 when $N = 3, 4$). Such a result is due to the coexistence of both FTC and ATC; also see Remark 4 and Figure 2.

3.3 | Joint equilibrium strategy

We are now ready to derive the joint equilibrium of B_0 and B_1 customers. We denote by the triplet $\mathcal{E} = (p^e, q_0^e, q_1^e)$ the joint joining-and-purchasing equilibrium strategy. By Definition 1 and Propositions 2–4, a strategy (p, q_0, q_1) is an equilibrium if and only if it satisfies

$$p \in p^e(q_0, q_1), \quad q_0 \in q_0^e(p), \quad q_1 \in q_1^e(p), \tag{26}$$

where $q_1^e(p)$, $q_0^e(p)$, and $p^e(q_0, q_1)$ are identified in Propositions 2, 3, and 4, respectively. For $p, q_1 \in [0,1]$, we call $\mathbf{0}_e \equiv (p, 0, q_1)$ the "zero" equilibrium strategy, under which no customer will join the system (because $q_0^e = 0$ means that no B_0-customer joins for service, the system will never be activated regardless of the values of p^e and q_1^e).

Characterizing the explicit form of equilibrium is challenging because (1) the joining-and-purchasing behavior depends on the values of all model parameters (such as the PTAS fee P and traffic intensity ρ) and (2) multiple equilibria can exist. We first focus on the case $5 \le N \le \mu R/C - 1$. This case is relatively straightforward because the customer utility is a monotone function when $N \ge 5$, which warrants the uniqueness of equilibrium, and the condition $N \le \mu R/C - 1$ ensures an active service even without PTAS.[5]

Theorem 1 (Joint joining-and-purchasing equilibrium strategy). *Consider the unobservable M/M/1 vacation queue with PTAS. Assume* $5 \le N \le \mu R/C - 1$, *the joint equilibrium strategy is given below:*

(*i*) *If* $\bar{P}_1 \le \underline{P}_2$, $\mathcal{E}$ *is given in the table below*

ρ	$(0,\rho_s(\tilde{p})]$	$(\rho_s(\tilde{p}),\rho_s(0)]$	$(\rho_s(0),\rho_l(0)]$	$(\rho_l(0),\rho_l(\tilde{p})]$	$(\rho_l(\tilde{p}),\rho_l(\bar{p})]$	$(\rho_l(\bar{p}),\rho_l(1)]$	$(\rho_l(1),\infty)$
$P \le \underline{P}_1$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,q_{11})$
$\underline{P}_1 < P \le \bar{P}_1$	$\mathbf{0}_e$	$(\tilde{p},1,1)$	$(\tilde{p},1,1)$	$(\tilde{p},1,1)$	(p',q_{01},q_{14})	(p',q_{01},q_{14})	$(1,1,q_{11})$
$\bar{P}_1 < P \le \underline{P}_2$	$\mathbf{0}_e$	$\mathbf{0}_e$	$(0,1,1)$	(p',q_{01},q_{14})	(p',q_{01},q_{14})	(p',q_{01},q_{14})	$(1,1,q_{11})$
$\underline{P}_2 < P \le \bar{P}_2$	$\mathbf{0}_e$	$\mathbf{0}_e$	$(0,1,1)$	(p',q_{01},q_{14})	(p',q_{01},q_{14})	$(\bar{p},1,q_{12})$	$(\bar{p},1,q_{12})$
$P > \bar{P}_2$	$\mathbf{0}_e$	$\mathbf{0}_e$	$(0,1,1)$	$(0,1,q_{13})$	$(0,1,q_{13})$	$(0,1,q_{13})$	$(0,1,q_{13})$

(*ii*) *If* $\bar{P}_1 > \underline{P}_2$, $\mathcal{E}$ *is given in the table below*

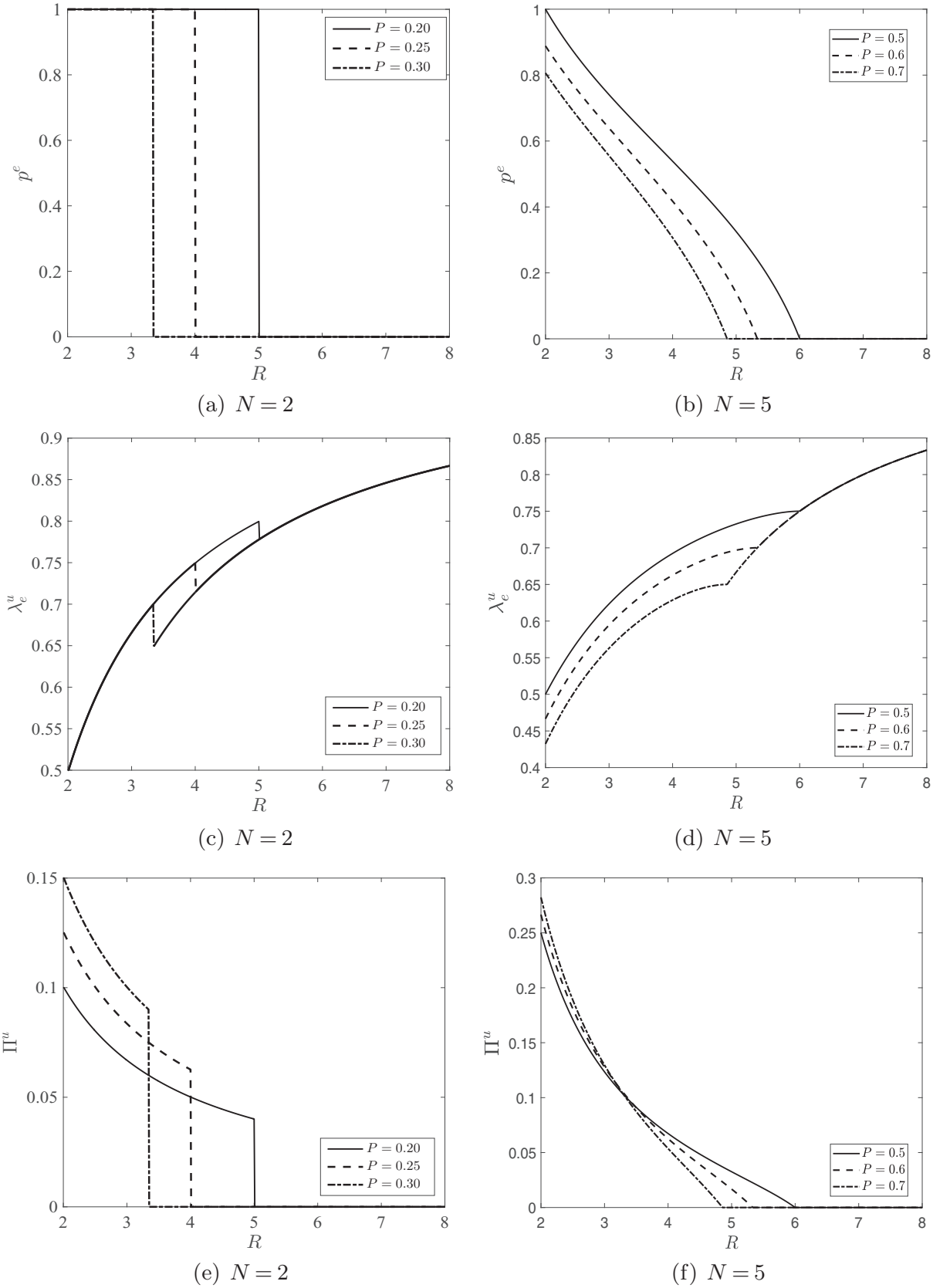
ρ	$(0,\rho_s(\tilde{p})]$	$(\rho_s(\tilde{p}),\rho_s(0)]$	$(\rho_s(0),\rho_l(0)]$	$(\rho_l(0),\rho_l(\tilde{p})]$	$(\rho_l(\tilde{p}),\rho_l(\bar{p})]$	$(\rho_l(\bar{p}),\rho_l(1)]$	$(\rho_l(1),\infty)$
$P \le \underline{P}_1$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,1)$	$(1,1,q_{11})$
$\underline{P}_1 < P \le \underline{P}_2$	$\mathbf{0}_e$	$(\tilde{p},1,1)$	$(\tilde{p},1,1)$	$(\tilde{p},1,1)$	(p',q_{01},q_{14})	(p',q_{01},q_{14})	$(1,1,q_{11})$
$\underline{P}_2 < P \le \bar{P}_1$	$\mathbf{0}_e$	$(\tilde{p},1,1)$	$(\tilde{p},1,1)$	$(\tilde{p},1,1)$	(p',q_{01},q_{14})	$(\bar{p},1,q_{12})$	$(\bar{p},1,q_{12})$
$\bar{P}_1 < P \le \bar{P}_2$	$\mathbf{0}_e$	$\mathbf{0}_e$	$(0,1,1)$	(p',q_{01},q_{14})	(p',q_{01},q_{14})	$(\bar{p},1,q_{12})$	$(\bar{p},1,q_{12})$
$P > \bar{P}_2$	$\mathbf{0}_e$	$\mathbf{0}_e$	$(0,1,1)$	$(0,1,q_{13})$	$(0,1,q_{13})$	$(0,1,q_{13})$	$(0,1,q_{13})$

where all relevant parameters are given by (S33)–(S35) in the Supporting Information.

In Theorem 1, all nonzero equilibria can be classified into three categories: $p^e = 0$, $p^e \in (0,1)$, and $p^e = 1$, which correspond to the three regions in the tables. As P increases, the probability to adopt PTAS decreases, and eventually $p^e = 0$ when P is large enough, resulting in a regular vacation queue operated under the N-policy. When P is small, it is never optimal for an arriving customer to wait for the future customers to activate the server ($p^e = 1$); this case reduces to a standard M/M/1 queue. When P is large, PTAS is almost never in effect so the only way to activate the service is by accumulating enough waiting customers. However, if in addition, ρ is small (so interarrival times are long), it will take a long time for the server's vacation to end. As a result, the overall delay cost becomes too high so that no one is willing to join the queue. See the lower left part of the tables in Theorem 1 with $\mathcal{E} = \mathbf{0}_e$.

Careful partition of the parameter space is less straightforward, and multiple equilibria may coexist due to the nonmonotonicity of $\Delta W_N(p)$. Following the criterion in Definition 2, we can identify the Pareto-dominant equilibrium by comparing customers' ex ante expected utilities (see (16)) under different equilibrium strategies. For example, when $N = 2$, $R = 4$, $P = 0.2$, and $\Lambda = \mu = C = 1$, $\mathcal{E}_1 = (0,1,0.6)$, and $\mathcal{E}_2 = (1,1,2/3)$ are both equilibrium strategies, of which the ex ante expected utilities are $\widehat{U}(0,1,0.6) = 0.569$ and $\widehat{U}(1,1,2/3) = 0.7$. Since $\widehat{U}(0,1,0.6) < \widehat{U}(1,1,2/3)$, $\mathcal{E}_1$ is strictly dominated by $\mathcal{E}_2$, it follows that $\mathcal{E}_2$ is the unique Pareto-dominant equilibrium.

We next conduct a numerical example with $\Lambda = \mu = C = 1$, $N = 2,5$, $0 \le R \le 8$, and $0.1 \le P \le 0.7$ to investigate the impact of service reward on the equilibrium outcomes. In Figure 3, we graph the equilibrium purchasing probability p^e, throughput λ_e^u, and PTAS revenue Π^u. We summarize our observations: First, both p^e and λ_e^u are (weakly) decreasing in P; as P increases, fewer customers adopt PTAS (i.e., $p^e \downarrow$). Consequently, the server's vacation time increases, discouraging future customers from joining the queue (i.e., $\lambda_e^u \downarrow$). Next, plots a and b of Figure 3 show that p^e decreases in R. To see this, we point out that, when R is small, the equilibrium effective arrival rate is small so the queue size almost never reaches the threshold N. Hence, the PTAS purchasing probability needs to increase in order to achieve an acceptable delay. On the other hand, a bigger R leads to a bigger effective arrival rate, which in turn reduces the individual PTAS purchasing probability. Hence, the PTAS purchasing probability is nonincreasing in R. This result stands in sharp contrast to the equilibrium strategy in priority queues where customers are more inclined to purchase priority when R increases (Wang et al., 2019). Furthermore, we observe from panel c of Figure 3 that the throughput λ_e^u is not always increasing in the service reward R. According to Remark 5, there exist multiple equilibrium purchasing probabilities when $N < 5$, thus the Pareto-dominant purchasing probability p^e may shift from one equilibrium to another when R varies (this explains why p^e is discontinuous in R as shown in panel a of Figure 3). In particular, when R is neither too large nor too small, p^e drops from a positive value to 0 as R increases, leading to surged vacation times, and hence a sharp fall in the throughput. However, when R is sufficiently small or large, the equilibrium PTAS purchasing probability p^e is unique and remains continuous in R. In fact, increasing R leads to two opposite effects: On the one hand, it reduces the PTAS purchasing probability p^e, and on the other hand, it attracts more customers to join the queue. In this case, the latter effect outweighs the former, so the throughput λ_e^u is increasing in R.



(a) $N = 2$ (b) $N = 5$

(c) $N = 2$ (d) $N = 5$

(e) $N = 2$ (f) $N = 5$

FIGURE 3 The unobservable case: equilibrium purchasing probability p^e, throughput λ_e^u, and revenue Π^u, with $\Lambda = \mu = C = 1$, $N = 2, 5$, $2 \le R \le 8$, $0.2 \le P \le 0.7$

4 | OBSERVABLE QUEUE

In this section, we investigate the M/M/1 vacation queue with PTAS when the real-time queue length is revealed to all arriving customers. Unlike the unobservable case where customers join the queue with a probability (independent with the queue length), their joining-and-purchasing decisions are now based on the real-time system state, see Naor (1969). When a customer (if joining) is indifferent between accepting and rejecting PTAS upon arrival, we assume for simplicity that she will choose to pay for PTAS. A similar assumption can be found in observable priority queues, see Wang et al. (2021b). Next, we characterize the equilibrium strategy, and then compute the system performance in equilibrium.

Since state-dependent decisions are made in the observable case, we consider this model with infinitely many decision makers (customers), each facing a state sampled from the state-space $\mathcal{S} = \{(0,n) : 0 \le n \le N-1\} \cup \{(1,n) : n \ge 1\}$. Each system state $s \in \mathcal{S}$ is associated with a set of actions $A(s)$ such that

$$A(s) = \begin{cases} \{J_0, B\}, & \text{if } s = (1,j) \text{ for } j \ge 1; \\ \{J_0, J_1, B\}, & \text{if } s = (0,j) \text{ for } 0 \le j \le N-1, \end{cases} \tag{27}$$

where B denotes "balking," and J_0 and J_1 denote "joining without purchasing PTAS" and "joining and purchasing PTAS," respectively. Then a pure strategy δ, specifies an action, $\delta(s) \in A(s)$, for every $s \in \mathcal{S}$. Using the preceding notations, we next give the formal definition of SPE strategy; also see Fudenberg and Tirole (1991) and Hassin and Haviv (2002) for discussions of SPE.

Definition 4 (SPE). A strategy δ_e is an SPE if $a \in \arg\max_{a \in A(s)} U_s(a; \delta_e)$ and $a = \delta_e(s) \in A(s)$ for every $s \in \mathcal{S}$.

To characterize an SPE, one needs to consider all the system states in $\mathcal{S}$; specifically, customers' best responses need to be determined in all scenarios (including transient states). SPE is especially useful in describing the dynamic evolution of the system, for example, in the case that a customer, due to some reason, deviates from her optimal strategy. It should be noted that an SPE specifies the best response actions in all states, even though not all of them can occur in the equilibrium. Hence, an SPE strategy must be an equilibrium strategy, but the inverse statement is not necessarily correct (Hassin & Haviv, 2002). To characterize the equilibrium strategy, we first need to identify customers' best responses to every system state, and then refine the unique SPE on the equilibrium path.

4.1 | Equilibrium analysis

Suppose a tagged customer arrives and finds an *active* server, along with n existing customers waiting in line (excluding herself), that is, $s = (1,n)$, for any strategy δ, her expected utility is

$$U_s(a; \delta) = \begin{cases} R - \dfrac{(n+1)C}{\mu}, & \text{if } a(s) = J_0; \\ 0, & \text{if } a(s) = B. \end{cases} \tag{28}$$

Hence, the tagged customer will join the system if and only if $R - (n+1)C/\mu \ge 0$, or equivalently, $n < \lfloor \mu R/C \rfloor$,[6] which gives

$$\delta_e(1,n) = \begin{cases} J_0, & \text{if } n \le \lfloor \mu R/C \rfloor; \\ B, & \text{if } n > \lfloor \mu R/C \rfloor. \end{cases} \tag{29}$$

Therefore, it remains to characterize the equilibrium strategy of a customer finding the server to be on vacation (i.e., $s = (0,n)$ for $0 \le n \le N-1$), which is what we shall do in the rest of this subsection. We consider two cases: (1) $P \le C/\Lambda$ (low PTAS fee) and (2) $P > C/\Lambda$ (high PTAS fee).

Suppose the server is *inactive* when the tagged customer arrives. When $P \le C/\Lambda$, that is, the PTAS fee is lower than the expected cost spent waiting for the next arrival (who may or may not be able to activate the server), then it is optimal for the customer to adopt PTAS (i.e., it is not worthy to wait for even a single arrival), provided that she decides to join the queue.

If, in addition $R \ge P + C/\mu$, then the tagged customer must join the system because doing so guarantees a nonnegative utility. Hence, under the two conditions $P \le C/\Lambda$ and $R \ge P + C/\mu$, the system reduces to a regular M/M/1 model (the server will be activated by the very first arriving customer in equilibrium). On the other hand, if $R < P + C/\mu$, joining and adopting PTAS will induce a negative utility so that it is optimal for all arrivals to balk (the system is never active). Below we formally describe the equilibrium strategy when $P \le C/\Lambda$.

Theorem 2 (Observable queue with low PTAS fee). *Consider an observable M/M/1 vacation queue with $P \le C/\Lambda$, the SPE on the equilibrium path is given as follows:*

- *High reward: If $R \ge P + C/\mu$, then $\delta_e(0,0) = J_1$ and*

$$\delta_e(1,n) = \begin{cases} J_0, & \text{if } 1 \le n \le \lfloor \mu R/C \rfloor - 1; \\ B, & \text{if } n \ge \lfloor \mu R/C \rfloor. \end{cases} \tag{30}$$

- *Low reward: If $R < P + C/\mu$, then $\delta_e(0,n) = B$ for $n \in \mathbb{N}$.*

According to Theorem 2, when the PTAS fee is sufficiently small, any joining customer purchases PTAS (if seeing an inactive server) so the system will be activated by the first arriving customer in equilibrium. Besides, the system reduces to a standard work-conservation queue in which customers join if and only if the queue length is below some threshold.

When the PTAS fee is higher than the cost of waiting for one future arrival, the tagged customer's best response has to take into account the behavior of future arrivals. Note that it may be worthwhile to wait for one arrival, but there is no guarantee that she will activate the server for sure. Let $I \equiv \min\{n : nC/\Lambda > P\} = \lceil \Lambda P/C \rceil$ be the minimum number of future arrivals a tagged customer awaits until her cumulative waiting cost exceeds P. We next present our results for the case $P > C/\Lambda$ (where I plays a critical role).

Theorem 3 (Observable queue with high PTAS fee). *Consider an observable M/M/1 vacation queue with $P > C/\Lambda$, the SPE on the equilibrium path is given as follows:*

$$\delta_e(0,n) = \begin{cases} J_0, & \text{if } n \le \bmod(\bar{n}, I) - 1; \\ J_1, & \text{if } \bmod(\bar{n}, I) = n. \end{cases}$$

$$\delta_e(1,n) = \begin{cases} J_0, & \text{if } \bmod(\bar{n}, I) + 1 \le n \le \lfloor \mu R/C \rfloor - 1; \\ B, & \text{if } \lfloor \mu R/C \rfloor \le n, \end{cases} \tag{31}$$

where $\bar{n} = N - 1$ if $R \ge \max\{u(i)\}$ and $\bar{n} = \lfloor (R-P)\mu/C \rfloor - 1$ otherwise; and

$$u(i) = \begin{cases} \dfrac{(i+1)C}{\mu} + P, & \text{if } \bmod(N-1-i, I) = 0; \\ \dfrac{(i+1)C}{\mu} + \dfrac{\bmod(N-1-i, I)C}{\Lambda}, & \text{if } \bmod(N-1-i, I) > 0 \end{cases} \tag{32}$$

for $i \in \{0, 1, \ldots, N-2\}$.

Remark 6. According to Theorem 3, the SPE on the equilibrium path is of a threshold type, that is, when the server is on vacation, arriving customers will join without purchasing PTAS until the queue size reaches a threshold $\bmod(\bar{n}, I)$, by that time the system will be activated by an arriving customer via PTAS. Afterwards, it remains active, and all future arrivals follow the standard Naor threshold strategy (see (29)). Unlike the standard vacation queues where the service-resumption threshold is exogenous, the threshold of the PTAS queue is dependent on the system parameters. We provide additional explanations regarding the structure of the threshold $\bmod(\bar{n}, I)$ in Section D of the Supporting Information.

4.2 | System performance

Next, we derive the system performance under the equilibrium strategy in the observable case. In particular, we compute the system throughput and the PTAS revenue, which is the rate at which customers pay for PTAS, multiplied by the PTAS fee P.

Theorem 4 (Throughput and PTAS revenue). *Consider an observable M/M/1 vacation queue with PTAS, the steady-state probabilities, throughput, and revenue are given below.*

(*i*) *If $P \le C/\Lambda$, the steady-state probabilities are*

$$\pi_{0,0} = \frac{1-\rho}{1-\rho^{n_1+1}}, \quad \pi_{1,i} = \frac{(1-\rho)\rho^i}{1-\rho^{n_1+1}}, \quad i = 1, 2, \ldots, n_1, \tag{33}$$

where $n_1 = \lfloor \mu R/C \rfloor$. The system throughput and PTAS revenue are given by

$$\lambda_e^o = \frac{\Lambda(1-\rho^{n_1})}{1-\rho^{n_1+1}} \quad \text{and} \quad \Pi^o = \frac{\Lambda P(1-\rho)}{1-\rho^{n_1+1}}. \tag{34}$$

(*ii*) *If $P > C/\Lambda$, the steady-state probabilities are*

$$\begin{aligned} \pi_{0,i} &= \frac{(1-\rho)^2}{(1-\rho)(n_2+1) + \rho^{n_1+1}(\rho - \rho^{-n_2})}, \\ \pi_{1,k+1} &= \frac{\rho(1-\rho)(1-\rho^{k+1})}{(1-\rho)(n_2+1) + \rho^{n_1+1}(\rho - \rho^{-n_2})}, \\ \pi_{1,j} &= \frac{\rho^{j-n_2}(1-\rho)(1-\rho^{n_2+1})}{(1-\rho)(n_2+1) + \rho^{n_1+1}(\rho - \rho^{-n_2})} \end{aligned} \tag{35}$$

for $i = 0, 1, \ldots, n_2$, $k = 0, 1, \ldots, n_2$, and $j = n_2+2, n_2+3, \ldots, n_1$, where $n_2 = \bmod(\bar{n}, I)$.
The system throughput and PTAS revenue are given by

$$\lambda_e^o = \frac{\Lambda[(1-\rho)(n_2+1) + \rho^{n_1}(\rho - \rho^{-n_2})]}{(1-\rho)(n_2+1) + \rho^{n_1+1}(\rho - \rho^{-n_2})} \quad \text{and}$$

$$\Pi^o = \frac{\Lambda P(1-\rho)^2}{(1-\rho)(n_2+1) + \rho^{n_1+1}(\rho - \rho^{-n_2})}. \tag{36}$$

Remark 7 (Queueing dynamics in equilibrium). The server, whenever on vacation, will be activated as soon as the queue length reaches a certain level. Unlike standard N-policy vacation queues having a designated threshold N, the activating threshold of our PTAS queue depends on several model parameters (i.e., R, P, μ, C, and Λ). Specifically, when P is small (Case (i)), it is optimal to purchase PTAS whenever the server is on vacation, and customers will join as long as the queue length is less than the Naor threshold $\lfloor \mu R/C \rfloor$. And the model reduces to the Naor model (Naor, 1969). When P is large (Case (ii)), the server remains inactive until an arriving customer finds $n_2 = \bmod(\bar{n}, I)$ existing customers in the queue, so the model reduces to an N-policy vacation queue with $N = n_2$.

We consider a numerical example to visualize results in Theorem 4. In Figure 4 we plot the throughput λ_e^o and PTAS revenue Π^o for $N = 2, 5, 10$. Intuitively, a bigger R drives more customers to join the system (so a bigger λ_e^o), making it less necessary for customers to adopt PTAS (so a smaller Π^o). However, Figure 4 indicates that neither λ_e^o nor Π^o is monotone in the service reward. Unlike regular vacation queues where the threshold N is independent of R, the equilibrium threshold of our PTAS queue is a function of the service reward, in particular, the equilibrium threshold $n_2(R) = \text{mod}(\lfloor (R-P)\mu/C \rfloor - 1, I)$, which itself is not monotone in R. This explains the cyclic "up-and-down" behavior of λ_e^o and Π^o (see cases $N = 5$ and $N = 10$). By contrast, when $N = 2$, n_2 does not vary much ($n_2 = 1$ or 2), so both λ_e^o and Π^o are monotone in R.

5 | COMPARISONS AND IMPLICATIONS

In this section, we first compare the system performance (e.g., throughput and PTAS revenue) and pricing implications under two information disclosure policies. Next, we benchmark the performance of our PTAS vacation model to that of a regular vacation queue without PTAS. Finally, we study how our PTAS model distinguishes from the pay-for-priority queues.

5.1 | Impact of service reward

Theorem 5. *For any fixed PTAS fee P, there exists a threshold $\underline{R}$ for the service reward such that $\Pi^u > \Pi^o$ if $R < \underline{R}$.*

In the observable case, the expected customer utility depends on their queueing positions. So a smaller service reward R discourages more customers from joining the queue, leading to a smaller throughput. In contrast, hiding the queueing position becomes an advantage in an unobservable queue because customers make their joining decisions based on the average queue length. As a result, the unobservable setting yields a higher revenue.

In Figure 5, we plot the PTAS revenue under two information policies as a function of the service reward. Consistent with Theorem 5, Figure 5 shows that, when R is small, a higher revenue can be achieved by hiding the queue-length information; on the other hand, when R is large, more customers join the system in the unobservable case, making it less necessary to purchase PTAS (hence a lower PTAS revenue).

5.2 | Impact of congestion level

Theorem 6. *For any fixed PTAS fee P, there exists a threshold $\underline{\Lambda}$ for the congestion level Λ such that $\lambda_e^u > \lambda_e^o$ if $\Lambda < \underline{\Lambda}$.*

Results in Theorem 6 are consistent with the general consensus: When the potential arrival rate is sufficiently small, all customers in the unobservable model join the system; but in the observable case, balking can still happen when customers observe a longer queue upon arrival. Next, we proceed to compare the system performance measures under two information levels relative to the case without PTAS. Note that the server in a standard vacation queue can never be activated if $R < CN/\mu$. To avoid triviality, we focus on the case $R \geq CN/\mu$ in the rest of this section. Let $\Pi^u(\Lambda)$ ($\Pi^o(\Lambda)$) be the maximum revenue collected from PTAS with demand volume Λ in the unobservable (observable) case, the following result reveals the impact of the congestion level on the system revenue.

Theorem 7. *Under both information policies, the PTAS revenue is nonmonotonic in the congestion level Λ, with $\Pi^o(0) = \Pi^o(\infty) = \Pi^u(0) = \Pi^u(\infty) = 0$.*

At a quick look, the fact that the PTAS revenue is not monotonically increasing in the potential demand seems to counter the conventional wisdom. In fact, the market size Λ impacts the revenue in two opposite directions. On the one hand, increasing Λ helps create bigger customer demand for purchasing PTAS; on the other hand, when Λ is sufficiently large, the high congestion level almost always warrants an active server, which impedes customers from paying for PTAS. When Λ is small, increasing the demand size yields a higher revenue because the first effect dominates. However, when Λ is already large enough, the server hardly has any vacation time, so the second effect prevails.

In the unobservable (observable) case, the revenue reaches its peak at some finite Λ^u (Λ^o). Let $P^u(\Lambda)$ ($P^o(\Lambda)$) be the optimal PTAS fee in the unobservable (observable) case with a demand volume Λ. We have $P^x(\Lambda) \in [0, R - C/\mu]$ for $x = u, o$, otherwise no one will ever purchase PTAS. The following theorem compares the optimal PTAS fees under the two information policies.

Theorem 8 (Optimal PTAS fee: observable queue vs. unobservable queue). *The optimal prices satisfy $P^u(0) = R - C/\mu$ and $P^u(\infty) = 0$. Furthermore, there exists a threshold $\bar{\Lambda}$ such that $P^o(\Lambda) > P^u(\Lambda)$ if $\Lambda > \bar{\Lambda}$.*

When the demand volume is sufficiently low, the only way to activate the server is through purchasing PTAS (because the queue length almost never reaches N). This motivates the service provider to set an increase in the PTAS fee to gain improved revenue. We next discuss the case of high-demand volume (with a large Λ). In the unobservable model, customers anticipate a low expected delay (because the average system size should not be far below N), so most customers are reluctant to purchase PTAS. As a result, the service provider needs to lower the PTAS fee in order to achieve improved revenue. In the observable case, customers who observe a shorter queue length (due to the stochastic nature of the queueing system) will likely use PTAS to mitigate their waiting costs. This should explain why the observable model has a higher PTAS fee.

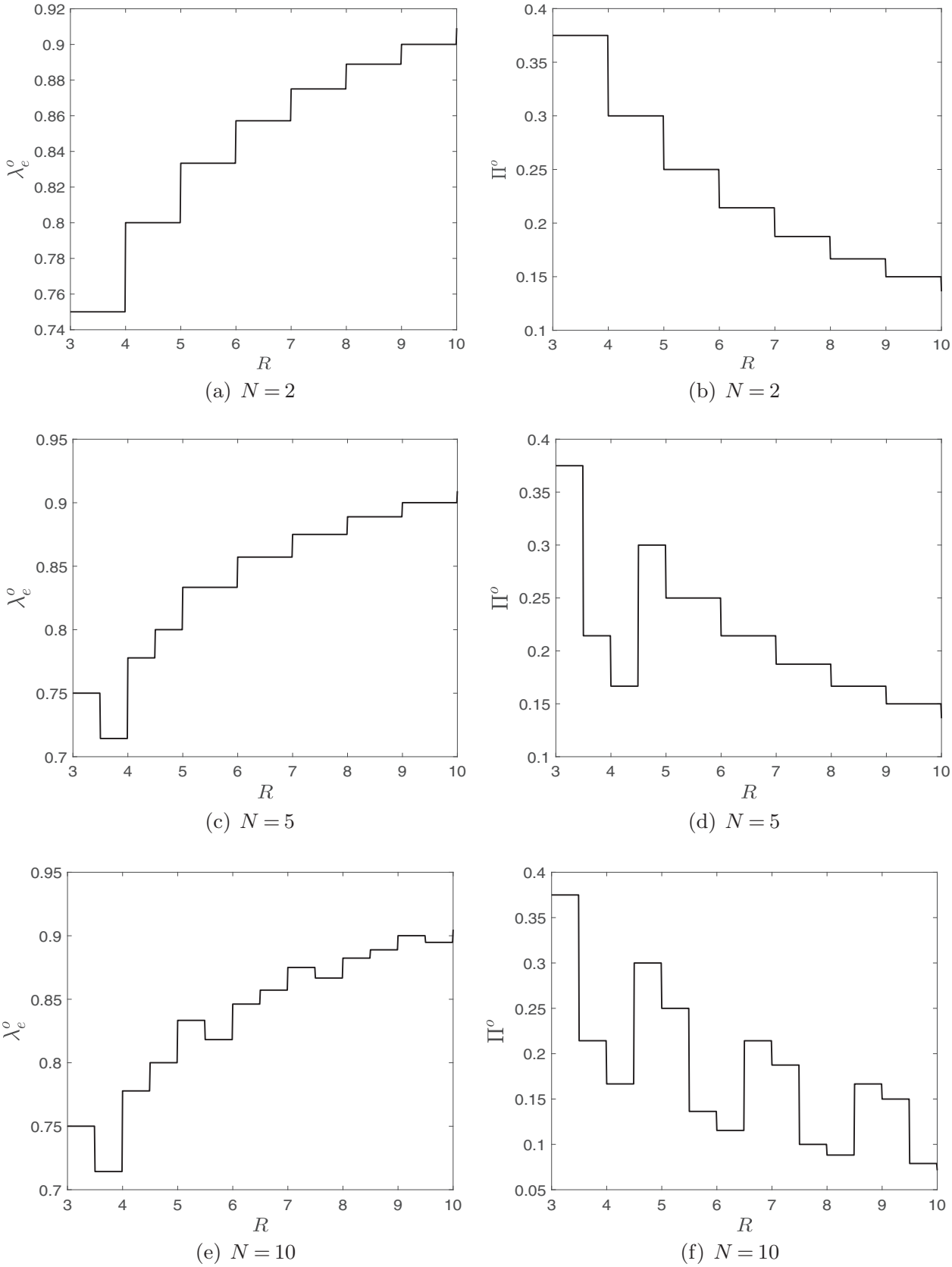
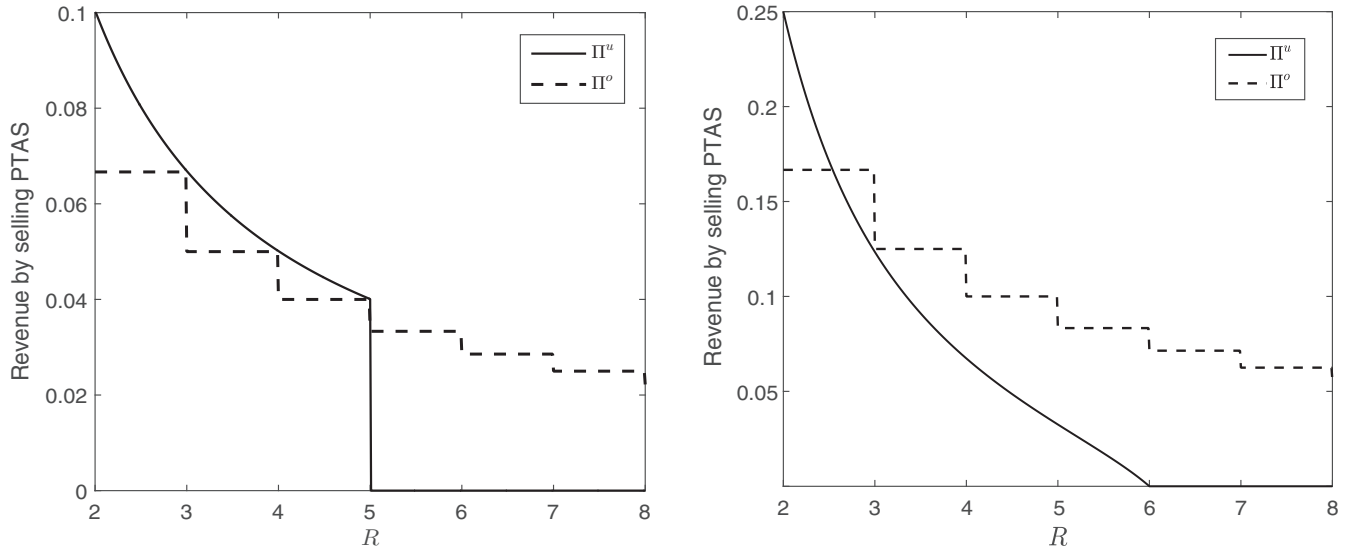
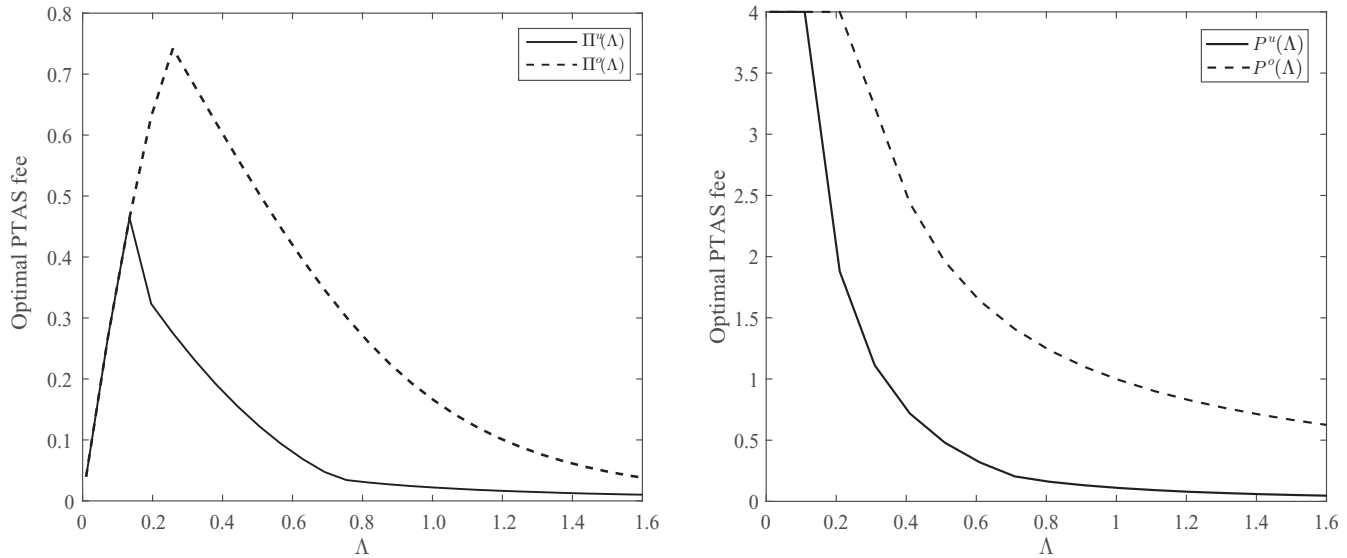


FIGURE 4 Throughput and PTAS revenue in the observable vacation queue for different R, with $\Lambda = \mu = C = 1$ and $P = 1.5$



(a) $N = 2, P = 0.2$ (b) $N = 5, P = 0.5$

FIGURE 5 Comparison of revenue and system throughput under two information structures for different R, with $\Lambda = \mu = C = 1$



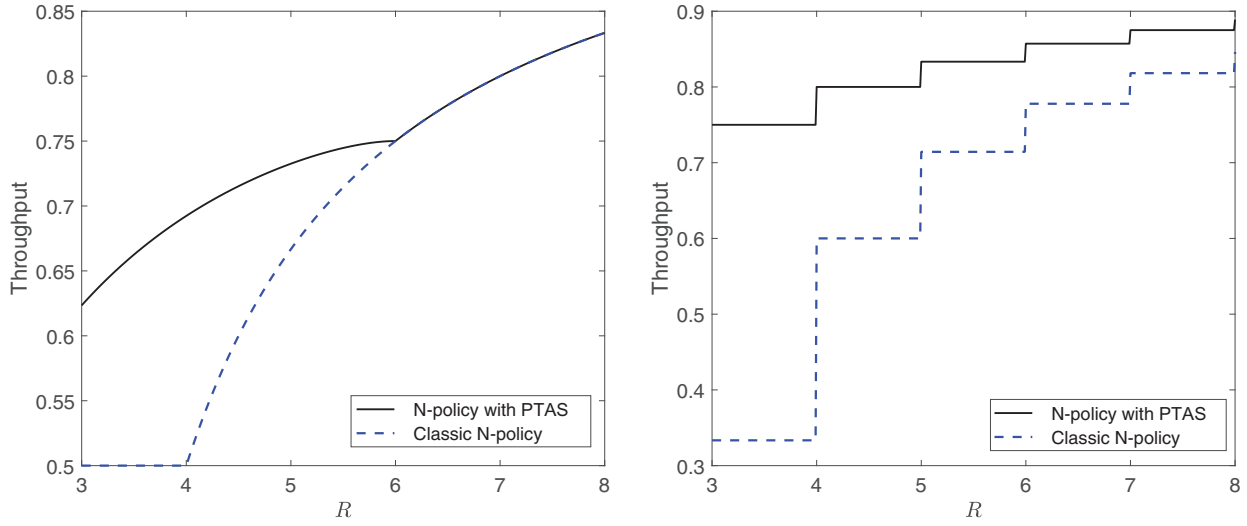
(a) PTAS revenue (b) PTAS fee

FIGURE 6 Comparisons of optimal PTAS revenue and corresponding PTAS fee under two information structures for different Λ, with $N = 2$, $R = 5$, and $\mu = C = 1$

In Figure 6 we use a numerical example to illustrate the PTAS revenue (left panel) and optimal PTAS fee (right panel) under the two information policies. Consistent with results in Theorem 8, Figure 6 shows that a higher PTAS fee should be set in the observable case. In addition, the PTAS revenue has a unimodal form in the demand volume, and the optimal PTAS fee is weakly decreasing in Λ (with $P^o \geq P^u$).

5.3 | Advantage of PTAS in vacation queues

In this subsection, we investigate how the PTAS mechanism benefits vacation queues. In particular, we provide a comparison of throughput in two models: an M/M/1 vacation queue with PTAS and an N-policy M/M/1 vacation queue without PTAS.



(a) Unobservable case (b) Observable case

FIGURE 7 Comparing the system's throughput functions in unobservable vacation queues with and without PTAS, with $\Lambda = \mu = C = 1$, $N = 5$, and $P = 0.5$ [Color figure can be viewed at wileyonlinelibrary.com]

Theorem 9 (PTAS improves throughput). *PTAS achieves improved system throughput for the M/M/1 vacation queue in both the observable and unobservable cases.*

PTAS improves the system throughput by allowing customers to activate service immediately upon their arrivals rather than awaiting future arrivals to increment the queue length to level N. Indeed, a customer adopting PTAS can not only reduce her own waiting time, it can also mitigate the delay cost for other customers (those present in the system and those yet to arrive). These effects work collectively to improve the system throughput.

In support of Theorem 9, we give a numerical example to compare the system throughput for vacation models with and without PTAS for different service reward R (see Figure 7a for the unobservable case and Figure 7b for the observable case). These results not only confirm that PTAS is useful in improving the system throughput, they also reveal some additional insights: PTAS is especially effective as long as the service reward R is relatively small. A bigger service reward attracts more customers to join for service so that queue size N is easily attained, which impedes joining customers from purchasing PTAS (Figure 7a confirms that, in the unobservable case, the system throughput of the PTAS model coincides with that of the vacation model under N-policy when R is large enough). In addition, Figure 7b shows that the superiority of PTAS remains in effect for both large and small Λ in the observable case.

5.4 | PTAS versus pay-for-priority

In Section 3, we have given a brief discussion on how the equilibrium strategy of our PTAS queue differs from that of the priority queue. To reiterate, a major distinction is that the priority queue exhibits a pure FTC behavior while PTAS shows a more sophisticated behavior that is the hybrid of both FTC and ATC. To further this discussion, we next compare the optimal revenue (i.e., revenue under the optimal fee) of these two models. We hereby restrict our attention to the unobservable setting (i.e., the server's state is observable but the queue length is not).[7] Denote by $\Pi^p(\Lambda)$ the optimal revenue by selling priorities, we can have the following result.

Theorem 10 (Comparison of optimal revenue: PTAS vs. pay-for-priority). *In the unobservable case, we have* $\Pi^u(\Lambda) < \Pi^p(\Lambda)$ ($\Pi^u(\Lambda) > \Pi^p(\Lambda)$) *as long as the market size* Λ *is sufficiently large* (*small*).

When the demand volume is low, customers in a priority queue anticipate a smaller expected delay so they intend not to pay for priority service, whereas in our PTAS model, customers are more inclined to purchase PTAS, because otherwise the server's vacation may last for a longer time. When the demand volume is high, customers are incentivized to mitigate their delay via the purchase of priority, while PTAS becomes less necessary because the queue is already long enough to reach level N (see Theorem 7). See Figure 8 for a numerical example, which shows a distinct structure of the two revenue functions: Π^u is unimodal in Λ while Π^p is increasing in Λ. In addition, there exists a cutoff point for the market size Λ, below (above) which the PTAS model yields a higher (lower) revenue than the priority model.

6 | CONCLUSION

In this paper, we study the equilibrium performance of a vacation queueing model with strategic customers. Unlike standard vacation queues in the extant literature where the

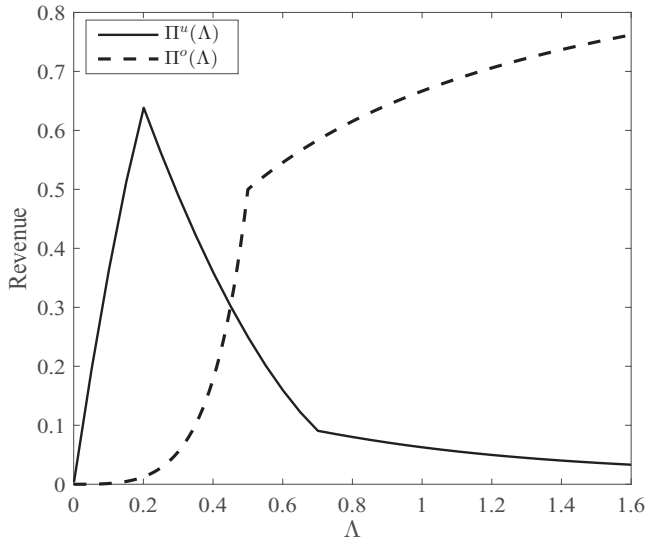


FIGURE 8 Comparison of $\Pi^u(\Lambda)$ and $\Pi^p(\Lambda)$ for different Λ, with $\mu = C = 1$, $N = 5$, and $R = 5$

server's vacation is ended whenever the queue length reaches a critical level, we introduce a new mechanism in that a customer, upon finding the server to be on vacation, may choose to pay a fee to end the server's vacation. This mechanism is referred to as PTAS. The ingenuity of PTAS lies in its ability to allow the server (when on vacation) to be activated immediately by arriving customers, which gives customers more active controls on the server's state (so earlier customer arrivals no longer need to passively wait for future customers to reach the critical queue threshold).

In the present model, customers seeing an inactive server need to make two decisions: (i) whether to join the queue and (ii) if yes, whether to pay for PTAS. We investigate customers' equilibrium joining-and-purchasing strategies, and study their responses to this mechanism under three information cases: (i) observable queue and server state, (ii) unobservable queue and observable server state, and (iii) unobservable queue and server state. Our theoretical analysis reveals results that are seemingly contrary to the conventional wisdom. For example, due to the coexistence of FTC and ATC behavior, a higher service reward does not always guarantee a higher system throughput. In addition, the PTAS revenue is a nonmonotone function in the demand volume (a higher potential demand may even yield a lower revenue). These findings provide quantitative and qualitative insights into the system design of vacation queue systems. We also conduct a careful performance comparison of different information policies.

There are several avenues for future research. One interesting direction is to study customers' rational abandonment behavior in response to the new PTAS mechanism. Another potential topic is to allow the service provider to dynamically adjust the PTAS fee based on the real-time queue length in order to further improve the system revenue.

ACKNOWLEDGMENTS

The authors thank the review team for their constructive comments that helped improve the paper. Zhongbin Wang acknowledges the financial support from the National Natural Science Foundation of China (Grants 72001118 and 72132007).

ORCID

Zhongbin Wang https://orcid.org/0000-0002-1154-5861
Yunan Liu https://orcid.org/0000-0001-9961-2610

ENDNOTES

[1] The NEV subsidy policy aims at contributing toward energy independence and addressing local air quality concerns by promoting the development of NEV (detailed benefits include purchase tax exemption, low interest loan, etc.; https://dieselnet.com/standards/cn/nev.php).

[2] Throughout the paper we refer to Λ as the *demand volume* as well as *market size*, and to μ as the *service capacity*.

[3] The term "system state" refers to different information under different policies: In the *observable* case it includes both the queue length and server's state, and in the *unobservable* case it means solely the server's state.

[4] In Supporting Information, we extend our base models by investigating the third case, called "no-information," where neither the server's state nor queue length is available.

[5] Equilibrium strategies in other cases (i.e., $N = 2, 3, 4$, and $N \geq \mu R/C$) are reported in Proposition 8 in Section C of the Supporting Information.

[6] In the observable case, when the server is active, customers will join (without purchasing PTAS) if and only if the queue length is less than a threshold $\lfloor \mu R/C \rfloor$ in Naor's model.

[7] Explicit equilibrium results in the observable priority queue are more complex and less tractable, see Adiri and Yechiali (1974) and Hassin and Haviv (1997).

REFERENCES

Abhishek, V., Kash, I. A. & Key, P. (2012). Fixed and market pricing for cloud services. In 2012 Proceedings IEEE INFOCOM Workshops (pp. 157–162). IEEE.

Adiri, I., & Yechiali, U. (1974). Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research*, *22*(5), 1051–1066.

Anunrojwong, J., Iyer, K. & Manshadi, V. (2020). Information design for congested social services: Optimal need-based persuasion. *Management Science*, forthcoming. arXiv preprint arXiv:2005.07253.

Aras, A. K., Chen, X., & Liu, Y. (2018). Many-server Gaussian limits for non-Markovian queues with customer abandonment. *Queueing Systems*, *89*(1), 81–125.

Cao, P., He, S., Huang, J., & Liu, Y. (2021). To pool or not to pool: Queueing design for large-scale service systems. *Operations Research*, *69*(6), 1866–1885.

Chen, H., & Frank, M. (2004). Monopoly pricing when customers queue. *IIE Transactions*, *36*(6), 569–581.

Cui, S., Wang, Z., & Yang, L. (2020). The economics of line-sitting. *Management Science*, *66*(1), 227–242.

Dierks, L., & Seuken, S. (2021). Cloud pricing: The spot market strikes back. *Management Science*, *68*(1), 105–122.

Economou, A., & Kanta, S. (2008). Equilibrium balking strategies in the observable single-server queue with breakdowns and repairs. *Operations Research Letters*, *36*(6), 696–699.

Edelson, N. M., & Hilderbrand, D. K. (1975). Congestion tolls for Poisson queuing processes. *Econometrica: Journal of the Econometric Society*, *43*(1), 81–92.

Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT Press.

Gao, J., Iyer, K., & Topaloglu, H. (2019). When fixed price meets priority auctions: Competing firms with different pricing and service rules. *Stochastic Systems*, *9*(1), 47–80.

Gao, S., & Liu, Z. (2013). An M/G/1 queue with single working vacation and vacation interruption under Bernoulli schedule. *Applied Mathematical Modelling*, *37*(3), 1564–1579.

Gavirneni, S., & Kulkarni, V. G. (2016). Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management*, *25*(6), 979–992.

Guo, P., & Hassin, R. (2011). Strategic behavior and social optimization in Markovian vacation queues. *Operations Research*, *59*(4), 986–997.

Guo, P., & Hassin, R. (2012). Strategic behavior and social optimization in Markovian vacation queues: The case of heterogeneous customers. *European Journal of Operational Research*, *222*(2), 278–286.

Guo, P., & Li, Q. (2013). Strategic behavior and social optimization in partially-observable Markovian vacation queues. *Operations Research Letters*, *41*(3), 277–284.

Hassin, R. (1986). Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society*, *54*(5), 1185–1195.

Hassin, R. (2016). *Rational queueing*. CRC Press, Taylor and Francis Group.

Hassin, R., & Haviv, M. (1997). Equilibrium threshold strategies: The case of queues with priorities. *Operations Research*, *45*(6), 966–973.

Hassin, R., & Haviv, M. (2002). Nash equilibrium and subgame perfection in observable queues. *Annals of Operations Research*, *113*(1–4), 15–26.

Hassin, R., & Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems* (Vol. 59). Springer Science & Business Media.

Hassin, R., & Koshman, A. (2017). Profit maximization in the M/M/1 queue. *Operations Research Letters*, *45*(5), 436–441.

Hassin, R., & Roet-Green, R. (2017). The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research*, *65*(3), 804–820.

Hassin, R., & Roet-Green, R. (2018). Cascade equilibrium strategies in a two-server queueing system with inspection cost. *European Journal of Operational Research*, *267*(3), 1014–1026.

Hu, M., Li, Y., & Wang, J. (2018). Efficient ignorance: Information heterogeneity in a queue. *Management Science*, *64*(6), 2650–2671.

Hu, M., Liu, J., & Zhai, X. (2021). Intertemporal segmentation via flexible-duration group buying. *Manufacturing & Service Operations Management*, *23*(5), 1005–1331.

Jacob, J., & Roet-Green, R. (2021). Ride solo or pool: Designing price-service menus for a ride-sharing platform. *European Journal of Operational Research*, *295*(3), 1008–1024.

Li, N. & Xu, J. (2020). Optimal production of a vehicle manufacturer without government subsidies. Available at SSRN 3538010.

Li, Q., Guo, P., Li, C.-L., & Song, J.-S. (2016). Equilibrium joining strategies and optimal control of a make-to-stock queue. *Production and Operations Management*, *25*(9), 1513–1527.

Lingenbrink, D., & Iyer, K. (2019). Optimal signaling mechanisms in unobservable queues. *Operations Research*, *67*(5), 1397–1416.

Liu, Y., & Whitt, W. (2014). Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability*, *24*(1), 378–421.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, *37*(1), 15–24.

Simhon, E., Hayel, Y., Starobinski, D., & Zhu, Q. (2016). Optimal information disclosure policies in strategic queueing games. *Operations Research Letters*, *44*(1), 109–113.

Stidham Jr, S. (2009). *Optimal design of queueing systems*. CRC press.

Sun, K., Liu, Y., & Li, X. (2022a). *Mail back or in-store dropoff? Optimal design of product-exchange policies in omnichannel retailing systems*. Working paper, North Carolina State University.

Sun, K., Liu, Y., & Yang, L. (2022b). Order ahead for pickup: Promise or peril? *Available at SSRN*: https://doi.org/10.2139/ssrn.3673617.

Tian, N., & Zhang, Z. G. (2006). *Vacation queueing models: Theory and applications* (Vol. 93). Springer Science & Business Media.

Wang, J., Cui, S., & Wang, Z. (2019). Equilibrium strategies in M/M/1 priority queues with balking. *Production and Operations Management*, *28*(1), 43–62.

Wang, J., Wang, Z., & Chen, Y. (2021a). Equilibrium strategies and optimal pricing in an online retailing queueing system. *Naval Research Logistics* (*NRL*), *68*(5), 556–576.

Wang, Z., & Fang, L. (2022). The effect of customer awareness on priority queues. *Naval Research Logistics* (*NRL*), https://doi.org/10.1002/nav.22049.

Wang, Z., & Wang, J. (2019). Information heterogeneity in a retrial queue: Throughput and social welfare maximization. *Queueing Systems*, *92*(1-2), 131–172.

Wang, Z., Yang, L., Cui, S., & Wang, J. (2021b). In-queue priority purchase: A dynamic game approach. *Queueing Systems*, *97*(3), 343–381.

Yadin, M., & Naor, P. (1963). Queueing systems with a removable service station. *Journal of the Operational Research Society*, *14*(4), 393–405.

Yang, L., Cui, S., & Wang, Z. (2022). Design of Covid-19 testing queues. *Production and Operations Management*, https://doi.org/10.1111/poms.13673.

Yang, L., & Debo, L. (2019). Referral priority program: Leveraging social ties via operational incentives. *Management Science*, *65*(5), 2231–2248.

Yang, L., Wang, Z., & Cui, S. (2021). A model of queue scalping. *Management Science*, *67*(11), 6803–6821.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wang, Z., Liu, Y., & Fang, L. (2022). Pay to activate service in vacation queues. *Production and Operations Management*, 1–19. https://doi.org/10.1111/poms.13705