APPENDIX

to

Pay to Activate Service in Vacation Queues

This appendix provides supplementary materials to the main paper. In Section A, we extend our base model in several directions. In Section B, we give all technical proofs omitted from the main paper. In Section C, we supplement Theorem 1 by developing the equilibrium strategies with N excluded from the condition in Theorem 1. In Section D, we further explains the cyclic form of the threshold mod $(\bar{n}, I)$ in the observable case.

Appendix A: Extensions

In this section, we extend our base model in several directions. In §A.1 we consider a vacation model under the third information provision rule - neither the queue or the server's state is observable. In §A.2 we consider the case of setup cost for a newly activated server.

A.1. Unobservable Server's State

In our base model, we assume that the server state (busy or on vacation) is available to all customers, and the operational flexibility lies solely in the information provision of the real-time queue length. In this subsection, we consider a new setting, the so-called *no-information*, i.e., both the server's state and the queue length are held unavailable. We continue to study the symmetric mixed strategy. Suppose customers will join the system with probability $q \in [0,1]$ (and balk with probability $1-q$) upon arrival, and a joining customer will purchase PTAS with probability $p \in [0,1]$. Then the strategy space of the customers can be described as a pair $(p,q) \in [0,1] \times [0,1]$ in the two-dimensional space. Under (p,q), the effective arrival rate to the system is $\lambda \equiv \Lambda q$.

A.1.1. System Performance For any given strategy (p,q), we are able to derive the steady-state performance using results in Proposition 1, specifically, with $q_0 = q_1 = q$, because the unavailability of the server's state makes it unnecessary to assign customers with B_0 or B_1 labels as in §3. The steady-state probabilities, expected queue length, and expected waiting time can be obtained immediately.

Corollary 1 (Steady-state performance in the no-information queue) *Consider the no-information M/M/1 vacation queue with PTAS, suppose customers adopt strategy (p,q), the expected number of customers in the system is given by*

$$\overline{N}(p,q) = \frac{\rho q}{1-\rho q} + Q_N(p).$$

The expected waiting time is

$$\overline{w}(p,q) = \frac{1}{\mu - \lambda} + \frac{Q_N(p)}{\lambda}.$$

The following lemma reports useful structural properties for the mean delay $\overline{w}(p,q)$.

Lemma 3 *Consider the no-information M/M/1 vacation queue with PTAS.*

(i). The mean delay $\overline{w}(p,q)$ *is decreasing in* $p \in [0,1]$.

(ii). The mean delay $\overline{w}(p,q)$ *is convex in* $q \in [0, \min\{1, \mu/\Lambda\})$.

a. If $p=1$, $\overline{w}(p,q)$ *is strictly increasing in* $q \in [0, \min\{1, \mu/\Lambda\})$.

b. If $p \in [0,1)$, $\overline{w}(p,q)$ *is increasing first and then decreasing in* q.

(iii). For a given $p \in [0,1]$, $\overline{w}(p,q)$ *can be minimized at* $\hat{q} = \frac{\sqrt{Q_N(p)}}{\rho\left(1+\sqrt{Q_N(p)}\right)}$. *Furthermore,* $\hat{q}$ *is decreasing in* p.

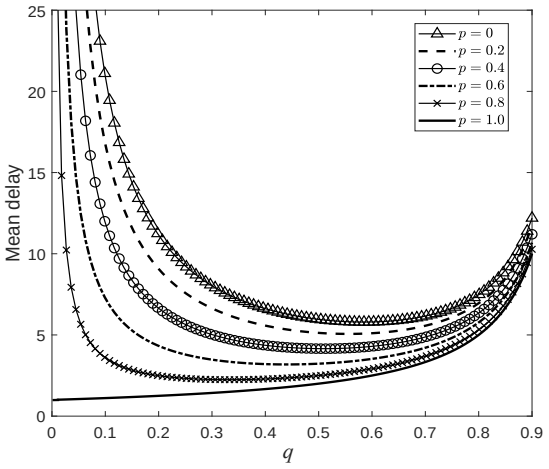


Figure 9 **The mean delay** $\overline{w}(p,q)$ **for different** p **and** q**, with** $\Lambda = 0.9$**,** $\mu = 1$**,** $N = 5$**.**

Figure 9 illustrates the structure of the mean delay for different values of p and q. First, for any given λ, $\overline{w}(p,q)$ is decreasing in p, because when p increases, the system is more likely to be activated. Next, for any $p < 1$, $\overline{w}(p,q)$ has a bathtub shape in q (consistent with Lemma 3). To understand this, recall that the delay in vacation queues may be expressed as the sum of two terms $\overline{w}(p,q) = \overline{w}_b(q) + \overline{w}_v(p,q)$, where the first term, representing the delay induced by existing waiting customers, is an increasing function in q, and the second term, representing the incremented delay due to the server's vacation time, is decreasing in q. When the effective arrival $\lambda = \Lambda q$ is small, increasing q is more efficient in reducing the vacation time (the decrease in $\overline{w}_v(p,q)$ outweighs the increases in $\overline{w}_b(q)$). On the other hand, when the system is already congested with a large q, the reduced vacation time (when q increase) is negligible and not enough to offset the customers' negative externalities. Once again, ATC and FTC co-exist. To understand part (iii) of Lemma 3, note that customer arrivals impact the system performance in two opposing directions: on the one hand, more arrivals can expedite the activation of service which helps relieve the incremented customer delay due to the server's vacations; on the other hand, they bring negative externalities by increasing the system congestion. When p is large, the server can be activated mostly by PTAS, so increasing the effective arrival rate (or equivalently q) leads to a prolonged customer delay. Therefore, a smaller q is required to minimize the average wait time. In particular, we can verify that $\lim_{p\to 1} \hat{q}(p) = \lim_{p\to 1} \frac{\sqrt{Q_N(p)}}{\rho[1+\sqrt{Q_N(p)}]} = 0$ by noticing that $\lim_{p\to 1} Q_N(p) = 0$. This is consistent with Figure 9.

A.1.2. Equilibrium Analysis

When all customers adopt strategy (p,q), the expected utility of an arbitrary customer is given by

$$U_{NI}(p,q) = R - pP - C\overline{w}(p,q).$$

Similar to the analysis in §3.2.2, we define $\Delta w_N(p,q) = \Delta W_N(p)|_{\rho_0=\rho_1=\rho q}$ as the reduced delay by purchasing PTAS when all other customers adopt strategy (p,q), and it is not difficult to verify that $\Delta w_N(p,q)$ is decreasing in q. Thus, when all other customers adopt strategy (p,q), the average reduced delay cost for tagged joining customer by deviating from purchasing probability p to p' is given by

$$\Delta(p';(p,q)) = \sum_{i=0}^{N-2}\left\{\pi_{0,i}\left[(1-p)\cdot\left(\frac{1}{\lambda}\cdot\mathbb{E}[N_i]+\frac{i+1}{\mu}\right)+p\cdot\frac{i+1}{\mu}\right]-\pi_{0,i}\cdot\frac{i+1}{\mu}\right\} = (p'-p)\Delta w_N(p,q). \quad (14)$$

Therefore, when all others adopt strategy $\boldsymbol{\alpha}=(p,q)$, if the tagged customer adopts strategy $\boldsymbol{\alpha}'=(p',q')$, her expected utility is given by

$$\widehat{U}_{NI}(\boldsymbol{\alpha}';\boldsymbol{\alpha}) = q'\left[R - p'P - C\overline{w}(p,q) + (p'-p)\Delta w_N(p,q)\right].$$

Unlike the case in §3, customers in the no-information model are homogeneous. Hence, a strategy profile $\boldsymbol{\alpha}^e=(p^e,q^e)$ is a symmetric Nash equilibrium strategy if and only if

$$\boldsymbol{\alpha}^e \in \underset{\boldsymbol{\alpha}\in[0,1]\times[0,1]}{\arg\max}\ \widehat{U}_{NI}(\boldsymbol{\alpha};\boldsymbol{\alpha}^e). \quad (15)$$

Theorem 11 (Equilibrium strategy in no-information case) *Consider the no-information M/M/1 vacation queue with PTAS. The joint equilibrium joining and purchasing strategy is given below:*

$$\boldsymbol{\alpha}^e = \begin{cases} (0,\min\{1,q_L\}), & \text{if } \Delta w_N(0,\min\{1,q_L\}) \le P; \\ \left(p^e,\min\left\{1,\frac{\mu(R-P)}{\Lambda(R-p^eP)}\right\}\right), & \text{if } \Delta w_N\left(p^e,\min\left\{1,\frac{\mu(R-P)}{\Lambda(R-p^eP)}\right\}\right)=P; \\ \left(1,\min\left\{1,\frac{\mu(R-P)-C}{\Lambda(R-P)}\right\}\right), & \text{if } \Delta w_N\left(1,\min\left\{1,\frac{\mu(R-P)-C}{\Lambda(R-P)}\right\}\right)\ge P, \end{cases}$$

where $q_L = \frac{\sqrt{C^2(N-3)^2-4C\mu(N+1)R+4\mu^2R^2}+C(N-3)+2\mu R}{4\Lambda R}$.

The equilibrium $\boldsymbol{\alpha}^e$ characterized in Theorem 11 is not necessarily unique, in which we can further refine these derived equilibria using criteria specified by Definitions 2-3.

Paralleling our steps to treat the base models, we derive the system throughput and PTAS revenue. Under the equilibrium strategy $\boldsymbol{\alpha}=(p,q)$, the system throughput is given by

$$\lambda^{NI}(p,q) = \Lambda q, \quad (16)$$

which only depends on the joining strategy (independent of the purchasing strategy p). The next result benchmarks the present model with a standard vacation queue.

Proposition 5 (PTAS improves throughput) *PTAS achieves improved system throughput for an M/M/1 vacation queue in the no-information case.*

Proposition 5 supplements Theorem 9 to confirm that PTAS is effective under all information revelation policies. For a given (p,q), the service provider's revenue collected via PTAS is a function of Λ

$$\Pi^{NI}(\Lambda) = p\Lambda q(1-\Lambda q/\mu)P, \tag{17}$$

where $1-\Lambda q/\mu$ is the steady-state probability that the server is inactive.

Proposition 6 (No information excels in high congestion) *When $R \geq CN/\mu$, there exists a threshold $\tilde{\Lambda}'$ such that $\Pi^{NI}(\Lambda) > \max\{\Pi^u(\Lambda), \Pi^o(\Lambda)\}$ if $\Lambda > \tilde{\Lambda}'$.*

We conduct a numerical example to compare the revenue under all three information policies, see Figure 10. Similar to the other two information cases, the PTAS revenue $\Pi^{NI}(\Lambda)$ under the no information policy is non-monotonic in Λ. The intuition is indeed similar (see discussions following Theorem 7). However, different from $\Pi^u(\Lambda)$ and $\Pi^o(\Lambda)$ both diminish to zero as Λ grows large, $\Pi^{NI}(\Lambda)$ quickly becomes stable and plateaus afterwards. We give some explanations: Unlike the other two information cases where the effective arrival rate increases in Λ, the effective arrival rate in the no-information case eventually becomes a constant (guaranteed by a decreasing joining probability q coping with the increasing Λ). This is consistent with the standard unobservable queue (Edelson and Hilderbrand 1975). Since increasing Λ (when it is already large enough) has no further impact on the system dynamics and does not increase the workload, a fraction of joining customers will continue to purchase PTAS and warrant a positive system revenue.

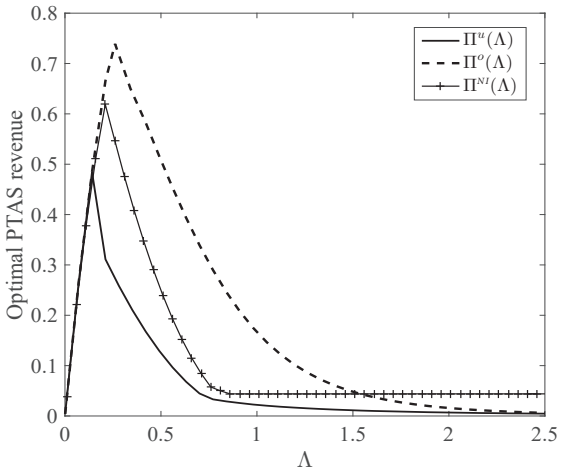


Figure 10 **The comparison of Π^u, Π^p and Π^{NI} for different Λ, with $\mu = C = 1$, $N = 2$ and $R = 5$.**

A.2. Server's Setup Cost

In our base model, the system incurs no cost when the server changes its state. However, in practice, it is often the case that when the server's vacation ends, it generates a inconvenient cost before returning to its normal working state. This is widely observed in make-to-order production systems with large machinery. In this section, we will extend our base model in this direction; we call this service-resumption cost the "setup

cost". And we will investigate when it is beneficial for the service provider to provide PTAS in order to appropriately balance the throughput, PTAS revenue and setup cost.

Let $K>0$ be the step cost, which can be incurred either by PTAS or the N^{th} customer arrival. For any fixed PTAS fee P, in equilibrium, we denote by $SC^u(P)$ and $SC^o(P)$ the setup cost per unit time in the unobservable and observable case, respectively. We next formulate the following optimization problem, where the service provider aims to maximize his net benefit by selecting the optimal PTAS fee.

$$\max_{P\in[0,R]} \quad \tilde{\Pi}^x(P)=\Pi^x(\Lambda,P)-SC^x(P), \quad x=u,o, \tag{18}$$

where $\Pi^x(\Lambda,P)$ is the PTAS revenue rate under information structure x in equilibrium when the demand volume and price are Λ and P, respectively. Let $\lambda^x(P)$ be the system throughput under information structure x when the PTAS fee is P. Because no customer has incentive to purchase PTAS when $P=R$, so $\lambda^x(R)$ is the system throughput without PTAS. Note that it is profitable to provide PTAS if and only if

$$\tilde{\Pi}^x(P)\geq\tilde{\Pi}^x(R), \quad x=u,o,$$

where $\tilde{\Pi}^x(R)=-SC^x(R)$. In the unobservable case, i.e., $x=u$, we have

$$SC^u(P)=K[\lambda_0\pi_{0,N-1}+\sum_{i=0}^{N-2}\pi_{0,i}\lambda_0 p^e]=\frac{p^e\rho_0\mu(1-\rho_1)K}{[1-(1-p^e)^N](1+\rho_0-\rho_1)},$$
$$\Pi^u(\Lambda,P)=\frac{(1-\rho_1)p^e\rho_0\mu P}{1+\rho_0-\rho_1}.$$

According to Proposition 1, it gives

$$\tilde{\Pi}^u(P)=\frac{(1-\rho_1)p^e\rho_0\mu}{1+\rho_0-\rho_1}\left[P-\frac{K}{1-(1-p^e)^N}\right], \quad \tilde{\Pi}^u(R)=-\frac{\rho_0'\mu(1-\rho_1')K}{N(1+\rho_0'-\rho_1')},$$

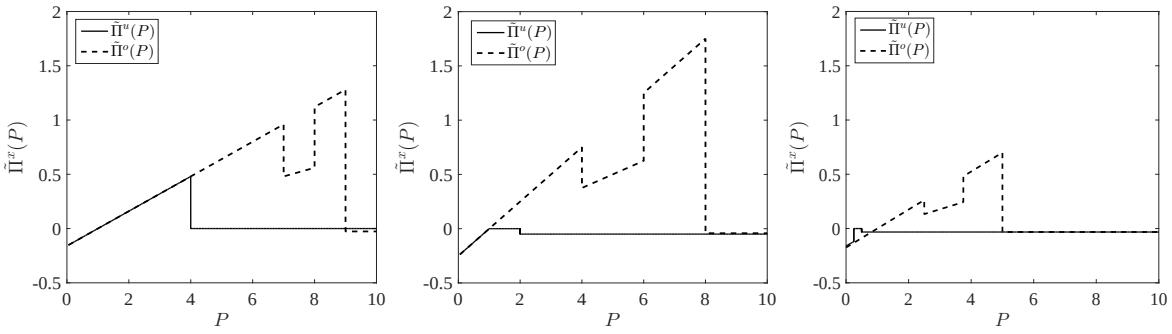
where (ρ_0',ρ_1') are the equilibrium pair without PTAS, see Li et al. (2016). On the other hand, in the observable case, we have (according to Theorem 4)

$$SC^o(P)=\frac{\Lambda K(1-\rho)^2}{(1-\rho)(\bar{n}_2+1)+\rho^{n_1+1}(\rho-\rho^{-\bar{n}_2})}, \quad SC^o(R)=\frac{\Lambda K(1-\rho)^2}{(1-\rho)(N+1)+\rho^{n_1+1}(\rho-\rho^{-N})}$$

It follows that

$$\tilde{\Pi}^o(P)=\frac{\Lambda(P-K)(1-\rho)^2}{(1-\rho)(\bar{n}_2+1)+\rho^{n_1+1}(\rho-\rho^{-\bar{n}_2})}, \quad \tilde{\Pi}^o(R)=-\frac{K\Lambda(1-\rho)^2}{(1-\rho)(N+1)+\rho^{n_1+1}(\rho-\rho^{-N})},$$

where $\bar{n}_2=1$ if $P\leq C\Lambda$ and $\bar{n}_2=n_2$ otherwise, n_1 and n_2 are specified in Theorem 4. We next provide a numerical example. In Figure 11 we plot $\tilde{\Pi}^u(P)$ and $\tilde{\Pi}^o(P)$ for different P and Λ. Figure 11 shows that PTAS is an effective measure for improving the system revenue when the demand volume Λ is intermediate. Such an observation is consistent with our base model (see Theorem 7). In addition, providing the real-time queue length information helps improve the PTAS revenue.



(a) $\Lambda = 0.2$ (b) $\Lambda = 0.5$ (c) $\Lambda = 0.8$

Figure 11 $\tilde{\Pi}^u(P)$ **and** $\tilde{\Pi}^o(P)$ **for different** P **and** Λ**, with** $\mu = C = K = 1$**,** $N = 5$ **and** $R = 10$**.**

A.3. Reward from Taking Vacations

We now consider a more general setting in that the service provider aims to optimize a reward that is the sum of two terms: (i) revenue by completing services, and (ii) gain from taking vacations. Let V denote the gain from vacation per unit time[8] and B be the revenue per service completion. The service provider's objective function is

$$\max_{P\in[0,R],N\in\mathbb{N}^*} \quad \tilde{\Pi}_N^x(P) = B\lambda_e^x(P) + V\cdot\pi_V^x, \quad x = u, o, \tag{19}$$

where π_V^x which is the steady-state fraction of time the server is inactive under information policy x. The system throughput is $\lambda_e^x(P) = \mu\pi_B^x$, where $\pi_B^x = 1 - \pi_V^x$ is the steady-state probability that the server is active. Thus, the above optimization problem is reformulated below.

$$\max_{P\in[0,R],N\in\mathbb{N}^*} \quad \tilde{\Pi}_N^x(P) = V + (B\mu - V)\pi_B^x, \quad x = u, o. \tag{20}$$

When $V < B\mu$, maximizing $\tilde{\Pi}_N^x(P)$ is equivalent to maximizing the steady-state probability that the server is active, or equivalently, maximizing the system throughput. Thus it is optimal to set a sufficiently low vacation threshold to attract as many customers as possible to join. When $V \geq B\mu$, maximizing $\tilde{\Pi}_N^x(P)$ is equivalent to maximizing π_V^x, or equivalently, minimizing the system throughput. Then it is optimal to set a sufficiently large vacation threshold to let all arriving customers balk. Solutions to the optimization problem (20) are summarized below.

Proposition 7 *If* $V < B\mu$*,* $(P^*, N^*) = (0, 1)$ *is an optimal solution to* (20)*; otherwise,* $(P^*, N^*) = (R, \infty)$ *is an optimal solution to* (20)*.*

The result in Proposition 7 is intuitive. When the benefit of taking a vacation is small compared to the benefit by serving a customer, it is optimal for the service provider to remain active as much as possible. Thus N should be selected at 1, which reduces to the work-conservation queues and the PTAS will never

[8] Here V may mean the reduction of the system's operating cost, or the improvement on the servers' mental and physical condition, from taking a vacation.

be used. By contrast, when the benefit of taking a vacation is relatively large, the service provider could benefit in taking vacation instead of working at any time. As a result, it is optimal to close down the service industry by setting a sufficiently large PTAS fee and vacation threshold.

The above results are due to the linearity of the problem structure of (20). We admit that it will be more interesting to consider nonlinear problems which should yield "non-trivial" solution for N. Nevertheless, Problem (20) is only an initial attempt of this new setting. The more in-depth investigation of this subject is beyond the scope of the present paper. We plan to consider the more general settings in future works.

Appendix B: Technical Proofs

Proof of Proposition 1. Combining (2)-(3) gives

$$\pi_{0,i} = \pi_{1,1} \cdot \frac{(1-p)^i}{\rho_0} \tag{21}$$

for $i = 0, 1, \ldots, N-1$. From (6), we can obtain that

$$\pi_{1,i} = \pi_{1,N} \cdot (\rho_1)^{i-N} \tag{22}$$

for $i \geq N$. Notice that (4) can be rewritten as

$$(\pi_{1,1} - 0)\rho_1 = \pi_{0,0}\rho_0 p + (\pi_{1,2} - \pi_{1,1}) \tag{23}$$

$$\vdots$$

$$(\pi_{1,N-1} - \pi_{1,N-2})\rho_1 = \pi_{0,N-2} \cdot \rho_0 p + (\pi_{1,N} - \pi_{1,N-1}). \tag{24}$$

Summing up from (23) to (24) and combining (21), we can obtain

$$\pi_{1,N} = \pi_{1,N-1} \cdot \rho q + \pi_{1,1} \cdot (1-p)^{N-1}. \tag{25}$$

Let $a_i = \pi_{1,i} - \pi_{1,i-1}$ for $i = 1, 2, \ldots, N-1$, by combining (21), equations (23) to (24) can be rewritten as

$$a_1\rho_1 = \pi_{1,1}(1-p)^0 p + a_2 \tag{26}$$

$$a_2\rho_1 = \pi_{1,1}(1-p)^1 p + a_3$$

$$\vdots$$

$$a_{N-1}\rho_1 = \pi_{1,1}(1-p)^{N-2} p + a_N. \tag{27}$$

Multiplying $(\rho q)^{i-k-1}$ to both sides of equation $a_k\rho q = \pi_{1,1}(1-p)^{k-1}p + a_{k+1}$ for $k = 1, 2, \ldots, i-1$, and summing them up from $k = 1$ to $i-1$ yield

$$\begin{aligned} a_i &= a_1\rho_1^{i-1} - \pi_{1,1}\sum_{k=1}^{i-1}(1-p)^{k-1}\rho_1^{i-k-1}p \\ &= \pi_{1,1}\rho_1^{i-1} - \pi_{1,1}\sum_{k=1}^{i-1}(1-p)^{k-1}\rho_1^{i-k-1}p \\ &= \pi_{1,1} \cdot \left[\frac{(1-\rho_1)\rho_1^{i-1} - p(1-p)^{i-1}}{1-p-\rho_1}\right]. \end{aligned} \tag{28}$$

Because $\pi_{1,0}=0$, $a_1=\pi_{1,1}$, and $\pi_{1,i}=\sum_{k=1}^{i}a_k$, it follows that

$$\begin{aligned}\pi_{1,i}&=\pi_{1,1}\cdot\sum_{k=1}^{i}\left[\frac{(1-\rho_1)\rho_1^{k-1}-p(1-p)^{k-1}}{1-p-\rho_1}\right]\\&=\pi_{1,1}\cdot\left[\frac{(1-\rho_1)}{1-p-\rho_1}\sum_{k=1}^{i}\rho_1^{k-1}-\frac{p}{1-p-\rho_1}\sum_{k=1}^{i}(1-p)^{k-1}\right]\\&=\pi_{1,1}\cdot\left[\frac{(1-p)^i-\rho_1^i}{1-p-\rho_1}\right]\end{aligned}\tag{29}$$

for $i=1,2,\ldots,N-1$. When $i=N-1$, by plugging (29) into (25), we can get

$$\pi_{1,N}=\pi_{1,1}\cdot\left[\frac{(1-p)^N-\rho_1^N}{1-p-\rho_1}\right].\tag{30}$$

Combining (22) and (30) gives

$$\pi_{1,i}=\pi_{1,1}\cdot\rho_1^{i-N}\left[\frac{(1-p)^N-\rho_1^N}{1-p-\rho_1}\right]\tag{31}$$

for $i\geq N$. So far, all steady-state probabilities have been expressed by $\pi_{1,1}$, then $\pi_{1,1}$ can be derived through normalization condition $\sum_{i=0}^{N-1}\pi_{0,i}+\sum_{i=1}^{N-1}\pi_{1,i}+\sum_{i=N}^{\infty}\pi_{1,i}=1$. From (21), (29) and (31), we have

$$\begin{aligned}\sum_{i=0}^{N-1}\pi_{0,i}&=\frac{1-(1-p)^N}{\rho_0 p}\cdot\pi_{1,1},\\\sum_{i=1}^{N-1}\pi_{1,i}&=\left[\frac{(1-p)(1-\rho_1)(1-(1-p)^{N-1})-\rho_1 p(1-\rho_1^{N-1})}{(1-\rho_1)p(1-p-\rho_1)}\right]\cdot\pi_{1,1},\\\sum_{i=N}^{\infty}\pi_{1,i}&=\left[\frac{(1-p)^N-\rho_1^N}{(1-p-\rho_1)(1-\rho_1)}\right]\cdot\pi_{1,1}.\end{aligned}$$

Therefore, it follows that

$$\pi_{1,1}=\frac{\rho_0(1-\rho_1)p}{[1-(1-p)^N](1+\rho_0-\rho_1)}.$$

Then all steady-state probabilities can be derived using $\pi_{1,1}$. On the other hand, the mean number of customers in the system can be expressed as $\overline{N}(p,q_0,q_1)=N_0+N_{1,1}+N_{1,2}$, where $N_0=\sum_{i=0}^{N-1}\pi_{0,i}\cdot i$, $N_{1,1}=\sum_{i=1}^{N-1}\pi_{1,i}\cdot i$ and $N_{1,2}=\sum_{i=N}^{\infty}\pi_{1,i}\cdot i$. After some algebraic manipulations, we can obtain

$$\begin{aligned}N_0&=\frac{(1-p)\left(1-(1-p)^N-N(1-p)^{N-1}p\right)(1-\rho_1)}{\left(1-(1-p)^N\right)p(1+\rho_0-\rho_1)},\\N_{1,1}&=\frac{\rho_0\left[\left(1-(1-p)^N\right)(1-\rho_1)^2-\left(1-(1-p)^N+N(1-p)^N\right)p(1-\rho_1)^2-p^2\left(\rho_1-N\rho_1^N+(N-1)\rho_1^{N+1}\right)\right]}{\left(1-(1-p)^N\right)p(1+\rho_0-\rho_1)(1-\rho_1)(1-p-\rho_1)},\\N_{1,2}&=\frac{p\rho_0(N(1-\rho_1)+\rho_1)\left((1-p)^N-\rho_1^N\right)}{\left(1-(1-p)^N\right)(1+\rho_0-\rho_1)(1-\rho_1)(1-p-\rho_1)},\end{aligned}$$

which yield

$$Q(p,q_0,q_1)=\frac{\rho_0}{(1-\rho_1)(1+\rho_0-\rho_1)}+Q_N(p).$$

The expected waiting time for the customers who find a busy server is given by

$$w_1(p,q_0,q_1)=\frac{1}{\mu}\left(\frac{N_{1,1}+N_{1,2}}{\sum_{i=1}^{\infty}\pi_{1,i}}+1\right)=\frac{1}{\mu}\left[\frac{1}{1-\rho_1}+1+Q_N(p)\right].$$

The expected waiting time for the customers who find the server to be on vacation is given by

$$w_0(p,q_0,q_1)=\frac{1}{\mu}\left(\frac{N_0}{\sum_{i=0}^{N-1}\pi_{0,i}}+1\right)+\frac{1-p}{\lambda_0\sum_{i=0}^{N-1}\pi_{0,i}}\cdot\sum_{i=0}^{N-1}\pi_{0,i}\cdot\mathbb{E}[N_i]=\frac{1}{\mu}+\frac{Q_N(p)}{\mu}\left[1+\frac{1}{\rho_0}\right],$$

where $\mathbb{E}[N_i]=[1-(1-p)^{N-i-1}]/p$. Accordingly, the system throughput is

$$\lambda^u(p,q_0,q_1)=\sum_{i=0}^{N-1}\pi_{0,i}\lambda_0+\sum_{i=0}^{\infty}\pi_{1,i}\lambda_1=\frac{\lambda_0}{1+(\lambda_0-\lambda_1)/\mu}.$$

Finally, the service provider's revenue collected from PTAS is given by

$$\Pi^u(p,q_0,q_1)=P\sum_{i=0}^{N-1}\pi_{0,i}\lambda_0 p=\frac{(1-\lambda_1/\mu)p\lambda_0 P}{1+(\lambda_0-\lambda_1)/\mu},$$

which completes this proof. ■

Proof of Lemma 1. By taking the derivative of $Q_N(p)$ with respect to p, we have

$$\frac{dQ_N(p)}{dp}=\frac{N^2(1-p)^{N-1}p^2-[1-(1-p)^N]^2}{[1-(1-p)^N]^2p^2}.$$

Then it is sufficient to show that $N^2(1-p)^{N-1}p^2\le[1-(1-p)^N]^2$ for all $p\in[0,1]$, i.e.,

$$N(1-p)^{\frac{N-1}{2}}p-[1-(1-p)^N]\le 0. \tag{32}$$

Taking the derivative of (32) with respect to p gives

$$\frac{dN(1-p)^{\frac{N-1}{2}}p-[1-(1-p)^N]}{dp}=-\frac{N(1-p)^{\frac{1}{2}(N-3)}}{2}\left(2\left[(1-p)^{\frac{1+N}{2}}-1\right]+(N+1)p\right)$$

$$\frac{d\left(2\left[(1-p)^{\frac{1+N}{2}}-1\right]+(N+1)p\right)}{dp}=(1+N)(1-(1-p)^{\frac{N-1}{2}})>0.$$

Thus $2\left[(1-p)^{\frac{1+N}{2}}-1\right]+(N+1)p$ is minimized at $p=0$. When $p=0$, $2\left[(1-p)^{\frac{1+N}{2}}-1\right]+(N+1)p=0$. Hence

$$\frac{dN(1-p)^{\frac{N-1}{2}}p-[1-(1-p)^N]}{dp}\le 0$$

for all $p\in[0,1]$, which implies that $N(1-p)^{\frac{N-1}{2}}p-[1-(1-p)^N]$ is decreasing in $p\in[0,1]$. When $p=0$, the left-hand side of (32) is 0, which implies that $N^2(1-p)^{N-1}p^2\le[1-(1-p)^N]^2$ for all $p\in[0,1]$, i.e., $dQ_N(p)/dp<0$. On the other hand, it follows from L'Hospital rule that $Q_N(0)=(N-1)/2$ and $Q_N(1)=0$, which completes this proof. ■

Proofs of Propositions 2-3. To simplify the notations, we define

$$\rho_s(p)=\frac{Q_N(p)}{\mu(R-pP)/C-Q_N(p)-1},\quad \rho_l(p)=1-\frac{1}{\mu R/C-Q_N(p)-1}.$$

For any given strategy $p\in[0,1)$, we can find that $w_0(p,q_0)$ and $w_1(p,q_1)$ are decreasing and increasing in $q_0\in[0,1]$, respectively. For the customers who find an inactive server, if $U_0(p,1)<0$, then the equilibrium arrival rate for the customers who find the server on vacation is $\lambda_0^e=0$, and the system can never be activated as all arriving customers will balk. If $U_0(p,1)\ge 0$, there exist two equilibria: $q_0^e=1$ and $q_0^e=\frac{Q_N(p)}{\rho[\mu(R-pP)/C-Q_N(p)-1]}$, and we can verify that $q_0^e=1$ is an ESS.

For the customers who find a busy server, if $Q_N(p)-\mu R/C+2>0$, we have $U_1(p,q_1)<0$ for all $q_1\in[0,\min\{\mu/\Lambda,1\}]$, then $q_1(p)^e=0$ is the unique equilibrium. If $Q_N(p)-\mu R/C+2\le 0$, we consider two subcases,

(i) when $\rho<1$ and $U_1(p,1)\geq 0$, i.e., $\rho\leq 1-\frac{1}{\mu R/C-Q_N(p)-1}$, which gives $\lambda_1^e=\Lambda$. Otherwise, (ii) when $\rho>1-\frac{1}{\mu R/C-Q_N(p)-1}$, it follows that $q_1^e=\frac{1}{\rho}-\frac{1}{\rho[\mu R/C-Q_N(p)-1]}$. In summary, we can get that

$$q_0^e(p)=\begin{cases}0, & \text{if } \rho\leq\rho_s(p);\\ \underline{1} \text{ or } \frac{Q_N(p)}{\rho[\mu(R-pP)/C-Q_N(p)-1]}, & \text{if } \rho>\rho_s(p),\end{cases}$$

$$q_1^e(p)=\begin{cases}1, & \text{if } \rho\leq\rho_l(p) \text{ and } Q_N(p)\leq\mu R/C-2;\\ \frac{1}{\rho}-\frac{1}{\rho[\mu R/C-Q_N(p)-1]}, & \text{if } \rho>\rho_l(p) \text{ and } Q_N(p)\leq\mu R/C-2;\\ 0, & \text{if } Q_N(p)>\mu R/C-2,\end{cases}$$

where $q_0^e(p)=1$ is the unique ESS. ■

Proof of Lemma 2. For any given $p\in[0,1]$, taking the derivative of $\Delta W_N(p)$ with respect to N, gives

$$\frac{\partial\Delta W_N(p)}{\partial N}=-\frac{C(1-\rho_1)(1-p)^{N-1}\left(1-(1-p)^N+N\ln(1-p)\right)}{\lambda_0(1+\rho_0+\rho_1)\left(1-(1-p)^N\right)^2},$$

$$\frac{d[1-(1-p)^N+N\ln(1-p)]}{dN}=[1-(1-p)^N]\ln(1-p)<0.$$

Thus, $1-(1-p)^N+N\ln(1-p)$ is decreasing in $N\geq 2$. Since $[1-(1-p)^N+N\ln(1-p)]\,|_{N=2}=(2-p)p+2\ln(1-p)\leq 0$ for all $p\in[0,1]$, we have $1-(1-p)^N+N\ln(1-p)<[1-(1-p)^N+N\ln(1-p)]\,|_{N=2}\leq 0$ for all $N\geq 2$. It follows that $\frac{\partial\Delta W_N(p)}{\partial N}>0$ for all $N\geq 2$, i.e., $\Delta W_N(p)$ is increasing in N.

Next, we consider the following cases according to N.

When $N=2$, we have $\Delta W_N(p)=\frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)(2-p)}$, which is obviously increasing in $p\in[0,1]$.

When $N=3$, we have $\Delta W_N(p)=\frac{3-2p}{3-3p+p^2}\frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$, by taking the derivative of $\Delta W_N(p)$ with respect to p, it gives

$$\frac{\partial\Delta W_N(p)}{\partial p}=\frac{3-6p+2p^2}{(3-3p+p^2)^2}\cdot\frac{1-\rho_1}{\lambda_0(1+\rho_0-\rho_1)},$$

$$\frac{\partial^2\Delta W_N(p)}{\partial p^2}=-\frac{2(3-p)p(3-2p)}{(3-3p+p^2)^3}\cdot\frac{1-\rho_1}{\lambda_0(1+\rho_0-\rho_1)}<0.$$

This implies that $\frac{d\Delta W_N(p)}{dp}>0$ for $p\in[0,\left(3-\sqrt{3}\right)/2]$ and $\frac{d\Delta W_N(p)}{dp}<0$ for $(\left(3-\sqrt{3}\right)/2,1]$. Then $\Delta W_N(p)$ is unimodal in $p\in[0,1]$.

When $N=4$, we have $\Delta W_N(p)=\frac{6-8p+3p^2}{4-6p+4p^2-p^3}\cdot\frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$, taking the derivative of $\Delta W_N(p)$ with respect to p gives

$$\frac{\partial\Delta W_N(p)}{\partial p}=\frac{h_4(p)}{(-4+6p-4p^2+p^3)^2}\cdot\frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)},$$

$$h_4'(p)=4(-6+16p-12p^2+3p^3),$$

$$h_4''(p)=4(16-24p+9p^2)>0,$$

where $h_4(p)=4-24p+32p^2-16p^3+3p^4$. It follows that $h_4'(p)$ is increasing in $p\in[0,1]$. As $h_4'(p)|_{p=0}=-24<0<h_4'(p)|_{p=1}=4$, we can get that $h_4(p)$ is decreasing first and then increasing in $p\in[0,1]$. It is not difficult to verify that $h_4(p)|_{p=0}=4>0>h_4(p)|_{p=1}=-1$, thus $\Delta W_N(p)$ is increasing first and then decreasing in $p\in[0,1]$, i.e., $\Delta W_N(p)$ is unimodal in $p\in[0,1]$.

When $N \geq 5$, we have

$$\frac{\partial \Delta W_N(p)}{\partial p} = \frac{N(1-p)^N[N-1+(1-p)^N]p^2 - (1-p)^2[1-(1-p)^N]^2}{(1-(1-p)^N)^2(1-p)^2p^2} \cdot \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}.$$

Then it suffices to show

$$N(1-p)^N[N-1+(1-p)^N]p^2 - (1-p)^2[1-(1-p)^N]^2 < 0 \quad \text{or} \quad \left[\frac{N}{1-(1-p)^N} - 1\right] \frac{N(1-p)^N}{1-(1-p)^N} < \frac{(1-p)^2}{p^2}.$$

Let $x = (1-p)^N$, it is sufficient to show

$$\frac{x \ln x}{1-x}\left(\frac{\ln x}{1-x} - \ln(1-p)\right) < \left[\frac{(1-p)\ln(1-p)}{p}\right]^2$$

holds for all $p \in [0,1]$ and $x \in [0,(1-p)^5]$. Let $f(x,p) = \frac{x\ln x}{1-x}\left(\frac{\ln x}{1-x} - \ln(1-p)\right)$, we have

$$\frac{\partial f(x,p)}{\partial x} = -\frac{(1-x)\ln(1-p)(1-x+\ln x) - \ln x(2-2x+(1+x)\ln x)}{(1-x)^3}.$$

Since $1-x+\ln x < 0$ and $\ln(1-p) = \ln x/N \geq \ln x/5$ when $N \geq 5$, it follows that $\partial f(x,p)/\partial x > h_5(x)/(1-x)^3$, where $h_5(x) = \ln x[9-8x-x^2+2(2+3x)\ln x]$. Notice that

$$\frac{dh_5(x)}{dx} = -2 + 4/x - 2x + 6\ln x, \qquad \frac{d^2h_5(x)\ln x}{dx^2} = -\frac{2\,(2-3x+x^2)}{x^2} < 0,$$

so that $\frac{dh_5(x)}{dx} > \frac{dh_5(x)}{dx}|_{x=1} = 0$. Then $9-8x-x^2+2(2+3x)\ln x < (9-8x-x^2+2(2+3x)\ln x < 0)|_{x=1} = 0$. Since $\ln x < 0$ for $x \in [0,(1-p)^5]$, it follows that

$$\frac{\partial f(x,p)}{\partial x} > \frac{h_5(x)}{(1-x)^3} > 0.$$

In other words, we can deduce that $\left[\frac{N}{1-(1-p)^N} - 1\right] \cdot \frac{N(1-p)^N}{1-(1-p)^N}$ is decreasing in $N \geq 5$. Then it is sufficient to show

$$\left(\left[\frac{N}{1-(1-p)^N} - 1\right] \frac{N(1-p)^N}{1-(1-p)^N}\right)|_{N=5} < \frac{(1-p)^2}{p^2}$$

$$\text{or equivalently} \quad (1-p)^2p^5\left(-50+140p-160p^2+95p^3-30p^4+4p^5\right) < 0.$$

Let $g(p) = -50+140p-160p^2+95p^3-30p^4+4p^5$, we have

$$\frac{dg(p)}{dp} = 140-320p+285p^2-120p^3+20p^4, \quad \frac{d^2g(p)}{dp^2} = -320+570p-360p^2+80p^3,$$
$$\frac{d^3g(p)}{dp^3} = 570-720p+240p^2, \quad \frac{d^4g(p)}{dp^4} = -720+480p < 0.$$

Therefore, $\frac{d^3g(p)}{dp^3}$ is decreasing in $p \in [0,1]$. Since $\frac{d^3g(p)}{dp^3}|_{p=1} > 0$, we have $\frac{d^3g(p)}{dp^3} > 0$ for $p \in [0,1]$. It implies that $\frac{d^2g(p)}{dp^2}$ is increasing in $p \in [0,1]$. As $\frac{d^2g(p)}{dp^2}|_{p=1} < 0$, then $\frac{d^2g(p)}{dp^2} < 0$ for all $p \in [0,1]$, i.e., $\frac{dg(p)}{dp}$ is decreasing in $p \in [0,1]$. Because $\frac{dg(p)}{dp}|_{p=1} = 5 > 0$, we can get that $g(p)$ is increasing in $p \in [0,1]$, which follows that $g(p) \leq 0$ by noticing that $g(p) < g(1) = -1$. Thus we can complete this proof. ■

Proof of Proposition 4. We consider the following four cases:

(1) When $N = 2$, $\Delta W_N(p)$ is increasing in p. (i) If $\Delta W_N(0) \geq P$, we have $\Delta W_N(p) \geq P$ for all $p \in [0,1]$, then purchasing PTAS is the unique equilibrium. Since $\lim_{p\to 0^+} \Delta W_N(p) = \frac{C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}$, purchasing PTAS is the

unique equilibrium if and only if $P \le \frac{C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}$. (ii) If $\Delta W_N(1) \le P$, we have $\Delta W_N(p) \le P$ for all $p \in [0,1]$, then never purchasing PTAS is the unique equilibrium. Since $\lim_{p\to 1} \Delta W_N(p) = C$, never purchasing PTAS is the unique equilibrium if and only if $P > \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$. (iii) Otherwise, if $P \in (\Delta W_N(0), \Delta W_N(1))$, there exist two pure equilibrium strategies $p_e^{(1)} = 0$, $p_e^{(2)} = 1$ and unique mixed equilibrium $p_e^{(3)} = \frac{2P\lambda_0(1+\rho_0-\rho_1)-C(1-\rho_1)}{P\lambda_0(1+\rho_0-\rho_1)}$ that uniquely solves $\Delta W_N(p) = P$ in $p \in (0,1)$. Among the three equilibria above, by Definition 3, we can verify that the mixed equilibrium $p_e^{(3)} = \frac{2P\lambda_0(1+\rho_0-\rho_1)-C(1-\rho_1)}{P\lambda_0(1+\rho_0-\rho_1)}$ is not ESS because when others adopt strategy $p_e^{(3)} + \delta$ for some small $\delta > 0$, we have $\Delta W_N(p_e^{(3)} + \delta) > P$. Thus for the tagged customer, the best response is to deviate from strategy $p_e^{(3)}$ to $p_{BR} = 1$.

(2) When $N = 3$, we have $\Delta W_N(p) = \frac{3-2p}{3-3p+p^2} \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$, which is unimodal in p. We can verify that $\Delta W_N(p)|_{p=0} = \Delta W_N(p)|_{p=1} = \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)} < \Delta W_N(p)|_{p\in(0,1)}$ and $\Delta W_N(p)$ is maximized at $\hat{p}_3 = (3-\sqrt{3})/2$, where $Q_N(\hat{p}_3) = \frac{2C\sqrt{3}(1-\rho_1)}{3\lambda_0(1+\rho_0-\rho_1)}$. Therefore, (i) if $P \le \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$, we have $\Delta W_N(p) \ge P$ for all $p \in [0,1]$, then purchasing PTAS is the unique equilibrium, i.e., $p_e = 1$. (ii) If $P > Q_N(\hat{p}_3)$, then never purchasing PTAS is the unique equilibrium by noticing that $Q_N(p) \le P$ for all $p \in [0,1]$, i.e., $p_e = 0$. (iii) Otherwise, if $P \in (\frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}, \frac{2C\sqrt{3}(1-\rho_1)}{3\lambda_0(1+\rho_0-\rho_1)})$, there exist one pure equilibrium strategy $p_e^{(1)} = 0$ and two mixed equilibrium strategies

$$\begin{aligned} p_e^{(2)} &= \frac{3P - 2(1-\rho_1)C/(\lambda_0[1+\rho_0-\rho_1]) - \sqrt{4[(1-\rho_1)C/(\lambda_0[1+\rho_0-\rho_1])]^2 - 3P^2}}{2P} \\ &< \frac{3P - 2(1-\rho_1)C/(\lambda_0[1+\rho_0-\rho_1]) + \sqrt{4[(1-\rho_1)C/(\lambda_0[1+\rho_0-\rho_1])]^2 - 3P^2}}{2P} = p_e^{(3)}. \end{aligned}$$

Similar to the argument in case (1), we can verify that $p_e^{(2)}$ is not ESS among the three equilibria.

(3) When $N = 4$, we have $\Delta W_N(p) = \frac{C[1-\rho q]}{\lambda} \frac{6-8p+3p^2}{4-6p+4p^2-p^3}$, which is unimodal in p. We can verify that $\Delta W_N(p)|_{p=0} = \frac{3C(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)} > \Delta W_N(p)|_{p=1} = \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$ and $\Delta W_N(p)$ is maximized at $\hat{p}_4$, where $\hat{p}_4 \in (0,1)$ uniquely solves $4 - 24p + 32p^2 - 16p^3 + 3p^4 = 0$ (see the proof of Lemma 2). Therefore, (i) if $P \le \Delta W_N(1) = \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$, we have $p_e = 1$. (ii) If $P > \Delta W_N(\hat{p}_4)$, then never purchasing PTAS is the unique equilibrium, i.e., $p_e = 0$. (iii) If $P \in (\Delta W_N(1), \Delta W_N(0))$, there exists unique equilibrium strategy $p_e \in (\hat{p}_4, 1)$, which uniquely solves $\Delta W_N(p) = P$. (iv) If $P \in [\Delta W_N(0), \Delta W_N(\hat{p}_4))$, there exist one pure equilibrium strategy $p_e^{(1)} = 0$ and two mixed equilibrium strategies $p_e^{(2)} < p_e^{(3)}$, where $p_e^{(2)}$ and $p_e^{(3)}$ are the solutions of $\Delta W_N(p) = P$ in $p \in (0, \hat{p}_4)$ and $p \in (\hat{p}_4, 1)$, respectively. And we can verify that $p_e^{(2)}$ and $p_e^{(3)}$ are not ESS.

(4) When $N \ge 5$, $\Delta W_N(p)$ is decreasing in $p \in [0,1]$. We can verify that $\lim_{p\to 0^+} \Delta W_N(p) = \frac{C(N-1)(1-\rho_1)}{2\lambda_0(1+\rho_0-\rho_1)}$ and $\lim_{p\to 1^-} \Delta W_N(p) = \frac{C(1-\rho_1)}{\lambda_0(1+\rho_0-\rho_1)}$ by using L'Hospital rule. (i) If $\Delta W_N(0) \le P$, then $p_e = 0$. (ii) If $\Delta W_N(1) \ge P$, then $p_e = 1$. (iii) Otherwise, if $P \in (\Delta W_N(1), \Delta W_N(0))$, the unique mixed equilibrium strategy is given by $p_e \in (0,1)$, which uniquely solves $\Delta W_N(p) = P$. And it is naturally identifies the ESS because of its uniqueness. By summarizing the results above, the proof is completed. ■

Proof of Theorem 1. We define the following thresholds for the PTAS fee

$$\underline{P}_1 = \frac{C(1-\rho)}{\Lambda}, \quad \bar{P}_1 = \frac{C(N-1)(1-\rho)}{2\Lambda}, \quad \underline{P}_2 = \frac{C}{\Lambda[(\mu R/C-1)\rho+1]}, \quad \bar{P}_2 = \frac{C(N-1)}{\Lambda[2+\rho(2\mu R/C-N-1)]}, \tag{33}$$

Let $\tilde{p}$, $\bar{p}$ and p' be the unique solutions to equations

$$\frac{CQ_N(\tilde{p})(1-\rho)}{\Lambda(1-\tilde{p})}=P,\quad \frac{CQ_N(\bar{p})}{\Lambda(1-\bar{p})[(\mu R/C-Q_N(\bar{p})-1)\rho+1]}=P,\quad \frac{CQ_N(p')}{\Lambda(1-p')(Q_N(p')+1)}=P. \tag{34}$$

And finally, we define

$$\begin{aligned}
&\rho_s(p)=\frac{Q_N(p)}{\mu(R-pP)/C-Q_N(p)-1},\quad \rho_l(p)=1-\frac{1}{\mu R/C-Q_N(p)-1},\\
&q_{01}=\frac{\mu Q_N(p')}{\Lambda[\mu(R-p'P)/C-Q_N(p')-1]},\quad q_{11}=\frac{\mu}{\Lambda}-\frac{\mu}{\Lambda(\mu R/C-1)},\quad q_{12}=\frac{\mu}{\Lambda}-\frac{\mu}{\Lambda[\mu R/C-Q_N(\bar{p})-1]},\\
&q_{13}=\frac{\mu}{\Lambda}-\frac{2\mu}{\Lambda[2\mu R/C-N-1]},\quad q_{14}=\min\left\{\Lambda,\mu-\frac{\mu}{\mu R/C-Q_N(p')-1}\right\}.
\end{aligned} \tag{35}$$

According to our preassumption that $\mu R/C\geq N+1$, we have $\mu R/C\geq Q_N(p)+2$ for all $p\in[0,1]$, which implies that $q_1^e>0$ by Proposition 2. Notice that $\Delta W_N(p)$ is decreasing in $p\in[0,1]$ when $N\geq 5$ by Lemma 2, then three cases are considered below.

(i) If $q_0^e=q_1^e=1$, by the definitions of $\rho_s(p)$ and $\rho_l(p)$, see Propositions 2-3, we consider the following three subcases.

- $p^e=0$ is the equilibrium if and only if (1) $\rho>\rho_s(0)$; (2) $\rho\leq\rho_l(0)$; and (3) $P>\bar{P}_1=\Delta W_N(0)=\frac{C(N-1)(1-\rho)}{2\Lambda}$.
- $p^e=1$ is the equilibrium if and only if (1) $\rho>\rho_s(1)$; (2) $\rho\leq\rho_l(1)$; and (3) $P\leq\underline{P}_1=\Delta W_N(1)=\frac{C(1-\rho)}{\Lambda}$.
- $p^e\in(0,1)$ is an equilibrium if and only if (1) $\rho>\rho_s(\tilde{p})$; (2) $\rho\leq\rho_l(\tilde{p})$; and (3) $P=\Delta W_N(\tilde{p})\in(\underline{P}_1,\bar{P}_1]$, where $\tilde{p}$ is the unique solution of

$$\frac{CQ_N(\tilde{p})(1-\rho)}{\Lambda(1-\tilde{p})}=P. \tag{36}$$

(ii) If $q_0^e=1>q_1^e$, we consider the following three subcases.

- $p^e=0$ is the equilibrium if and only if (1) $\rho>\rho_s(0)$; (2) $\rho>\rho_l(0)$; and (3) $P>\bar{P}_2=\Delta W_N(0)=\frac{C(N-1)(1-\rho_1^e)}{2\Lambda(1+\rho-\rho_1^e)}$, where $\rho_1^e=1-\frac{2}{2\mu R/C-N-1}$, i.e., $q_1^e=\frac{\mu}{\Lambda}-\frac{2\mu}{\Lambda[2\mu R/C-N-1]}$.
- $p^e=1$ is the equilibrium if and only if (1) $\rho>\rho_s(1)$; (2) $\rho>\rho_l(1)$; and (3) $P\leq\underline{P}_2=\Delta W_N(1)=\frac{C(1-\rho_1^e)}{\Lambda(1+\rho-\rho_1^e)}$, where $\rho_1^e=1-\frac{1}{\mu R/C-1}$, i.e., $q_1^e=\frac{\mu}{\Lambda}-\frac{\mu}{\Lambda(\mu R/C-1)}$.
- $p^e\in(0,1)$ is the equilibrium if and only if (1) $\rho>\rho_s(\bar{p})$; (2) $\rho>\rho_l(\bar{p})$; and (3) $P=\Delta W_N(\bar{p})\in(\underline{P}_2,\bar{P}_2]$, where $\rho_1^e=1-\frac{1}{\mu R/C-Q_N(\bar{p})-1}$, i.e., $q_1^e=\frac{\mu}{\Lambda}-\frac{\mu}{\Lambda[\mu R/C-Q_N(\bar{p})-1]}$, and $\bar{p}$ is the unique solution of

$$\frac{CQ_N(\bar{p})(1-\rho_1^e)}{\Lambda(1-\bar{p})(1+\rho-\rho_1^e)}=P. \tag{37}$$

(iii) If $0\leq q_0^e<1$, then the equilibrium can be determined by $q_0^eU_0(p',q_0^e)=0$, $q_1^e=\max\{q_1|U_1(p',q_1)\geq 0\}$ and $\Delta W_N(p')=P$. Solving the two equations above gives

$$q_0^e=\left(\frac{\mu Q_N(p')/\Lambda}{\mu(R-p'P)/C-Q_N(p')-1}\right)^+\quad\text{and}\quad q_1^e=\min\left\{1,\frac{1}{\rho}-\frac{1}{\rho[\mu R/C-Q_N(p')-1]}\right\}.$$

It is not difficult to verify that $\rho_l(p)$ is increasing in $p\in[0,1]$ because of the monotonicity of $Q_N(p)$. In addition, notice that $\rho_s(p)$ can be rewritten as

$$\rho_s(p)=\frac{Q_N(p)/(1-p)}{(\mu(R-P)/C-1)/(1-p)+\mu P/C-Q_N(p)/(1-p)},$$

where $Q_N(p)/(1-p)$ is decreasing in $p\in[0,1]$ when $N\geq 5$ by Lemma 2. Then $\rho_s(p)$ is decreasing in $p\in[0,1]$. Since $\rho_s(0)<\rho_l(0)\Leftrightarrow \frac{\mu R}{C}>N+1$ (holds naturally due to our preassumption that $\mu R/C\geq N+1$), we have $\rho_s(1)=0<\rho_s(\tilde{p})<\rho_s(0)<\rho_l(0)<\rho_l(\tilde{p})<\rho_l(1)$. On the other hand, we can verify that $\bar{P}_1<\bar{P}_2$ and $\underline{P}_1<\underline{P}_2$. Combining the monotonicity of $Q_N(p)/(1-p)$ and the fact that

$$\frac{CQ_N(p)(1-\rho_1^e)}{\Lambda(1-p)(1+\rho-\rho_1^e)}>\frac{CQ_N(p)(1-\rho)}{\Lambda(1-p)},$$

we can verify that $\bar{p}>\tilde{p}$ by using equations (36)-(37). Combining the three cases above, we can get the two tables in Theorem 1 by comparing $\bar{P}_1$ and $\underline{P}_2$, which completes this proof. ■

Proof of Theorem 2. (1) When $R-P-C/\mu\geq 0$, consider the tagged customer who finds an unavailable server and n customers. If $n=N-1$, she can activate this system if she joins, then she will join if and only if $R\geq CN/\mu$ according to (12). If $n\leq N-2$, her expected utility is given by

$$U_{(0,n)}(a)=\begin{cases} R-P-\frac{(n+1)C}{\mu}, & \text{if } a(0,n)=J_1;\\ R-CT_n-\frac{(n+1)C}{\mu}, & \text{if } a(0,n)=J_0;\\ 0, & \text{if } a(0,n)=B,\end{cases}$$

where T_n is the expected waiting time of the tagged customer before the system is activated. Obviously, we have $T_n\geq 1/\Lambda$. As $P\leq C/\Lambda$, it is optimal for her to purchase the PTAS if she decides to join. Thus she will join and purchase the PTAS if and only if $n\leq\min\{N-2,\lfloor\mu(R-P)/C\rfloor-1\}$. Therefore, the best response of the tagged customer for each system state is given by

(i) If $s=(0,N-1)$, then the customer will join without purchasing if and only if $R\geq NC/\mu$, i.e.,

$$\delta_e(0,N-1)=\begin{cases} J_0, \text{ if } R\geq NC/\mu;\\ B, \text{ if } R<NC/\mu.\end{cases}$$

(ii) If $s=(0,n)$ for $n\leq N-2$, the customer who finds the server on vacation will join and purchase the PTAS if and only if $n\leq\lfloor\mu(R-P)/C\rfloor-1$, i.e.,

$$\delta_e(0,n)=\begin{cases} J_1, \text{ if } n\leq\lfloor\mu(R-P)/C\rfloor-1;\\ B, \text{ if } n>\lfloor\mu(R-P)/C\rfloor-1.\end{cases}$$

In summary, when all customers adopt the best response, i.e., on the equilibrium path, the system degenerates to the M/M/1 queue since it will be activated by the first arriving customer. Combining (12) and the analysis above gives $\delta_e(0,0)=J_1$ and then

$$\delta_e(1,n)=\begin{cases} J_0, \text{ if } 1\leq n\leq\lfloor\mu R/C\rfloor-1;\\ B, \text{ if } n\geq\lfloor\mu R/C\rfloor.\end{cases}$$

(2) When $R-P-C/\mu<0$, the expected utility of the first arriving customer will not join, thus $\delta_e(0,n)=B$ for $n\in\mathbb{N}$. ■

Proof of Theorem 3. When an arriving customer finds the server to be on vacation, we must have $n\leq N-1$, otherwise the server has been activated. We consider the following two cases.

(i) If all customers will join when they find the server is on vacation, i.e., $\delta((0,n)) = J_0$ or $\delta((0,n)) = J_1$ for $n = 0, 1, \ldots, N-1$. Then the purchasing strategy of customers can be described in the following order:

$$\psi(i) = \begin{cases} 1, \text{ if she purchases at } n = i; \\ 0, \text{ if she does not purchase at } n = i. \end{cases}$$

for $i = 1, 2, \ldots, N-1$. Denote by $\psi^*(i)$ the best response of customers at state $(0, i)$. We firstly show that $\psi^*(i) \cdot \psi^*(i+1) = 0$ for all $i = 1, 2, \ldots, N-2$. That is, $\psi^*(i) = 1$ implies that $\psi^*(i-1) = 0$ and $\psi^*(i+1) = 0$. Otherwise, if $\psi^*(i) = \psi^*(i+1) = 1$. Consider the tagged customer who finds i customers upon arrival, by following this strategy, her expected cost is given by $\frac{(i+1)C}{\mu} + P$. If she deviates to not purchase the PTAS, the server would be activated by the next arriving customer, then her expected cost is $\frac{(i+1)C}{\mu} + \frac{C}{\Lambda} > \frac{(i+1)C}{\mu} + P$. That is, by deviating this strategy $\psi^*(i)$, a lower cost can be derived. Thus $\psi(i)$ cannot be the best response, i.e., under equilibrium strategy, we must have $\psi^*(i) \cdot \psi^*(i+1) = 0$ for all $i = 1, 2, \ldots, N-2$. It should be noted that the customer who finds $n = N-1$ would never purchase the PTAS because the server has been activated if she joins. Then it gives $\psi^*(N-1) = 0$. Consider the one who finds $N-2$ and decides to join, her expected cost is given by

$$cost(N-2; \psi^*(N-1) = 0) = \begin{cases} \frac{(N-1)C}{\mu} + P, \text{ if } \psi(N-2) = 1; \\ \frac{(N-1)C}{\mu} + \frac{C}{\Lambda}, \text{ if } \psi(N-2) = 0. \end{cases}$$

Since $\frac{(N-1)C}{\mu} + P > \frac{(N-1)C}{\mu} + \frac{C}{\Lambda}$, it is never for her to purchase PTAS, which gives $\psi^*(N-2) = 0$. Similarly, consider the one who finds $N-3$, her expected cost is given by

$$cost(N-3; \psi^*(N-2) = 0, \psi^*(N-1) = 0) = \begin{cases} \frac{(N-2)C}{\mu} + P, \quad \text{if } \psi(N-3) = 1; \\ \frac{(N-2)C}{\mu} + \frac{2C}{\Lambda}, \text{ if } \psi(N-3) = 0. \end{cases}$$

Based on the definition of I, if $I \geq 2$, we have $\frac{(N-2)C}{\mu} + P > \frac{(N-2)C}{\mu} + \frac{2C}{\Lambda}$, then it gives $\psi^*(N-3) = 0$. Similar to this argument, it follows that $\psi^*(N-1-j) = 0$ for all $1 \leq j \leq I-1$. Consider the one who finds $n = N-(I+1)$, her expected cost is given by

$$cost(N-(I+1); \psi^*(N-1-j) = 0, \ j \in [0, I-1]) = \begin{cases} \frac{(N-I)C}{\mu} + P, \quad \text{if } \psi(N-(I+1)) = 1; \\ \frac{(N-I)C}{\mu} + \frac{IC}{\Lambda}, \text{ if } \psi(N-(I+1)) = 0. \end{cases}$$

Then we must have $\psi^*(N-(I+1)) = 1$ since $\frac{(N-I)C}{\mu} + P \leq \frac{(N-I)C}{\mu} + \frac{IC}{\Lambda}$. Analogically, we can obtain that

$$\psi^*(i) = \begin{cases} 1, \text{ if } \mathrm{mod}(N-1-i, I) = 0 \ ; \\ 0, \text{ if } \mathrm{mod}(N-1-i, I) > 0 \ . \end{cases}$$

for $i = N-2, N-3, \ldots, 0$. Under strategy $\{\psi^*(i), i \in [0, N-2]\}$, the expected cost of customer is given by $u(i)$ (see (13)). Therefore, if $R \geq \max\{u(i)\}$, all customer will join when they find $n \leq N-1$ under the best response. Then the SPE $\delta_e(s)$ for each state s can be characterized accordingly.

(ii) If $R < \max\{u(i)\}$, $\{\psi^*(i), i \in [1, N-1]\}$ cannot be an equilibrium strategy in purchasing PTAS. And the server can only be activated by receiving a PTAS request from customer (otherwise all customers will join, which contradicts to $R < \max\{u(i)\}$). Denote by n_J ($n_J < N-1$) the threshold that customers join if and only if $n \leq n_J$ under equilibrium. Then we must have $\psi^*(n_J) = 1$ (otherwise the server cannot be activated).

Thus it gives $n_J \le \frac{(R-P)\mu}{C} - 1$, which implies that $n_J = \lfloor \frac{(R-P)\mu}{C} \rfloor - 1$. Based on the similar argument in case (i), we can obtain that

$$\psi^*(i) = \begin{cases} 1, \text{ if } \mathrm{mod}(n_J - i, I) = 0 \ ; \\ 0, \text{ if } \mathrm{mod}(n_J - i, I) > 0 \ . \end{cases}$$

for $i = 0, 1, \ldots, n_J$, sequently. Under strategy $\{\psi^*(i), i \in [0, n_J]\}$, we can also obtain the corresponding expected cost of customers:

$$\tilde{u}(i) = \begin{cases} \frac{(i+1)C}{\mu} + P, & \text{if } \mathrm{mod}(n_J - i, I) = 0; \\ \frac{(i+1)C}{\mu} + \frac{\mathrm{mod}(n_J - i, I)C}{\Lambda}, & \text{if } \mathrm{mod}(n_J - i, I) > 0. \end{cases} \tag{38}$$

for $i = 0, 1, \ldots, n_J$. It is not difficult to verify that $R \ge \tilde{u}(i)$ for $i = 0, 1, \ldots, n_J$ since $\tilde{u}(i) \le (n_J + 1)C/\mu + P$ for $i = 0, 1, \ldots, n_J$.

That is, under the equilibrium strategy, all customers who find that $n \le n_J$ will definitely join. Thus, we can fully characterize the SPE $\boldsymbol{\delta}_e$ as follows:

(i) If $R \ge \max\{u(i)\}$, customers join the queue for all $n \le N - 1$ and adopt PTAS if and only if $n = N - 1 - k \cdot I$ for some $k = 1, 2, ..., \lfloor (N-1)/I \rfloor$, i.e., for all $n \le N - 1$,

$$\delta_e(0, n) = \begin{cases} J_1, \text{ if } n = N - 1 - k \cdot I \text{ for some } k = 1, 2, ..., \lfloor (N-1)/I \rfloor; \\ J_0, \text{ if } n \ne N - 1 - k \cdot I \text{ for all } k = 1, 2, ..., \lfloor (N-1)/I \rfloor. \end{cases}$$

(ii) If $R < \max\{u(i)\}$, customers join the queue if and only if $n \le n_J = \lfloor (R-P)\mu/C \rfloor - 1$ and adopt PTAS if and only if $n = n_J - k \cdot I$ for some $k = 1, 2, ..., \lfloor n_J/I \rfloor$, i.e.,

$$\delta_e(0, n) = \begin{cases} J_1, \text{ if } n \le n_J \text{ and } n = n_J - k \cdot I \text{ for some } k = 1, 2, ..., \lfloor n_J/I \rfloor; \\ J_0, \text{ if } n \le n_J \text{ and } n \ne n_J - k \cdot I \text{ for all } k = 1, 2, ..., \lfloor n_J/I \rfloor; \\ B, \text{ if } n > n_J. \end{cases}$$

Similar to Theorem 2, on the equilibrium path, some states will not appear, then the equilibrium strategy of customers is given by

(i) If $R \ge \max\{u(i)\}$,

$$\delta_e(0, n) = \begin{cases} J_0, \text{ if } n \le \mathrm{mod}(N-1, I) - 1; \\ J_1, \text{ if } \mathrm{mod}(N-1, I) = n. \end{cases} \quad \delta_e(1, n) = \begin{cases} J_0, \text{ if } \mathrm{mod}(\bar{n}, I) + 1 \le n \le \lfloor \mu R/C \rfloor - 1; \\ B, \text{ if } \lfloor \mu R/C \rfloor \le n. \end{cases}$$

(ii) If $R < \max\{u(i)\}$,

$$\delta_e(0, n) = \begin{cases} J_0, \text{ if } n \le \mathrm{mod}(n_J, I) - 1; \\ J_1, \text{ if } \mathrm{mod}(n_J, I) = n. \end{cases} \quad \delta_e(1, n) = \begin{cases} J_0, \text{ if } \mathrm{mod}(n_J, I) + 1 \le n \le \lfloor \mu R/C \rfloor - 1; \\ B, \text{ if } \lfloor \mu R/C \rfloor \le n. \end{cases}$$

■

Proof of Theorem 4. (1) When $P \le C/\Lambda$, under equilibrium, the customer who finds an empty system would purchase the PTAS (see Theorem 2). Thus the system degenerates to the classic work-conservation queueing system, which follows that

$$\pi_{0,0}\Lambda = \pi_{1,1}\mu, \quad \pi_{1,i}\Lambda = \pi_{1,i+1}\mu$$

for $i=1,2,\ldots,n_1-1$, $n_1=\lfloor \mu R/C \rfloor$. By solving the equations above, we can obtain the steady-state probabilities. The system throughput is $\lambda_e^o=\Lambda(1-\pi_{0,n_1})=\frac{\Lambda(1-\rho^{n_1})}{1-\rho^{n_1+1}}$. And the revenue of service provider by selling PTAS is $\Pi^o=\Lambda\pi_{0,0}P=\frac{\Lambda P(1-\rho)}{1-\rho^{n_1+1}}$.

(2) When $P>C/\Lambda$, under equilibrium, the system can be activated by the (n_2+1)-th customer, where

$$n_2=\begin{cases}\mathrm{mod}(N-1,I), & \text{if } R\geq \max\{u(i)\};\\ \mathrm{mod}(n_J,I), & \text{if } R<\max\{u(i)\}.\end{cases}$$

by Theorem 3. Then the state transition equations are given by

$$\pi_{0,0}\Lambda=\pi_{1,1}\mu,\quad \pi_{0,i}\Lambda=\pi_{0,i-1}\Lambda,$$
$$\pi_{1,n_2+1}(\Lambda+\mu)=(\pi_{1,n_2}+\pi_{0,n_2})\Lambda+\pi_{1,n_2+2}\mu,\quad \pi_{1,j}\Lambda=\pi_{1,j+1}\mu,$$
$$\pi_{1,1}(\Lambda+\mu)=\pi_{1,2}\mu,\quad \pi_{1,k}(\Lambda+\mu)=\pi_{1,k-1}\Lambda+\pi_{1,k+1}\mu,$$

where $i=1,2,\ldots,n_2$, $j=n_2+2,n_2+3,\ldots,n_1-1$ and $k=1,2,\ldots,n_2$. Then it gives

$$\pi_{0,i}=\pi_{0,0},\quad \pi_{1,1}=\rho\pi_{0,0},$$
$$\pi_{1,k+1}=\pi_{1,k}\rho+\pi_{1,1},\quad \pi_{1,j}=\pi_{1,n_2+1}\rho^{j-n_2-1},$$

where $i=1,2,\ldots,n_2$, $j=n_2+2,n_2+3,\ldots,n_1$ and $k=1,2,\ldots,n_2$. Then we have $\pi_{1,k+1}-\frac{\pi_{1,1}}{1-\rho}=(\pi_{1,1}-\frac{\pi_{1,1}}{1-\rho})\rho^k$, which implies that $\pi_{1,k+1}=\frac{\rho\pi_{0,0}(1-\rho^{k+1})}{1-\rho}$ for $k=0,1,\ldots,n_2$. And it gives $\pi_{1,j}=\frac{\rho^{j-n_2}\pi_{0,0}(1-\rho^{n_2+1})}{1-\rho}$ for $j=n_2+2,n_2+3,\ldots,n_1$. Combining the normalization condition follows that

$$\pi_{0,0}\left(n_2+1+\sum_{k=0}^{n_2}\frac{\rho(1-\rho^{k+1})}{1-\rho}+\sum_{j=n_2+2}^{n_1}\frac{\rho^{j-n_2}(1-\rho^{n_2+1})}{1-\rho}\right)=1,$$

which gives $\pi_{0,0}=\frac{(1-\rho)^2}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$. Then we can get

$$\pi_{0,i}=\frac{(1-\rho)^2}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})},$$
$$\pi_{1,k+1}=\frac{\rho(1-\rho)(1-\rho^{k+1})}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})},$$
$$\pi_{1,j}=\frac{\rho^{j-n_2}(1-\rho)(1-\rho^{n_2+1})}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$$

for $i=0,1,\ldots,n_2$, $k=0,1,\ldots,n_2$ and $j=n_2+2,n_2+3,\ldots,n_1$. The system throughput is $\lambda_e^o=\Lambda(1-\pi_{1,n_1})=\frac{\Lambda[(1-\rho)(n_2+1)+\rho^{n_1}(\rho-\rho^{-n_2})]}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$. The PTAS revenue is given by $\Pi^o=\Lambda P\pi_{0,n_2}=\frac{\Lambda P(1-\rho)^2}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$, which completes this proof. ■

Proof of Theorem 5. When $\rho>\rho_l=1-\frac{1}{\mu R/C-1}$ (i.e., $R<\frac{(2-\rho)C}{\mu}$ if $\rho<1$) and $P<\overline{P}_2=\frac{C}{\Lambda[(\mu R/C-1)\rho+1]}$ (i.e., $R<\frac{\Lambda P-(1-\rho)C}{\Lambda}$), we have $\Pi^u=\frac{\Lambda P(1-\rho_1)}{1+\rho-\rho_1}=\frac{\Lambda P}{(\mu R/C-1)\rho+1}$ by the proof of Theorem 1. Therefore, let

$$\underline{R}=\begin{cases}\min\left\{\frac{(2-\rho)C}{\mu},\frac{\Lambda P-(1-\rho)C}{\Lambda}\right\}, & \text{if } \rho<1;\\ \frac{\Lambda P-(1-\rho)C}{\Lambda}, & \text{if } \rho\geq 1,\end{cases}$$

we have

$$\Pi^u=\frac{\Lambda P(1-\rho_1)}{1+\rho-\rho_1}=\frac{\Lambda P}{(\mu R/C-1)\rho+1}\quad \text{when } R<\underline{R}.$$

On the other hand, the revenue in observable case is given by

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

since $\frac{\Lambda P(1-\rho)^2}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$ is maximized at $n_2=0$. To show $\Pi^o<\Pi^u$ when $R<\underline{R}$, it suffices to prove

$$\frac{\Lambda P}{(\mu R/C-1)\rho+1} > \frac{\Lambda P(1-\rho)}{1-\rho^{n_1+1}}.$$

According to the definition of n_1, we have $\mu R/C-1<n_1$, then it is sufficient to show

$$\frac{1}{n_1\rho+1} \ge \frac{1-\rho}{1-\rho^{n_1+1}}.$$

We consider the following three cases.

(1) If $\rho<1$, then

$$\frac{1}{n_1\rho+1} > \frac{1-\rho}{1-\rho^{n_1+1}} \Leftrightarrow \rho^{n_1}-1-n_1(1+\rho)<0.$$

Since $\rho^{n_1}-1-n_1(1+\rho)$ is decreasing in n_1, we have $\rho^{n_1}-1-n_1(1+\rho)<\rho^0-1-0(1+\rho)=0$, which implies that $\Pi^o<\Pi^u$.

(2) If $\rho=1$, then $\frac{1}{n_1\rho+1}-\frac{1-\rho}{1-\rho^{n_1+1}}\equiv 0$ for any n_1, then we can still have $\Pi^o<\Pi^u$.

(3) If $\rho>1$, then

$$\frac{1}{n_1\rho+1} > \frac{1-\rho}{1-\rho^{n_1+1}} \Leftrightarrow \rho^{n_1}-1-n_1(1+\rho)>0,$$

which increases in $\rho>1$. Since $\lim\limits_{\rho\to 1^+}\rho^{n_1}-1-n_1(1+\rho)=0$, we can get that $\Pi^o<\Pi^u$, which completes this proof. ■

Proof of Theorem 6. By the proof of Theorem 1, we have $\lambda_e^u=\Lambda$ when $\rho<\rho_l(0)\Leftrightarrow\Lambda<\underline{\Lambda}\equiv\mu-\frac{\mu}{\mu R/C-(N+1)/2}$. Notice that there always have some customers balk in observable case due to the endogenous threshold strategy of customers, under which the system throughput satisfies $\lambda_e^o<\Lambda$. It implies that $\lambda_e^u>\lambda_e^o$ when $\Lambda<\underline{\Lambda}$. ■

Proof of Theorem 7. (1) In the unobservable case, by (9), it is directly to derive that

$$\lim_{\Lambda\to 0}\Pi^u(\Lambda)=\frac{(1-\lambda_1^e/\mu)p^e\Lambda P}{1+\rho-\lambda_1^e/\mu}=0.$$

On the other hand, for any $\Lambda>\mu$, we have $R>CN/\mu>C\left(\frac{i}{\mu}+\frac{N-i}{\Lambda}\right)$ for any $i=0,1,\ldots,N$. Then B_0-customers will definitely join, i.e., $q_0^e=1$. Also, in equilibrium, we must have $q_1^e<1$ to ensure the stability of system, thus it gives

$$\Delta W_N(p)=\frac{C\left[1-[1+p(N-1)](1-p)^{N-1}\right](1-\rho_1)}{\lambda_0 p[1-(1-p)^N](1+\rho_0-\rho_1)}<\frac{C\mu\left[1-[1+p(N-1)](1-p)^{N-1}\right]}{p[1-(1-p)^N]\Lambda^2}.$$

Therefore, for any given PTAS fee P, if $\Lambda>\sqrt{\frac{C\mu\max\{\frac{N-1}{2},2\}}{P}}>\max\limits_{p\in[0,1]}\{\sqrt{\frac{C\mu\left[1-[1+p(N-1)](1-p)^{N-1}\right]}{p[1-(1-p)^N]P}}\}$, we have $P>\max\limits_{p\in[0,1]}\{\Delta W_N(p)\}$, i.e., $p^e=0$ is the unique equilibrium. Then we have $\lim\limits_{\Lambda\to\infty}\Pi^u(\Lambda)<\lim\limits_{\Lambda\to\infty}\Lambda\cdot$

$\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda^2}=0$. Thus we can get that $\Pi^u(0)=\Pi^u(\infty)=0$. When $\Lambda\in(0,\mu)$, we can always obtain a positive PTAS revenue by charging a relatively small PTAS fee, i.e., $\Pi^u(\Lambda)>0$ for $\Lambda\in(0,\mu)$, that is, the PTAS revenue is non-monotone in Λ.

(2) In the observable case, when $P\le C/\Lambda$, we have $\Pi^o(\Lambda)=\frac{\Lambda P(1-\rho)}{1-\rho^{n_1+1}}<\Lambda R$, then it gives $\lim\limits_{\Lambda\to 0}\Pi^o(\Lambda)=0$. Also, $\Lambda P\le C$ implies that $\Pi^o(\Lambda)=\frac{\Lambda P(1-\rho)}{1-\rho^{n_1+1}}<\frac{C(1-\rho)}{1-\rho^{n_1+1}}$. Therefore, we conclude that

$$\lim_{\Lambda\to\infty}\Pi^o(\Lambda)<\lim_{\rho\to\infty}\frac{C(1-\rho)}{1-\rho^{n_1+1}}=\lim_{\rho\to\infty}\frac{C}{(n_1+1)\rho^{n_1}}=0.$$

When $P>C/\Lambda$, we have

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

by noticing that $\frac{\Lambda P(1-\rho)^2}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$ is decreasing in n_2. Then it follows that $\lim\limits_{\Lambda\to 0}\Pi^o(\Lambda)=0$ and

$$\lim_{\Lambda\to\infty}\Pi^o(\Lambda)<\lim_{\Lambda\to\infty}\frac{\mu R\rho(1-\rho)}{1-\rho^{n_1+1}}=\lim_{\Lambda\to\infty}\frac{2}{(n_1+1)n_1\rho^{n_1-1}}=0.$$

Therefore, the revenue in the observable case satisfies $\Pi^o(0)=\Pi^o(\infty)=0<\Pi^o(\Lambda)$ for any $\Lambda\in(0,\infty)$, i.e., $\Pi^o(\Lambda)$ is non-monotone in Λ. ■

Proof of Theorem 8. (1) In the unobservable case, we have

$$\Delta W_N(p)=\frac{C\left[1-[1+p(N-1)](1-p)^{N-1}\right](1-\rho_1)}{\lambda_0 p[1-(1-p)^N](1+\rho_0-\rho_1)}\ge\frac{C(1-\Lambda/\mu)\left[1-[1+p(N-1)](1-p)^{N-1}\right]}{p[1-(1-p)^N]\Lambda}$$

since $\lambda_0\le\Lambda$. For any $P<R-C/\mu$, if $\Lambda<\frac{\mu}{\mu(R-C/\mu)/A+1}$, where $A=\min\limits_{p\in[0,1]}\frac{C\left[1-[1+p(N-1)](1-p)^{N-1}\right]}{p[1-(1-p)^N]}=C/2$, we can get

$$P<R-C/\mu<\frac{C(1-\Lambda/\mu)\left[1-[1+p(N-1)](1-p)^{N-1}\right]}{p[1-(1-p)^N]\Lambda}\le\Delta W_N(p)$$

for any $p\in[0,1]$. In equilibrium, it follows that $p^e=1$ for any $P<R-C/\mu$. Recall that when $\Lambda<\min\{\frac{\mu}{\mu(R-C/\mu)/A+1},\underline{\Lambda}\}$, $\Pi^u(\Lambda)=\frac{(1-\lambda_1^e/\mu)p^e\Lambda P}{1+\rho-\lambda_1^e/\mu}$ is increasing in $P\in(0,R-C/\mu)$, where $\lambda_1^e=\min\{\mu-\frac{\mu}{\mu R/C-1},1\}$. That is to say, $P^u(\Lambda)=R-C/\mu$ when $\Lambda<\min\{\frac{\mu}{\mu(R-C/\mu)/A+1},\underline{\Lambda}\}$, which implies that $P^u(0)=R-C/\mu$.

On the other hand, for any $\Lambda>\mu$, by the proof of Theorem 7, for any given PTAS fee P, we have $p^e=0$ if $\Lambda>\sqrt{\frac{C\mu\max\{\frac{N-1}{2},2\}}{P}}\Leftrightarrow P>\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda^2}$, we have $p^e=0$. Therefore, to have a positive PTAS revenue, we must have $P^u(\Lambda)<\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda^2}$, which implies that $P^u(\infty)=0$.

(2) In the observable case, for any $\Lambda>0$, when $P\le C/\Lambda$, the PTAS revenue is $\Pi^o(\Lambda)=\frac{\Lambda P(1-\rho)}{1-\rho^{n_1+1}}$, which increases in $P\in[0,C/\Lambda]$, i.e., $P^o(\Lambda)\ge C/\Lambda$.

Combining the results above, for $\Lambda\ge\mu$, we have $P^u(\Lambda)<\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda^2}$ and $P^o(\Lambda)\ge C/\Lambda$. That is to say, when $\frac{C}{\Lambda}>\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda^2}\Leftrightarrow\Lambda>\bar{\Lambda}=\mu\max\{\frac{N-1}{2},2\}$, we have $P^o(\Lambda)>P^u(\Lambda)$, which completes this proof.

■

Proof of Theorem 9. We just consider the case that $R>CN/\mu$, since when $R\le CN/\mu$, no customer will join in the regular vacation queues.

(1) In the unobservable case, by Proposition 1, the throughput is $\lambda_e^u(p^e)=\frac{\lambda_0(p^e)}{1+\rho_0(p^e)-\rho_1(p^e)}$. Then it suffices to show $q_i^e(p)\geq q_i^e(0)$ for any $p\in(0,1]$, $i=0,1$. By Proposition 2, we have

$$q_1^e(p)>q_1^e(0)\Leftrightarrow\rho_l(p)>\rho_l(0)\Leftrightarrow Q_N(p)\leq(N-1)/2,$$

which holds naturally by the definition of $Q_N(p)$. By Proposition 3, $q_0^e(p)=1$ is an ESS if and only if $\rho>\rho_s(p)$, then it follows that

$$q_0^e(p)>q_0^e(0)\Leftrightarrow\rho_s(p)<\rho_s(0).$$

By the proof of Theorem 1, we have that $\rho_s(p)$ is decreasing in $p\in[0,1]$, which gives $\rho_s(p)<\rho_s(0)$.

(2) In the observable case, the throughput by providing PTAS is $\lambda_e^o=\frac{\Lambda[(1-\rho)(n_2+1)+\rho^{n_1}(\rho-\rho^{-n_2})]}{(1-\rho)(n_2+1)+\rho^{n_1+1}(\rho-\rho^{-n_2})}$, while the throughput in regular vacation queues is $\lambda_c^o=\frac{\Lambda[(1-\rho)(N+1)+\rho^{n_1}(\rho-\rho^{-N})]}{(1-\rho)(N+1)+\rho^{n_1+1}(\rho-\rho^{-N})}$ by Guo and Hassin (2011). It is not difficult to verify that λ_e^o is decreasing in n_2. Thus, $\lambda_e^o\geq\lambda_c^o$ if and only if $n_2\leq N$. By the definition of n_2 that

$$n_2=\begin{cases}\mathrm{mod}(N-1,I), & \text{if } R\geq\max\{u(i)\};\\ \mathrm{mod}(n_J,I), & \text{if } R<\max\{u(i)\},\end{cases}$$

it follows that $n_2<N$ immediately. ■

Proof of Theorem 10. We first establish the benchmark of non-preemptive priority queues when the server's state is disclosed. Denote by F the priority fee and Λ the potential arrival rate. And let the effective arrival rate of priority and ordinary customers be λ_1^p and λ_1^o, respectively. Then the steady-state probability that the server is inactive is given by

$$\pi_0=\frac{1-\rho_1}{1-\rho_1+\rho},$$

where $\rho=\Lambda/\mu$ and $\rho_1=(\lambda_1^o+\lambda_1^p)/\mu$. When the server is found to be inactive, all arriving customer will join without purchasing priority since their expected utility is $R-C/\mu>0$ and they will not be preempted by future priority customers. When the server is active, one can check that the average queue length is

$$L=\frac{\rho}{(1-\rho_1)(1-\rho_1+\rho)}=\pi_0\cdot0+\pi_1\cdot L_1,$$

where $\pi_1=1-\pi_0$ is the steady-state probability that the server is active, and L_1 is the average queue length (including ordinary and priority customers) when the server is active. Thus it gives

$$L_1=\frac{1}{1-\rho_1}=\frac{\mu}{\mu-\lambda_1^o-\lambda_1^p}.$$

Since the priority customers own absolute priorities over the ordinary customers arrive at system when the server is active, they will not be affected by the ordinary ones, the expected queue length of priority queue (excluding the one in service) when the server is active is $L_1^p=\frac{\lambda_1^p}{\mu-\lambda_1^p}$. Thus, the expected queue length for ordinary queue when the server is active is $L_1^o=(L_1-1)-L_1^p=\frac{\lambda_1^o\mu}{(\mu-\lambda_1^p)(\mu-\lambda_1^p-\lambda_1^o)}$ (excluding the one in service). By Little's law, the expected waiting time in the queue for priority and ordinary customers who find an active server are given by

$$w_1^p=\frac{L_1^p}{\lambda_1^p}=\frac{1}{\mu-\lambda_1^p},\quad w_1^o=\frac{L_1^o-1}{\lambda_1^o}=\frac{\mu}{(\mu-\lambda_1^p)(\mu-\lambda_1^p-\lambda_1^o)}.$$

Thus, under the optimal price fee P, in equilibrium, we have

$$R-C(w_1^o+1/\mu)\geq 0,\quad R-F-C(w_1^p+1/\mu)\geq 0,\quad F=C(w_1^o-w_1^p)=\frac{C(\lambda_1^p+\lambda_1^o)}{(\mu-\lambda_1^p)(\mu-\lambda_1^p-\lambda_1^o)}.$$

And the revenue collected by selling priority is given by $\Pi^p(\Lambda)=\lambda_1^p\pi_1 F$. Next, we consider the following two cases to compare the revenue by selling PTAS and priorities.

(1) When Λ is relatively small, by Theorem 8, when $\Lambda<\min\{\frac{\mu}{2\mu R/C+1},\underline{\Lambda}\}$, we have $P^u=R-C/\mu$, $\lambda_e^0=\Lambda$ and $p_e=1$. If $U_1(1,\rho)=R-\frac{C}{\mu}\left[\frac{1}{1-\rho}+1+Q_N(1)\right]>0\Leftrightarrow\Lambda<\mu\left(1-\frac{1}{\mu R/C-1}\right)$, then we have $\lambda_e^1=\Lambda$, and the PTAS revenue is given by $\Pi^u(\Lambda)=(1-\Lambda/\mu)\Lambda(R-C/\mu)$. On the other hand, in priority queues, when all customers join the system in equilibrium, we have $F=\frac{C\Lambda}{(\mu-\Lambda)(\mu-\lambda_1^p)}$, then $\Pi^p(\Lambda)=\frac{C\Lambda\pi_1\lambda_1^p}{(\mu-\Lambda)(\mu-\lambda_1^p)}\leq\frac{C\Lambda^2}{(\mu-\Lambda)^2}$. Notice that $\lambda_1^p=\Lambda$ is an equilibrium if and only if $\frac{C}{\mu-\Lambda}+\frac{C\Lambda}{(\mu-\Lambda)^2}<R\Leftrightarrow\Lambda<\mu-\sqrt{C\mu/R}$. Therefore, when $\Lambda<\min\left\{\frac{\mu}{2\mu R/C+1},\underline{\Lambda},\mu\left(1-\frac{1}{\mu R/C-1}\right),\mu-\sqrt{C\mu/R}\right\}$, we have

$$\Pi^u(\Lambda)=\Lambda\left(1-\frac{\Lambda}{\mu}\right)\left(R-\frac{C}{\mu}\right),\quad \Pi^p(\Lambda)\leq\frac{C\Lambda^2}{(\mu-\Lambda)^2}.$$

We can verify that $\Pi^u(\Lambda)>\Pi^p(\Lambda)$ if

$$R>\frac{C}{\mu}\left[\frac{\rho}{(1-\rho)^3}+1\right]\Leftrightarrow\Lambda<\hat{\rho}\mu,$$

where $\hat{\rho}$ uniquely solves $R=\frac{C}{\mu}\left[\frac{\rho}{(1-\rho)^3}+1\right]$. Let $\underline{\Lambda}'\equiv\min\left\{\frac{\mu}{2\mu R/C+1},\underline{\Lambda},\mu\left(1-\frac{1}{\mu R/C-1}\right),\mu-\sqrt{C\mu/R},\hat{\rho}\mu\right\}$, we have $\Pi^u(\Lambda)>\Pi^p(\Lambda)$ for all $\Lambda<\underline{\Lambda}'$.

(2) When $\Lambda>\mu$, by the proof of Theorems 7-8, for any fixed $P<R-C/\mu$, we have $P^u(\Lambda)<\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda^2}$, which implies that $\Pi^u(\Lambda)<\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda}$. In the pay-for-priority system, in equilibrium, we have

$$F=\frac{C\lambda}{(\mu-\lambda)(\mu-\lambda_1^p)},\quad R=F+\frac{C}{\mu-\lambda_1^p}+\frac{C}{\mu},$$

where λ_1^p and λ are the effective arrival rates of priority customers and total customers, respectively, when the server is active. Consider the equilibrium that $\lambda=\lambda_1^p$, i.e., all customers purchase priority, we can have $R=\frac{C\mu}{(\mu-\lambda)^2}+\frac{C}{\mu}$, which gives $\lambda=\mu-\sqrt{C\mu/(R-C/\mu)}$. Then the priority revenue is given by $\lambda_1^p\pi_1F=\frac{C\rho(1-\sqrt{C/(\mu R-C)})^2}{(\sqrt{C/(\mu R-C)}+\rho)C/(\mu R-C)}\geq\frac{C(1-\sqrt{C/(\mu R-C)})^2}{(\sqrt{C/(\mu R-C)}+1)C/(\mu R-C)}$, which it is independent of the potential arrival rate Λ. Therefore, we can get that $\Pi^p(\Lambda)\geq\frac{C(1-\sqrt{C/(\mu R-C)})^2}{(\sqrt{C/(\mu R-C)}+1)C/(\mu R-C)}>\frac{C\mu\max\{\frac{N-1}{2},2\}}{\Lambda}>\Pi^u(\Lambda)$ if and only if $\Lambda>\bar{\Lambda}'\equiv\frac{C\mu(\sqrt{C/(\mu R-C)}+1)\max\{\frac{N-1}{2},2\}}{(\mu R-C)(1-\sqrt{C/(\mu R-C)})^2}$, which completes this proof.

■

Proof of Lemma 3 (1) Since $Q_N(p)$ is decreasing in p by Lemma 1 and $\overline{w}(p,q)=\frac{1}{\mu-\lambda}+\frac{Q_N(p)}{\lambda}$, which is obviously decreasing in p. (2) Taking the derivative of $\overline{w}(p,q)$ with respect to q gives

$$\frac{d\overline{w}(p,q)}{dq}=-\frac{1}{\Lambda q^2}\cdot\frac{(1-p)[1+(N-1)(1-p)^N-N(1-p)^{N-1}]}{p[1-(1-p)^N]}+\frac{1}{(\mu-\lambda)^2}$$
$$\frac{d^2\overline{w}(p,q)}{dq^2}=\frac{2}{\Lambda q^3}\cdot\frac{(1-p)[1+(N-1)(1-p)^N-N(1-p)^{N-1}]}{p[1-(1-p)^N]\lambda}+\frac{2}{(\mu-\lambda)^3}>0.$$

Thus $\overline{w}(p,q)$ is strictly convex in λ. In particular, if $p=1$, we have $\overline{w}(p,q)=\frac{1}{\mu-\lambda}$, which is strictly increasing in λ.

When $p=1$, $\overline{w}(p,q)$ is minimized at $\hat{q}=0$. Otherwise, if $p>0$, the minimum of $\overline{w}(p,q)$ can be attained at the unique extreme point that satisfying $\frac{d\overline{w}(p,q)}{d\lambda}=0$, i.e., $\frac{1}{(\mu-\lambda)^2}-\frac{1}{\lambda^2}\cdot\frac{(1-p)[1+(N-1)(1-p)^N-N(1-p)^{N-1}]}{p[1-(1-p)^N]}=0$. By solving this equation, we get $\hat{q}=\frac{\mu\sqrt{Q_N(p)}}{\Lambda(1+\sqrt{Q_N(p)})}$, where $Q_N(p)=\frac{(1-p)[1+(N-1)(1-p)^N-N(1-p)^{N-1}]}{p[1-(1-p)^N]}$. Also, it is not difficult to verify that $\lim\limits_{p\to 1}\frac{\mu\sqrt{Q_N(p)}}{\Lambda(1+\sqrt{Q_N(p)})}=0$, i.e., $\hat{q}=0$ when p approaches to 1. Since $\frac{dQ_N(p)}{dp}\le 0$, it follows that $\frac{d\hat{q}}{dQ_N(p)}\cdot\frac{dQ_N(p)}{dp}\le 0$, i.e., $\hat{q}$ is decreasing in $p\in[0,1]$, which completes this proof. ■

Proof of Theorem 11 We consider the following two case according to the value of q^e.

Case (1) When $q^e=1$ (must have $\rho<1$):

(a) $(0,1)$ is an equilibrium if and only if $\Delta w_N(0,1)\le P$ and $R-C\overline{w}(0,1)\ge 0\Leftrightarrow R\ge\frac{C}{\mu-\Lambda}+\frac{C(N-1)}{2\Lambda}$.

(b) $(p^e,1)$ (where $p^e\in(0,1)$) is an equilibrium if and only if $\Delta w_N(p^e,1)=P$ and $R-C\overline{w}(p^e,1)-p^eP\ge 0\Leftrightarrow R-p^eP-\frac{C}{\mu-\Lambda}-\frac{CQ_N(p^e)}{\Lambda}\ge 0$. That is, $\Lambda\le\frac{\mu(R-P)-C}{R-p^eP}$, where p^e solves $\Delta w_N(p^e,1)=P$.

(c) $(1,1)$ is an equilibrium if and only if $\Delta w_N(1,1)\ge P$ and $R-P-\frac{C}{\mu-\Lambda}\ge 0\Leftrightarrow\Lambda\le\mu-\frac{C}{R-P}$.

Case (2) When $q^e<1$:

(a) $(0,q^e)$ is an equilibrium if and only if $\Delta w_N(0,q^e)\le P$ and q^e is determined by $R-C\overline{w}(0,q^e)=0\Leftrightarrow R=\frac{C}{\mu-\Lambda q^e}+\frac{C(N-1)}{2\Lambda q^e}$. It should be noted that $R=\frac{C}{\mu-\Lambda q^e}+\frac{C(N-1)}{2\Lambda q^e}$ has at most two solutions, and only the larger one is an ESS. Then it gives $q^e=\frac{\sqrt{C^2(N-3)^2-4C\mu(N+1)R+4\mu^2R^2}+C(N-3)+2\mu R}{4\Lambda R}$.

(b) (p^e,q^e) (where $p^e\in(0,1)$) is an equilibrium if and only if $\Delta w_N(p^e,q^e)=P$ and $R-C\overline{w}(p^e,q^e)-p^eP=0\Leftrightarrow R-p^eP-\frac{C}{\mu-\Lambda q^e}-\frac{CQ_N(p^e)}{\Lambda q^e}=0$. That is, $\Lambda q^e=\frac{\mu(R-P)-C}{R-p^eP}$ and $\Delta w_N(p^e,q^e)=P$.

(c) $(1,q^e)$ is an equilibrium if and only if $\Delta w_N(1,q^e)\ge P$ and $R-P-\frac{C}{\mu-\Lambda q^e}=0\Leftrightarrow q^e=\frac{\mu-\frac{C}{R-P}}{\Lambda}$.

By combining all the results above, we can complete this proof. ■

Proof of Proposition 5. In no-information case, denote by q^c the equilibrium joining probability in regular vacation queues, it suffices to prove $q^e\ge q^c$.

(1) When $q^c=0$, i.e., the system cannot be activated, we must have $q^e\ge q^c$.

(2) When $0<q^c<1$, it satisfies $U_{NI}(0,q^c)=0$. If $q^e=1$, then it immediately follows that $q^e>q^c$. Otherwise, consider the equilibrium $(1,q^e)$, which satisfies $R-P-C\bar{w}(1,q^e)=0$. Notice that

$$q^e\ge q^c\Leftrightarrow R-P-C\bar{w}(1,q^c)\ge R-C\bar{w}(0,q^c)=0\Leftrightarrow P\le\frac{\Delta w_N(0,q^c)}{1-\Lambda q^c/\mu}.$$

Thus, if $P\le\frac{\Delta w_N(0,q^c)}{1-\Lambda q^c/\mu}$, the equilibrium $(1,q^e)$ can induce a higher throughput. On the other hand, if $P>\frac{\Delta w_N(0,q^c)}{1-\Lambda q^c/\mu}$, we will show that $(0,q^c)$ is also an equilibrium when PTAS is introduced, then it also induces that $q^e\ge q^c$. When all others adopt strategy $\boldsymbol{\alpha}=(0,q^c)$, assume that the tagged customer adopts strategy $\boldsymbol{\alpha}'=(p',q')$, her expected utility is given by

$$\widehat{U}_{NI}(\boldsymbol{\alpha}';\boldsymbol{\alpha})=q'\left[R-p'P-C\overline{w}(0,q^c)-p'\Delta w_N(0,q^c)\right]=-p'q'\left[P-\Delta w_N(0,q^c)\right]<0.$$

Then the best response of the tagged customer is $p'=0$. It follows that $(0,q^c)$ is also an equilibrium.

(3) When $q^c=1$, then $U_{NI}(0,1)>0$. If $q^e<1$, then $U_{NI}(p^e,q^e)=0$, which is Pareto-nominates by strategy $(0,1)$. Thus the customers can shift their strategy from (p^e,q^e) to $(0,1)$ to improve their expected utilities, which also induces $q^e=q^c=1$. ■

Proof of Proposition 6. In fully unobservable case, for any $\Lambda \geq \mu$, we must have $q^e < 1$ to ensure the stability of system, thus it gives

$$\Delta w_N(p^e, q^e) = \frac{C\left[1-[1+p^e(N-1)](1-p^e)^{N-1}\right](1-\rho q^e)}{\Lambda q^e p^e[1-(1-p^e)^N]}.$$

Since $p^e = 0$ and $p^e = 1$ cannot be optimal for service provider ($p^e = 0$ is not optimal for the firm because the revenue would be zero; $p^e = 1$ is not optimal, either, because the firm can always strictly improve its revenue by increasing the price). In equilibrium, we must have $\Delta w_N(p^e, q^e) = P$ and $R - C\overline{w}(p^e, q^e) - p^e P = 0 \Leftrightarrow R - p^e P - \frac{C}{\mu - \Lambda q^e} - \frac{CQ_N(p^e)}{\Lambda q^e} = 0$. That is, $\Lambda q^e = \frac{\mu(R-P)-C}{R-p^e P}$ and $\Delta w_N(p^e, q^e) = P$. That is to say, p_e is determined by

$$\Delta w_N(p^e, q^e) = P \Leftrightarrow \frac{C\left[1-[1+p^e(N-1)](1-p^e)^{N-1}\right]\left[(1-p^e)P + C/\mu\right]}{(\mu(R-P)-C)p^e[1-(1-p^e)^N]} = P.$$

It follows that p^e is independent of Λ. Then the revenue via PTAS is given by

$$\Pi^{NI}(\Lambda) = \frac{\mu(R-P)-C}{R-p^e P}\left(1 - \frac{\mu(R-P)-C}{R-p^e P}\right)p^e P > 0,$$

which is also independent of Λ. Thus, we have $\Pi^{NI}(\Lambda) = \Pi^{NI}(\mu) > 0$ for all $\Lambda \geq \mu$. Recall that $\lim_{\Lambda\to\infty} \Pi^u(\Lambda) = \Pi^o(\Lambda) = 0$. There must exists a sufficiently large $\tilde{\Lambda}'$ such that $\Pi^{NI}(\Lambda) > \Pi^u(\Lambda)$ and $\Pi^{NI}(\Lambda) > \Pi^o(\Lambda)$ when $\Lambda > \tilde{\Lambda}'$. ■

Appendix C: Equilibrium Strategies with $N = 2, 3, 4$ and $N \geq \mu R/C$

We hereby supplement Theorem 1 by establishing the equilibrium strategies in the unobservable case for N not satisfying conditions as required in Theorem 1.

Proposition 8 *Consider the unobservable $M/M/1$ vacation queue with PTAS. When $N = 2, 3, 4$ or $N \geq \mu R/C$, the joint equilibrium strategy is given below:*

$$\mathcal{E} = \begin{cases} (1,1,1), & \text{if } \rho_s(1) < \rho \leq \rho_l(1) \text{ and } P \leq m(\Lambda,\Lambda); \\ (\tilde{p},1,1), & \text{if } \rho_s(\tilde{p}) < \rho \leq \rho_l(\tilde{p}) \text{ and } m(\Lambda,\Lambda) < P \leq M(\Lambda,\Lambda); \\ (0,1,1), & \text{if } \rho_s(0) < \rho \leq \rho_l(0) \text{ and } P > M(\Lambda,\Lambda); \\ (1,1,q_{11}), & \text{if } \rho > \rho_l(1) \text{ and } P \leq m(\Lambda, q_{11}\Lambda); \\ (\bar{p},1,q_{12}), & \text{if } \rho > \rho_l(\bar{p}) \text{ and } m(\Lambda, q_{12}\Lambda) < P \leq M(\Lambda, q_{12}\Lambda); \\ (0,1,q_{13}), & \text{if } \rho > \rho_l(0) \text{ and } P > M(\Lambda, q_{13}\Lambda); \\ (p', q_{01}, q_{14}), & \text{otherwise,} \end{cases}$$

where $M(\lambda_0,\lambda_1) \equiv \max_p \Delta W_N(p)$ and $m(\lambda_0,\lambda_1) \equiv \min_p \Delta W_N(p)$. $\tilde{p}, \bar{p}, p'$ and q_{i1} for $i = 0,1,2,3,4$ are defined in the proof of Theorem 1.

Proof. When $N = 2, 3, 4$ or $N \geq \mu R/C$, $\Delta W_N(p)$ is not necessarily decreasing in $p \in [0,1]$ by Lemma 2, thus for any fixed pair (λ_0, λ_1), we let $M(\lambda_0,\lambda_1) = \max_p \Delta W_N(p)$ and $m(\lambda_0,\lambda_1) = \min_p \Delta W_N(p)$ to characterize the equilibrium. Similar to Theorem 1, three cases are considered below:

(i) If $\lambda_0^e = \lambda_1^e = \Lambda$, we have the following three subcases.

- $p_e = 0$ is the equilibrium if and only if (1) $\rho > \rho_s(0)$; (2) $\rho \le \rho_l(0)$; and (3) $P > M(\Lambda, \Lambda)$.
- $p_e = 1$ is the equilibrium if and only if (1) $\rho > \rho_s(1)$; (2) $\rho \le \rho_l(1)$; and (3) $P \le m(\Lambda, \Lambda)$.
- $p_e \in (0,1)$ is an equilibrium if and only if (1) $\rho > \rho_s(\tilde{p})$; (2) $\rho \le \rho_l(\tilde{p})$; and (3) $m(\Lambda, \Lambda) < P \le M(\Lambda, \Lambda)$, any $\tilde{p}$ satisfies

$$\frac{CQ_N(\tilde{p})(1-\rho)}{\Lambda(1-\tilde{p})} = P. \tag{39}$$

is a mixed equilibrium.

(ii) If $\lambda_0^e = \Lambda > \lambda_1^e$, we consider the following three subcases.

- $p_e = 0$ is the equilibrium if and only if (1) $\rho > \rho_s(0)$; (2) $\rho > \rho_l(0)$; and (3) $P > M(\Lambda, \lambda_1^e)$, where $\rho_1^e = 1 - \frac{2}{2\mu R/C - N - 1}$, i.e., $\lambda_1^e = \frac{\mu}{\Lambda} - \frac{2\mu}{\Lambda[2\mu R/C - N - 1]}$.
- $p_e = 1$ is the equilibrium if and only if (1) $\rho > \rho_s(1)$; (2) $\rho > \rho_l(1)$; and (3) $P \le m(\Lambda, \lambda_1^e)$, where $\rho_1^e = 1 - \frac{1}{\mu R/C - 1}$, i.e., $\lambda_1^e = \frac{\mu}{\Lambda} - \frac{\mu}{\Lambda(\mu R/C - 1)}$.
- $p_e \in (0,1)$ is the equilibrium if and only if (1) $\rho > \rho_s(\bar{p})$; (2) $\rho > \rho_l(\bar{p})$; and (3) $m(\Lambda, \lambda_1^e) < P \le M(\Lambda, \lambda_1^e)$, where $\rho_1^e = 1 - \frac{1}{\mu R/C - Q_N(\bar{p}) - 1}$, i.e., $\lambda_1^e = \frac{\mu}{\Lambda} - \frac{\mu}{\Lambda[\mu R/C - Q_N(\bar{p}) - 1]}$, and $\bar{p}$ satisfies

$$\frac{CQ_N(\bar{p})(1-\rho_1^e)}{\Lambda(1-\bar{p})(1+\rho-\rho_1^e)} = P. \tag{40}$$

(iii) If $\lambda_0^e < \Lambda$, the equilibrium can be determined by $U_0(p', \lambda_0^e/\mu) = 0$, $\lambda_1^e = \max\{\lambda | U_1(p, \lambda_1^e/\mu) \ge 0\}$ and $\Delta W_N(p') = P$, which gives $\lambda_0^e = \frac{\mu Q_N(p')}{\Lambda[\mu(R - p'P)/C - Q_N(p') - 1]}$ $\lambda_1^e = \min\left\{\Lambda, \mu - \frac{\mu}{\mu R/C - Q_N(p') - 1}\right\}$.

In summary, when $N = 2, 3, 4$ or $N \ge \mu R/C$, the joint equilibrium can be obtained in Proposition 8. ■

Appendix D: Additional Discussions on the Observable Case

It should be noted that in a standard M/M/1 observable queues, the Naor-threshold joining strategy is not only an equilibrium strategy, but also a dominant strategy. That is, the best response of a tagged customer is unaffected by other customers' actions; this is true even when their actions deviate from the equilibrium strategy. By contrast, in our PTAS model, a tagged customer's best response will be altered when some other customers do not follow the equilibrium strategy; and such an impact leads to a cyclic strategy structure in the system state (see the proof of Theorem 3). Specifically, in case customers seeing $mod(\bar{n}, I)$ customers do not purchase PTAS, the best response of future customers is to join without paying for PTAS until there are in total $mod(\bar{n}, I) + I$ customers in the system. This is a major distinction from the standard model in Naor (1969). It should be noted that the SPE characterized in Theorems 2–3 is unique because a customer's best response (in the pure strategy case) is unique for all states.

To illustrate how the SPE reduces to a threshold-type strategy, we consider a numerical experiment with $\mu = 2$, $N = 10$, $R = 7$, $P = 1.5$, $\Lambda = 1.5$, $C = 1$, $n_J = \lfloor (R - P)\mu/C \rfloor - 1 = 13$, $\bar{n} = 9$. In Figure 12 we describe customers' best responses when in different system states (panel (a)) and the induced SPE on the equilibrium path (panel (b)). When $n < N$, we only discuss the best responses of those finding the server to be inactive. As depicted in panel (a), the joining strategy is of a threshold type, and the best response of PTAS purchasing strategy is cyclic with a cycle $I = \lceil \Lambda P/C \rceil = 3$. We explain in the backward order of all arrival customers:

- An arrival finding an inactive server and $N - 1$ existing customers will automatically activate the server and has no incentive to purchase PTAS.

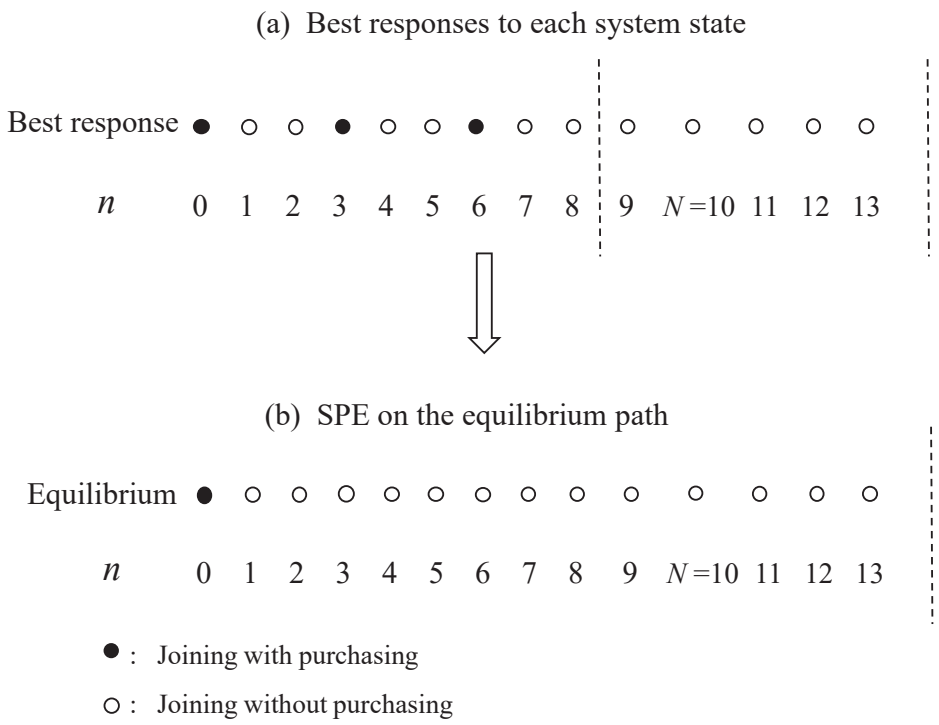


Figure 12 **An illustration of the SPE in an observable vacation queue with PTAS when** $\mu = 2$**,** $N = 10$**,** $R = 7$**,** $P = 1.5$**,** $C = 1$ **and** $\Lambda = 1.5$**,** $\bar{n} = 9$**,** $n_J = 13$**.**

• A customer seeing n existing customers satisfying $N-1-I < n \leq N-2$ will not adopt PTAS either, because her expected delay disutility (i.e., waiting cost for future customers to activate the service) is lower than P.

• A customer finding $n = N-1-I$ existing customers will have to pay for PTAS because the expected waiting cost until the server is activated by future customers is $IC/\Lambda > 1.5 = P$.

• Similarly, those observing a queue length $n = N-2-I$ will not adopt PTAS because she "knows" that the server will be activated by the next customer arrival (an arrival observing a queue length $n = N-1-I$). Best responses for customers observing a queue length $n \in \{N-1-2I, N-2I, \ldots, N-2-I\}$ can be obtained in a similar way.

Following this analysis, the best response of the first arriving customer is to purchase PTAS. But when all customers adopt the best responses (See panel (a) of Figure 12), states state $(0,1), \ldots, (0, N-1)$ will not occur. Therefore, on an equilibrium path, the SPE reduces to the simple threshold-type form (see panel (b) of Figure 12).