Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

Strategic Behavior and Social Optimization in Markovian Vacation Queues

Pengfei Guo
Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hung Hom, Hong Kong, lgtpguo@polyu.edu.hk

Refael Hassin
Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel, hassin@post.tau.ac.il

We consider a single server queueing system in which service shuts down when there are no customers present, and is resumed only when the queue length reaches a given critical length. We analyze the strategic response of customers to this mechanism and compare it to the overall optimal behavior, with and without information on delay. The results are significantly different from those obtained when the server is continuously available. We show that there may exist multiple equilibria in such a system and the optimal arrival rate may be greater or smaller than that of the decentralized equilibrium. Finally, the critical length is taken as a decision variable and the optimal operations policy is discussed by taking strategic customers into consideration.

Key words: vacation queue, strategic customers, balking queue, equilibrium analysis

1. Introduction

Customers who arrive at a queueing system can respond strategically to delay by deciding whether they wish to *join* or *balk*, maximizing their individual welfare. Such decentralized behavior results in an *equilibrium* arrival pattern. From the viewpoint of society as a whole, it is well known that this equilibrium may be suboptimal. When a customer joins a first-come first-served queue, his/her decision does not affect earlier arrivals, but may increase the delay for future arrivals. This effect is called *negative externality*. Social welfare maximization takes such effects into consideration and a toll can be used by social planners to induce optimal decentralized behavior.

In some systems, an increase in congestion may actually benefit the customers. Our daily experiences exemplify this; for example, a shuttle may only leave after all of the seats are occupied. Consequently, a shuttle passenger may anxiously wait for more passengers to arrive. When we wait in a bank lineup, a long queue may induce the opening of additional service counters and thus,

the queue will move progressively faster. We may also have similar experiences at check out in a grocery store. In the border-crossing system, where both security and customer satisfaction need to be considered, managers usually use *congestion-based staffing policy* (see Zhang, 2006) where some extra inspection servers are opened when the size of a queue is larger than the upper threshold level and closed when the size drops to a lower threshold level.

This phenomenon also exists in service systems with unobservable queues. For example, many call centers adopt call blending systems in which an agent can make both inbound and outbound calls. Inbound calls to a call center may have to wait for the agent to switch back to service after finishing an outbound call. A longer inbound call queue can stimulate the server to switch back faster and thus benefit the customers in that queue. (In some call centers, customers can be provided with information on the real-time delay such as queue length or expected waiting time. The queue in those types of call centers is still 'observable'.) In a combination of make-to-stock and make-to-order systems, the facility is not switched back to process customer orders until the number of orders reaches a critical level. More examples can be found in a survey on polling systems by Takagi (2000).

Such service systems belong to a broad class of queues called *vacation queues* or *queues with removable servers*. The book of Tian and Zhang (2006) is devoted to such systems and their applications. Although there exist multiple settings for such service systems, our focus here is on the analysis of the joining behavior of customers in such systems. Therefore, we have focused on a relatively simple one: a vacation queue with an N-policy and *exhaustive service*; that is, the server starts to work when the queue reaches a size N and once the server starts to work, it finishes all the work in the system before taking its next 'vacation'. With this device, we can derive many insights on the decentralized behavior and social optimal requirements of customers. Our results shed light on decentralized customer behavior and social optimization in general systems where congestion could be beneficial for customers.

There exist fundamental differences between a regular and vacation queue with an N-policy. When a customer arrives and sees the server on vacation, s/he must consider both the customers

waiting in front of him/her and future arrivals. Therefore, s/he may prefer to join a longer queue, in anticipation that the service will start sooner. Moreover, a customer may be more compelled to join a queue with an idle server if s/he knows that *future* arrivals have a higher tendency to join the queue when the server is idle. This type of behavior is dubbed by Hassin and Haviv (1997) as *follow the crowd* (FTC), which is in contrast to *avoid the crowd* (ATC), a common behavior in which a customer tries to avoid others and finds it increasingly less attractive to join as the number of people who join increases. Several examples of FTC behavior are described by Hassin and Haviv (2003). Our model is special in that both types of behaviors exist in our system: It may be FTC in some states (namely, those with an idle server) and ATC in others. The result is an unusual pattern of behavior, both in the decentralized case and under social optimization. It is interesting to note that in the work of Hassin and Haviv (1997), the number of equilibrium solutions is in general unbounded, whereas in our model, there can only be a small number of solutions, due to the existence of the ATC region.

By joining a queue with an idle server, the customer shortens the time that current customers need to wait for the server to start working and hence *reduces* their waiting time. Therefore, *positive externalities* exist in such a system. Moreover, joining at a busy state can bring both negative and positive externalities to future customers. On the one hand, they might have to wait as a joining customer receives service, but on the other hand, this act keeps the server busy for a longer time and increases the chances for future customers to avoid an idle server. Multiple equilibria may exist because the best response function of a player may increase with the actions of other players (see Hassin and Haviv 1997, 2003). A socially optimal arrival rate could be larger or smaller than the equilibrium arrival rate. We note that some of these features are common to our model and the *shuttle model* in Hassin and Haviv (2003) §1.5. The shuttle can be viewed as a server with vacations and instantaneous service.

We first study customer equilibrium behavior in a vacation queue. We assume that customers are identical and consider two cases of information availability. With *no information*, the queue is *unobservable* and customers have no real-time information on the status of the server. The

anticipation of waiting time is based on their long-term experience. With *full information*, the queue length and the status of the server are *observable* to the arriving customers. We will also study the socially optimal behavior for both cases. Finally, we compare the decentralized equilibrium with a socially optimal solution.

With no information, the expected waiting time first decreases with the arrival rate and then, after a certain level, increases. Therefore, both FTC and ATC behaviors exist, which depend on the heaviness of the traffic flow. Consequently, multiple equilibria exist; among which, some are stable and others are not. With full information, customers adopt a threshold strategy to join the system. We obtain closed-form expressions for the equilibrium thresholds.

We show that there exists a unique optimal arrival rate for the social optimization problem in the no-information case. Due to the existence of positive externality, the socially optimal arrival rate could be larger than the equilibrium rates. There is an interesting difference between the observable and unobservable models. In the former, it may happen that the socially optimal solution requires that some customers join even when their expected utility is negative. This cannot happen in the latter case. We also observe that a simple uniform price may not be enough to induce the optimal behavior of customers in an observable case.

After obtaining the decentralized and optimal arrival rates under each scenario, we study how the server can increase social welfare by controlling the arrival rate and activation value N. For example, in the call blending system, the system designer may need to decide on the optimal trigger point for an agent to switch from making outbound call to handling inbound calls. If both variables can be costlessly controlled then it is obvious that in the resulting first-best solution $N = 1$ and nothing can be gained by increasing N. However, given that customers are strategic and the arrival rates are decentralized, it is of greater interest to compute the second best optimal threshold value N which maximizes social welfare. Analytically, we will show that with no information and no set-up and shut-down costs, the maximization of social welfare in the system requires a paradoxical operating policy for the server: When the utilization is light, the server should be always available

to shorten customers' waiting time; when the utilization is large, N should be large. We find that it is optimal to set the server as always available in the full information case.

The paper is organized as follows: Section 2 provides the literature review. Section 3 introduces the model and assumptions. Sections 4-5 study the equilibrium and optimal arrival rates and optimal N with no information and full information, respectively. Section 6 provides the concluding remarks. The supplement contains proofs and other technical material, and is available in the e-companion.

2. Related Literature

Study on customers' decentralized behavior and socially optimal control of arrivals was pioneered by Naor (1969) with a single-server system with an *observable* queue, i.e., upon arrival, a customer is informed about the queue length before a decision is made to join. Edelson and Hildebrand (1975) considered the *unobservable* case. There is more related work in the survey book by Hassin and Haviv (2003).

Most of the literature on vacation queues does not allow for the strategic behavior of customers. We know of two exceptions. Burnetas and Economou (2007) assumed $N = 1$ and an exponential setup time when the server starts a new busy period. They considered the strategic behavior of customers under different levels of information which may include the queue length and/or the state of the server (during setup or busy). In particular, if only the queue length is known and the setup time is considerably long, the FTC behavior of customers is observed. In this paper we assume that customers are aware of the service policy, in particular the threshold N, and react to it in a strategic way. We also consider a social optimization problem. Economou and Kanta (2008) considered strategic customer behavior in an observable queue with server vacations due to breakdowns. In this model, the length of the vacation is independent of the queue size.

This paper is also related to some recent work on the decentralized equilibrium in service systems where both negative and positive externalities exist. Veeraraghavan and Debo (2009) considered situations where queue length indicates not only congestion but also service quality. They showed

that when the service rates and unknown service values are negatively correlated, customers prefer to join longer queues. When the service rates are positively correlated with unknown service values, customers might join shorter queues. Johari and Kumar (2008) considered a type of network service such as an on-line gaming system, where users form a club. In such a system, both negative and positive externalities exist. They characterize the size of the club for self-interested users to form autonomously and they showed that the decentralized size is always smaller than the one chosen by the service manager.

3. Formulation and Preliminaries

We assume that potential customers arrive in accordance to a Poisson process with rate Λ. There is a single server, and the service times are independent and exponentially distributed with mean μ^{-1}. The server uses an N-policy, that is, it shuts down when the system becomes empty of customers and resumes service after N arrivals. Unless otherwise stated, assume that $N > 1$, otherwise, it becomes a regular queue.

Suppose that the utility of a customer consists of a *reward* for receiving service minus a *waiting cost*. This waiting cost is linear and depends on the customer-specific parameter and the *waiting time*. The waiting time means the total sojourn time in the system. We also consider an additive social utility composed of the sum of individual utilities of all served customers. A solution that maximizes the social utility is *socially optimal*, or simply *optimal*. Specifically, define

- W = expected waiting time in system.
- W_i= conditional expected waiting time given information state i.
- θ = customer-type delay-sensitivity parameter, indicating his cost per time unit spent in the system.
- R = reward to the customer for receiving service, $R \geq 0$.
- U = the expected utility *for a customer who joins the system*: $U = R - \theta W$ in the unobservable queue; $U = R - \theta W_i$ for a customer who joins at state i in the observable case.
- $\nu = \frac{R\mu}{\theta}$ the upper bound on the number of service epochs that a customer is ready to wait.
- $\rho = \frac{\Lambda}{\mu}$ the system utilization factor when every potential customer joins.

4. Unobservable vacation queues with N-policy

Consider symmetric equilibrium strategies. For the pure strategies, either all of the customers join the queue or all balk. With a mixed strategy, an arriving customer joins with a certain probability, α, and the *effective arrival rate*, or *joining rate*, is $\lambda = \Lambda\alpha$.

Obviously, "all balk" is always an equilibrium strategy; if all others choose balk, the server will never return from vacation and the best choice for the customer is to balk too. (Here, we use the assumption that $N > 1$.) Therefore, $\lambda = 0$ is always an equilibrium arrival rate. Below, we restrict our attention to positive equilibrium arrival rates.

4.1. Equilibrium

In a single server vacation queue, the stationary waiting time can be decomposed into a sum of two independent random variables. One is the corresponding waiting time in a regular queue without vacation and the other is additional delay due to vacation (see, e.g., Doshi 1986). For an $M/M/1$ queue with an N-policy and arrival rate λ ($\lambda < \mu$), we can express the waiting time explicitly (see Yadin and Naor, 1963) as:

$$W(\lambda) = \frac{1}{\mu - \lambda} + \frac{N-1}{2\lambda}, \tag{1}$$

where $\frac{1}{\mu-\lambda}$ is the expected waiting time in a standard $M/M/1$ queue and $\frac{N-1}{2\lambda}$ is the expected extra waiting time for the server to begin to work. This can be interpreted as follows. The former term means that an increasing arrival rate increases the average waiting time such as that in a regular queue $M/M/1$ (negative externality). The second term means that an increase in the arrival rate decreases the idle time for the server and reduces the average waiting time (positive externality).

The function $W(\lambda)$ is strictly convex in λ, with a minimum value of:

$$W(\tilde{\lambda}) = \frac{1}{\mu}\left(1 + \sqrt{\frac{N-1}{2}}\right)^2 \quad \text{at} \quad \tilde{\lambda} = \frac{\mu\sqrt{\frac{N-1}{2}}}{1 + \sqrt{\frac{N-1}{2}}}. \tag{2}$$

Using (1), Λ is an equilibrium arrival rate if $\Lambda < \mu$ and

$$R \geq \theta\left(\frac{1}{\mu - \Lambda} + \frac{N-1}{2\Lambda}\right).$$

The above inequality can be equivalently written as

$$\nu \geq \frac{1}{1-\rho} + \frac{N-1}{2\rho}.$$

Otherwise, an equilibrium arrival rate $0 < \lambda < \Lambda$ solves the equation:

$$R = \theta \left(\frac{1}{\mu - \lambda} + \frac{N-1}{2\lambda} \right), \tag{3}$$

or

$$\nu = \frac{1}{1-\rho} + \frac{N-1}{2\rho}.$$

Equation (3) may have no, one or two solutions. In the latter case, the two solutions λ_1 and λ_2 satisfy $0 \leq \lambda_1 \leq \tilde{\lambda} \leq \lambda_2$, where

$$\lambda_1 = \frac{R\mu - \frac{3-N}{2}\theta - \sqrt{R^2\mu^2 + \frac{(3-N)^2}{4}\theta^2 - (N+1)\mu\theta R}}{2R},$$

and

$$\lambda_2 = \frac{R\mu - \frac{3-N}{2}\theta + \sqrt{R^2\mu^2 + \frac{(3-N)^2}{4}\theta^2 - (N+1)\mu\theta R}}{2R}.$$

The feasibility of these solutions to be equilibrium arrival rates depends on their value relative to μ. Using (2), the condition $R > (=, <)\theta W(\tilde{\lambda})$ is equivalent to $\nu > (=, <) \left(1 + \sqrt{\frac{N-1}{2}}\right)^2$. This gives the following characterization of the equilibrium solutions:

Proposition 1.

(a) If $\nu < \left(1 + \sqrt{\frac{N-1}{2}}\right)^2$*, there exists no positive equilibrium arrival rate;*

(b) if $\nu = \left(1 + \sqrt{\frac{N-1}{2}}\right)^2$*, there exist one positive equilibrium arrival rate* $\lambda_e = \tilde{\lambda}$ *iff* $\tilde{\lambda} \leq \Lambda$*;*

(c) if $\nu > \left(1 + \sqrt{\frac{N-1}{2}}\right)^2$*,* λ_1, λ_2 *and* Λ *could all be an equilibrium arrival rate. Specifically, there exist two positive equilibrium arrival rates* $\lambda_e = (\lambda_1, \lambda_2)$ *if* $\lambda_2 \leq \Lambda$*; two positive equilibrium arrival rates* $\lambda_e = (\lambda_1, \Lambda)$ *if* $\lambda_1 < \Lambda < \lambda_2$ *(which reduces to one if* $\lambda_1 = \Lambda$*); and no positive equilibrium arrival rate if* $\Lambda < \lambda_1$*.*

For the two equilibrium arrival rates in Case (c), the solution with λ_2 or Λ are stable; that is, if there is a small perturbation to λ, the system will converge back to λ. The equilibrium with 0

is also stable. However, the equilibrium with λ_1 is unstable. With any small increase of the arrival rate, the expected waiting time decreases and more customers will arrive in the system. This will further increase λ.

We illustrate Case (c) of Proposition 1 in Figure 1. We observe two positive equilibrium solutions for the continuous curve where $\Lambda > \lambda_2$. The equilibrium solutions change when we move the vertical line at Λ to the left. In the rang of $\lambda_1 < \Lambda \leq \lambda_2$, the larger one becomes Λ; when $\Lambda = \lambda_1$, the two positive equilibrium solutions reduce to be only one; when $\Lambda < \lambda_1$, there exist no positive equilibrium solutions. Note that the positive equilibrium on the left is not stable.

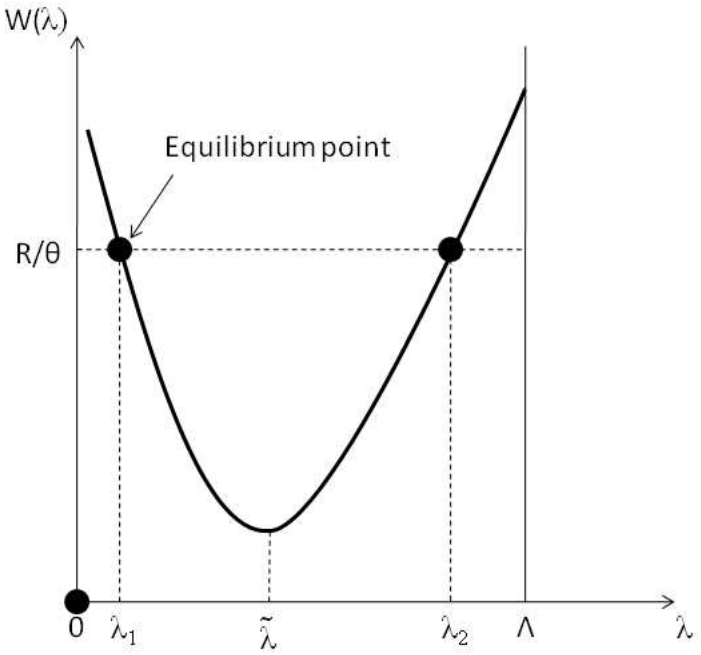

Figure 1 Equilibrium arrival rates in the unobservable case

4.2. Optimal arrival rate

The goal of a social planner is to maximize overall social welfare; that is, the sum of customer utility and the payoff of the server. Here, the control variables are the arrival rate and N.

We first assume a fixed N and consider that the decision maker will set an arrival rate λ so that social welfare, $SW(\lambda)$, is maximized, where:

$$SW(\lambda) = \lambda \left[R - \theta \left(\frac{1}{\mu - \lambda} + \frac{N-1}{2\lambda} \right) \right]. \tag{4}$$

Note that the price, p, does not appear in the above objective function because it is considered an internal transfer of welfare in the system.

This social welfare function is concave and the optimality condition of (4) is:

$$SW'(\lambda) = R - \frac{\mu\theta}{(\mu-\lambda)^2} = 0.$$

Solving this equation yields the unique optimal solution:

$$\bar{\lambda} = \mu - \sqrt{\mu\theta/R}. \tag{5}$$

It is interesting that $\bar{\lambda}$ does not depend on N. Note that $W(\lambda)$ in (1) is the sum of two terms $\frac{1}{\mu-\lambda}$ and $\frac{N-1}{2\lambda}$ where the second term measures the positive externality of increasing λ. In the social objective function, the positive externality part becomes a constant $\frac{N-1}{2}$, and does not affect the decision on λ.

We now give bounds on $\bar{\lambda}$.

PROPOSITION 2. *If* $R > \theta W(\tilde{\lambda})$ *then* $\tilde{\lambda} < \bar{\lambda} < \lambda_2$.

To understand the bounds on $\bar{\lambda}$ intuitively, note that $\lambda < \tilde{\lambda}$ is not optimal since $\lambda = \tilde{\lambda}$ has a greater arrival rate with a smaller expected waiting time for each customer. $\lambda > \lambda_2$ is also not optimal since the individual utilities are negative in this range.

Social welfare at $\bar{\lambda}$ is:

$$\begin{aligned} SW(\bar{\lambda}) &= \mu R - \sqrt{\mu\theta R} + \theta \left(1 - \sqrt{\frac{\mu R}{\theta}} \right) - \theta \frac{N-1}{2} \\ &= \theta \left(\frac{\mu R}{\theta} - 2\sqrt{\frac{\mu R}{\theta}} - \frac{N-3}{2} \right). \end{aligned} \tag{6}$$

We see that $SW(\bar{\lambda})$ is monotone decreasing in N. Therefore, even though the optimal solution $\bar{\lambda}$ does not depend on N, the welfare does.

The optimal arrival rate for the system can be expressed as a function of $\bar{\lambda}$, which depends on the different conditions in the parameters.

PROPOSITION 3.

(a) If $\nu < \left(1+\sqrt{\frac{N-1}{2}}\right)^2$, *there exists a unique optimal arrival rate* $\lambda^* = 0$;

(b) if $\nu = \left(1+\sqrt{\frac{N-1}{2}}\right)^2$, *there exist two optimal arrival rates* $\lambda^* = \{0, \tilde{\lambda}\}$;

(c) if $\nu > \left(1+\sqrt{\frac{N-1}{2}}\right)^2$, *there exists a unique optimal arrival rate* $\lambda^* = \min\{\bar{\lambda}, \Lambda\}$.

In Cases (a) and (b) of Proposition 3, the decentralized equilibrium is optimal. Since $\lambda_1 < \bar{\lambda} < \lambda_2$, the optimal arrival rate in Case (c) may be smaller or greater than the equilibrium rate. Therefore, it is unclear whether the social planner wants a tax to discourage arrivals or a subsidy to encourage arrivals. However, if only a stable and positive equilibrium arrival rate is of interest, a toll is needed as in the $M/M/1$ model.

Also, since $\bar{\lambda} < \lambda_2$, no customers obtain negative utility in a socially optimal solution. This is easy to understand. Since customers are identical, if one customer obtains a negative utility, everyone gets a negative utility. This will obviously not happen in social optimization.

4.3. Optimal N

Often, control over arrival rates may not be feasible. A more convenient and implicit way to regulate arrivals is through the use of an activation value N to maximize social welfare. We now study this problem.

Unlike the price variable which does not appear in the social welfare function, as it is just an internal transfer of welfare in the system, N shall be included as it is associated with the operation costs of the server. A larger N decreases the operation cost of the server but also increases customer waiting time and turns away some customers. Therefore, an optimal decision on N needs to consider the trade-off between saving on operation cost and losing customers.

Of course, the operation costs associated with N can be complex and may include a set-up and shut-down cost. To simplify our analysis, we assume zero set-up and shut-down costs, and only

consider a busy-period cost, c_b. Thus, the average operating cost per unit time, $C(N)$, can be expressed as:

$$C(N) = c_b P_{\text{busy}},$$

where P_{busy} is the busy probability for the server.[1]

In previous sections, we obtained the expression for customer welfare given a certain N, $SW(N)$. The social welfare function here can be expressed as the difference between $SW(N)$ and $C(N)$; that is, $SW(N) - C(N)$.

Proposition 1 has three cases. In the first case, $\lambda_e = 0$. The second case assumes that ν and N exactly satisfy a specific relation, which is unlikely to hold with general data. In the third case, $\nu > \left(1 + \sqrt{\frac{N-1}{2}}\right)^2$ or equivalently, $N \le \bar{N}$, where:

$$\bar{N} = \left\lfloor 2\left(\sqrt{\nu} - 1\right)^2 \right\rfloor + 1.$$

In this case, there are three equilibrium solutions. Among them, 0 is obviously not interesting, and λ_1 is unstable. A small perturbation which is caused, for example, by a short-term promotion, will attract more arrivals which will reduce waiting time even further, and subsequently, induce even more arrivals. This continues until a stable equilibrium $\min(\lambda_2, \Lambda)$ is reached. Therefore, we assume the third equilibrium when we compute an optimal N. The optimization problem becomes:

$$\max_N \lambda_e \left[R - \theta \left(\frac{1}{\mu - \lambda_e} + \frac{N-1}{2\lambda_e} \right) \right] - C(N), \tag{7}$$

where $\lambda_e = \min(\lambda_2, \Lambda)$ and

$$\lambda_2 = \frac{R\mu - \frac{3-N}{2}\theta + \sqrt{R^2\mu^2 + \frac{(3-N)^2}{4}\theta^2 - (N+1)\mu\theta R}}{2R}.$$

The busy probability for the $M/M/1$ with an N-policy and arrival rate λ_e is:

$$P_{\text{busy}} = \frac{\lambda_e}{\mu},$$

[1] Suppose that there is also a positive idle-period cost c_i. Then $C(N) = c_b P_{\text{busy}} + c_i(1 - P_{\text{busy}}) = (c_b - c_i)P_{\text{busy}} + c_i$. The only term that affects the decision on N is the difference $(c_b - c_i)$. Therefore, assuming $c_i = 0$ is without loss of generality.

which is independent of N (N just changes the length of the busy and idle cycles, but not their ratio). Thus, the cost $C(N)$ can be expressed as:

$$C(N) = c_b \frac{\lambda_e}{\mu}. \tag{8}$$

We then have the following analytical result.

PROPOSITION 4. *If* $\rho \le 1 - \frac{1}{\sqrt{\nu}}$, *then* $N^* = 1$; *if* $\rho \ge 1 - \frac{1}{\nu}$, *then* $N^* = \bar{N}$.

The first part of Proposition 4 means that when the arrival rate is small enough and all of the customers join with any $N \le \bar{N}$, the optimal decision is to set the server as always available. The second part means that when the system workload is heavy such that there exist balking customers even if the server is set as always available, the optimal decision is to set N as large as possible while sustaining the equilibrium. This is sort of paradoxical and reflects the inefficiency of regulating arrivals with N. When the utilization is heavy and customer waiting time is long, we want to decrease the waiting time by using a small N, but in doing so, more customers will come and everybody still obtains 0 utility. For instance, in the shuttle bus example, N corresponds to the capacity of the bus. When the traffic flow is light, the bus company will use small-sized buses; otherwise, it will use large-sized buses. One can expect a similar two-operations regime when customers are heterogeneous on delay sensitivity. On the one hand, when the utilization is so light that even the most impatient customer joins with a large N, N can be further reduced to shorten customers' waiting time while keeping the operation cost unchanged; on the other hand, if the utilization is very heavy such that there exists balking customers for high type of impatient customers when $N = 1$, N can be increased to a certain level so that that impatient customers all balk while the expected waiting time is unchanged. In doing so, customers' utilities are unchanged for each type but the operation cost is reduced.

We are still unclear about the optimal N in the range $\left[1 - \frac{1}{\sqrt{\nu}}, 1 - \frac{1}{\nu}\right]$. Therefore, we conducted a numerical study and observed that there exists a threshold for ρ in this range, such that below this point, $N^* = 1$ and above it, $N^* = \bar{N}$.

Proposition (4) relies on the assumption of zero set-up and shut-down cost for the server and the fact that N does not affect the busy probability in an $M/M/1$ queue with an N-policy. If such set-up and shut-down costs exist, then $C(N)$ is a function of the regeneration cycle length, which is increasing in N. In that case, it is likely that $N^* > 1$.

5. Observable vacation queues with an N-policy

In this section we assume that customers have information on the queue length when they make their decision to join or balk. We assume that *a customer who is indifferent between joining and balking joins*. We use superscripts $-$ and $+$ to distinguish between states with idle and busy server, respectively, when both states are possible. Thus, the set of the states is

$$\left\{0, 1^-, \ldots, (N-1)^-, 1^+, 2^+, \ldots, (N-1)^+, N, N+1, \ldots\right\},$$

where m^- means that the system occupancy is m and the server is idle, and m^+ means that the system occupancy is m and the server is busy.

In general, the strategy to never join is always an equilibrium when $N > 1$, since if this policy is adopted by others, the expected wait for a customer who joins at state 0 is infinite, and thus balking at this state is the best response. In this case, the server is never active.[2] We concentrate on the existence of other equilibrium strategies in which the server is busy for at least a positive fraction of the time. We refer to such as a solution with an *active server*.

A *threshold strategy* with a threshold n is a strategy where customers join if and only if they find at most $n-1$ customers in the system upon arrival. Thus the maximum number of customers in the system at any time is n. We will see that indeed, the equilibrium solutions are threshold strategies. However, the optimal strategy may have a more general structure, although it also involves a threshold.

[2] The conditions for equilibrium are actually milder; for example, it suffices that customers balk at state 1^- to ensure that balking at 0 is an optimal response. We avoid a thorough analysis of this subject and in particular, a description of all equilibria and subgame perfect solutions. The reader is referred to §1.5 in Hassin and Haviv (2003) for an example of such analysis.

5.1. Equilibrium

The next sentence is not exact. In a regular queue we require naturally that $R > \theta/\mu$ and this is sufficient for the server to be busy a positive fraction of the time. Without this condition nobody joins. So in the vacation queue we also have a condition but it is stronger as explained below

In a regular queue, an incoming customer always joins an idle queue, as the waiting time is zero. However, this need not hold in the observable vacation queue. For the server to be active, the set of states in which customers join must include all the states where the server is idle. This requires that the expected utility in each of these states is nonnegative, given that all of the customers join in these states. This condition depends on the relationship between Λ and μ. When $\Lambda > \mu$, the longest expected waiting time for a customer who arrives when the server is idle, is when there are $N-1$ customers waiting in front of him/her; hence, the waiting time is $\frac{N}{\mu}$. When $\Lambda < \mu$, the longest waiting happens when a customer joins an empty system; the waiting time is $\frac{N-1}{\Lambda} + \frac{1}{\mu}$. Customers join an idle queue if their reward is at least as great as the worst waiting cost. We summarize the sufficient conditions for a server to be active in the following proposition.

PROPOSITION 5. *There exists an equilibrium solution with an active server if and only if either (i) $\rho \geq 1$, and $\nu \geq N$, or (ii) $\rho \leq 1$ and $\nu \geq \frac{N-1}{\rho} + 1$.*

We now assume that the conditions for an active server as given in Proposition 5 are satisfied. Observe that an arrival at state m^+ is associated with a lower expected waiting time than arriving at m^-. It follows that in equilibrium with an active server, all join when the number of customers observed upon arrival is at most $N-1$. The equilibrium strategy is therefore characterized by a threshold value $n_e \geq N$. A customer who observes state m^+ is motivated to join if $R \geq \theta \frac{m+1}{\mu}$, or $m \leq \frac{R\mu}{\theta} - 1$. Therefore, $n_e = \left\lfloor \frac{R\mu}{\theta} \right\rfloor$. The conditions of Proposition 5 indeed guarantee that this value is at least N. We summarize the equilibrium threshold solution in the following proposition.

PROPOSITION 6. *Suppose that the conditions of Proposition 5 are satisfied. Then the unique equilibrium threshold is $n_e = \lfloor \nu \rfloor \geq N$.*

5.2. Optimal arrival rates

We now show that, unlike in the unobservable case, an optimal strategy may induce customers to join even in some states where their individual expected utility is negative. The reason is that when a customer joins, s/he may reduce the expected waiting time of previous arrivals by shortening the time until the server turns on. This can even bring positive externality to future customers as the joining behavior prevents the server from idling. In particular, it may so happen that all arrivals join when the server is idle, whereas balking is more easily tolerated when the server is busy. An extreme case is when Λ and N are both large. When the server is idle, all should join, but when it is busy, balking should start at low state values due to the high expected arrival rate.

By the assumption, a customer joins if his/her utility is 0. Since it is assumed that the customers are identical, the state-dependent arrival rates are either 0 or Λ. In the idle states, obviously, it is never optimal to set any arrival rate to be 0; otherwise, the system will stay idle forever. Therefore, the choice for the social planner is left to decide which states will have Λ when the server is busy. Consider two adjacent states n^+ and $(n+1)^+$. Obviously, when $n \geq N$, it will never be possible that the arrival rate on n^+ is 0 while the one on $(n+1)^+$ is Λ as the state $(n+1)^+$ can not be reached. However, when $n < N$, state $(n+1)^+$ is reachable from state $(n+2)^+$ even though the arrival rate at n^+ is 0. But the case with the arrival rate on n^+ to be 0 and on $(n+1)^+$ to be Λ is not optimal, as we can simply consider a similar system with the arrival rate being Λ on both n^+ and $(n+1)^+$. The latter system can generate higher average utility for customers, as those who join at n^+ obtain the minimal expected waiting time, and also it keeps the system away from idleness. Similarly, it is better to have the arrival rate of Λ on $(n-1)^+$ too and so on. This is an intuitive argument. One can also follow the approach in Stidham and Weber (1989) by directly obtaining the optimality equations on arrival rates through uniformization technique and then show that the optimal arrival rates are non-increasing in the state variable. Hence, the second-best control over arrival rates requires customers to join when the queue is below a threshold.

Denote the optimal threshold by n^*. If social welfare is positive, then there are two cases:1. $n^* \geq N$; 2. $n^* \in \{1, \ldots, N-1\}$.

Case 2 means that if the server is idle, customers always join, and when the server is busy, customers are not allowed to join if the system occupancy is n^* or larger. To find the threshold, we first solve the stable distribution for the system states and then obtain the expression for the social welfare. Then, we can solve the optimization problem. The maximum social welfare obtained in Cases 1 and 2 is denote by SW_1 and SW_2, respectively. The optimal solution value is then $SW = \max\{SW_1, SW_2\}$. In the following analysis, we present the results with $\rho \neq 1$. One can easily derive the results for the degenerate case with $\rho = 1$.

Case 1, $n^* \geq N$

The state transition diagram is shown in Figure 2. The corresponding balance equations and the derivation of the expected number of customers are detailed in the Appendix.

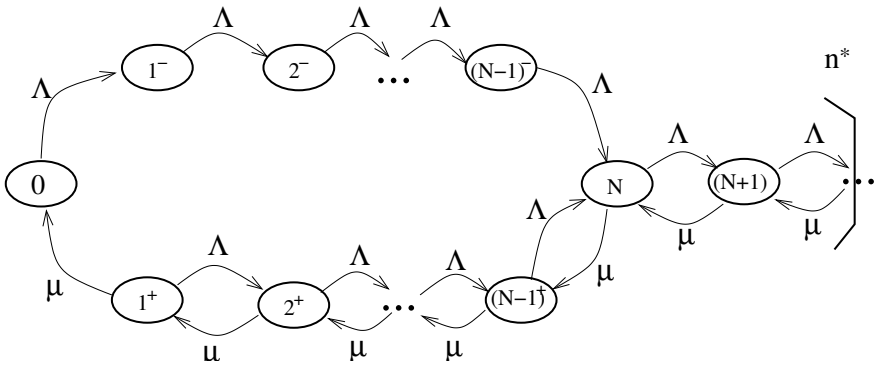

Figure 2 State Transition Diagram of Case 1

The social welfare problem is to compute a threshold n which maximizes

$$SW_1(n) = (R - \theta W)\Lambda(1 - p_n) = R\Lambda(1 - p_n) - \theta L,$$

where[3]

$$L = \frac{p_0}{1-\rho} \cdot \frac{(N-1)N}{2} + p_0 \rho \frac{(1-\rho)N + \rho^{n+1}(1-\rho^{-N})(n(1-\rho)+1)}{(1-\rho)^3}$$

[3] For the computational study, we use parameters ρ, v, N, which have $SW_1 = \theta[\rho v(1-p_n) - L]$, and normalize $\theta = 1$.

is the expected number of customers in the system,

$$p_n = \frac{\rho^{n-N+1}(1-\rho^N)}{1-\rho} p_0$$

is the steady state probability that the queue is at its maximum size, and

$$p_0 = \frac{(1-\rho)^2}{N - N\rho - \rho^{n-N+2} + \rho^{n+2}}$$

is the steady state probability that the system is empty.

Case 2, $n^* \in \{1, \ldots, N-1\}$

Note that this case is possible when ρ is very large. In fact when ρ and N are large, we might have $n^* = 2$ to keep the server active. This is in contrast to Naor's $M/M/1$ analysis (Naor, 1969) where $n^* = 1$ when ρ is large.

The state transition diagram is shown in Figure 3.

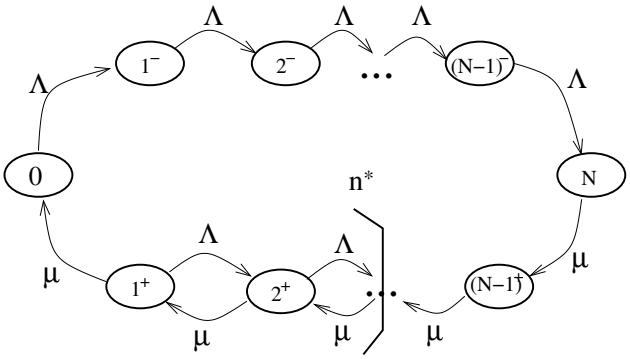

Figure 3 State Transition Diagram of Case 2

The social welfare problem (see supplement for details) is to compute a threshold n which maximizes[4]

$$\begin{aligned} SW_2(n) &= (R - \theta W)\Lambda(1 - p_{(n)^+} - p_{(n+1)^+} - \cdots - p_N) \\ &= R\Lambda\left(1 - \frac{\rho(1-\rho^n)}{1-\rho} p_0 - \rho(N-n)p_0\right) - \theta L, \end{aligned}$$

[4] For the computational study we use $SW_2 = \theta\left\{\rho\nu\left[1 - \frac{\rho(1-\rho^n)}{1-\rho} p_0 - \rho(N-n)p_0\right] - L\right\}$ and normalize $\theta = 1$.

where

$$L = p_0 \cdot \frac{(N-1)N}{2} + \frac{p_0\rho}{1-\rho}\left(\frac{n(n+1)}{2} + \frac{-\rho+(n+1)\rho^{n+1}-n\rho^{n+2}}{(1-\rho)^2}\right) + \rho p_0 \frac{(n+N+1)(N-n)}{2}.$$

is the expected number of customers in the system, and

$$p_0 = \frac{(1-\rho)^2}{N - N\rho - (N-n+1)\rho^2 + (N-n)\rho^3 + \rho^{n+2}}$$

is the steady state probability that the system is empty.

Define $SW(n)$ as $SW_1(n)$ if $n \geq N$ and $SW_2(N)$ otherwise. The following proposition shows that $SW(n)$ is unimodal and therefore, n^* is unique.

PROPOSITION 7. *The social welfare function $SW(n)$ is unimodal.*

It is interesting to observe that a simple tax (positive or negative, uniform over all customers and states) may not be sufficient to induce customers to behave in accordance to the optimal strategy. In particular, suppose that $n^* \in \{1, \ldots, N-1\}$. This means that we want customers to join at state $(N-2)^-$ and balk at $(N-2)^+$. However, with any tax that is uniformly imposed on all joining customers, the expected utility of a customer who joins at state $(N-2)^-$ is strictly lower than his or her utility if s/he joins at state $(N-2)^+$. Hence, in contrast with Naor's findings for the corresponding $M/M/1$ model, a simple tax cannot induce customers to behave optimally. The tax should discriminate, for example, customers who join the system while the server is busy from those that join in idle states. **PG: The next several sentences are required to be expanded to include more details. Please see the following paragraph on it.**

It seems that there is some repetition in the beginning

Even when $n^* > N$, a uniform tax may not be sufficient. The tax needs to deter arrivals in state n^* while allowing arrivals at state $n^* - 1$ and at any state when the server is idle. If $\Lambda > \mu$, the longest waiting when the server is idle occurs at state $(N-1)^-$ and the expected waiting time is N/μ, which is not greater than the expected waiting time at state $n^* - 1$. In this case, a uniform

tax p such that $R-\frac{\theta n^*}{\mu}<p\le R-\frac{\theta(n^*-1)/}{\mu}$ achieves socially optimal behavior. If $\Lambda<\mu$, the longest waiting when the server is idle is at state 0 and the expected waiting time is $\frac{N-1}{\Lambda}+\frac{1}{\mu}$. In this case, only when $n^*\ge\mu\frac{N-1}{\Lambda}+2$, $\frac{N-1}{\Lambda}+\frac{1}{\mu}\le\frac{n^*-1}{\mu}$ and the uniform tax will be adequate.

Even when $n^*>N$, a uniform tax may not be enough. To be sufficient, a uniform tax needs to satisfy three conditions: 1. deterring arrivals at state n^*, that is, $p>R-\frac{\theta n^*}{\mu}$; 2. allowing arrivals at state n^*-1, that is, $p\le R-\frac{\theta(n^*-1)}{\mu}$; 3. allowing arrivals at any state when the server is idle. Since the longest waiting when the server is idle occurs at either state $(N-1)^-$ or state 0, the third condition can be expressed as $p\le R-\theta\max\{\frac{N-1}{\Lambda}+\frac{1}{\mu},\frac{N}{\mu}\}$. If $\Lambda\ge\mu$, $\frac{n^*-1}{\mu}\ge\frac{N}{\mu}\ge\frac{N-1}{\Lambda}+\frac{1}{\mu}$. In this case, condition 2 implies condition 3. In this case, a uniform tax p such that $R-\frac{\theta n^*}{\mu}<p\le R-\frac{\theta(n^*-1)}{\mu}$ achieves socially optimal behavior. If $\Lambda<\mu$, the longest waiting when the server is idle is at state 0 and the expected waiting time is $\frac{N-1}{\Lambda}+\frac{1}{\mu}$. In this case, when $n^*\le 1+\frac{(N-1)\mu}{\Lambda}$, $R-\theta\left(\frac{N-1}{\Lambda}+\frac{1}{\mu}\right)\le R-\frac{\theta n^*}{\mu}$, and conditions 1 and 3 can never be satisfied simultaneously with a uniform tax; when $1+\frac{(N-1)\mu}{\Lambda}<n^*\le 2+\frac{(N-1)\mu}{\Lambda}$,

$$\begin{aligned}R-\frac{\theta n^*}{\mu}&<R-\theta\left(\frac{N-1}{\Lambda}+\frac{1}{\mu}\right)\\&\le R-\theta\frac{n^*-1}{\mu},\end{aligned}$$

the uniform tax such that $R-\frac{\theta n^*}{\mu}<p\le R-\theta\left(\frac{N-1}{\Lambda}+\frac{1}{\mu}\right)$ is adequate; when $n^*>2+\frac{(N-1)\mu}{\Lambda}$, $R-\theta\left(\frac{N-1}{\Lambda}+\frac{1}{\mu}\right)>R-\theta\frac{n^*-1}{\mu}$ and the uniform tax such that $R-\frac{\theta n^*}{\mu}<p\le R-\theta\frac{n^*-1}{\mu}$ is adequate.

Figures 4- 6 present some numerical results (in Figure 6, the equilibrium threshold is represented by the the integer part of the 45-degree line). We see that for larger values of N, Case 2 with $n^*<N$ is obtained. As expected, social welfare under n^* decreases with N and increases with v and ρ. As in queues without vacations ($N=1$), $n^*\le n_e$ in spite of the positive externalities. The difference increases with v and decreases with N. In Figure 5, the server is busy in equilibrium for points (ρ,Λ) above the upper curve. The server is busy in the optimal solution for points above the lower curve. We see that the solution with a busy server is obtained less easily in equilibrium, and larger values of Λ or ρ are required for this to happen in contrast to those under an optimal solution.

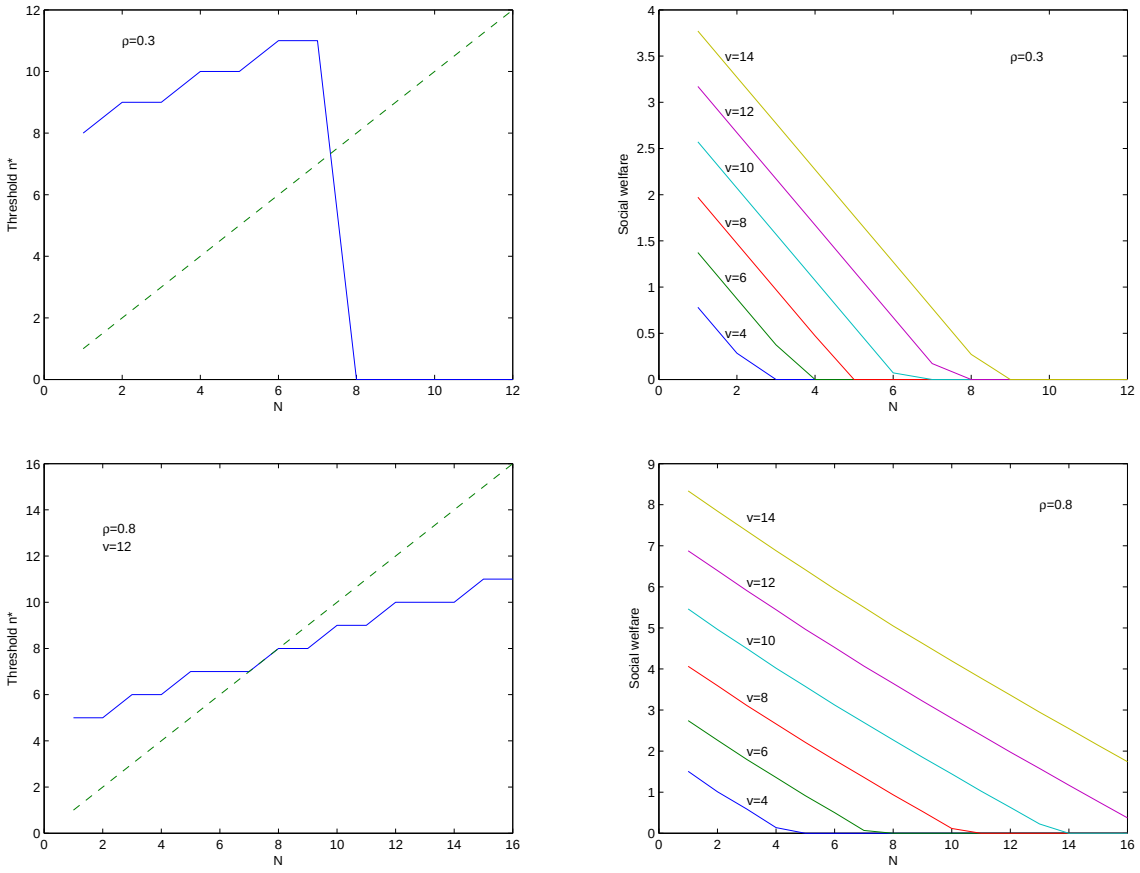

Figure 4 Social welfare and optimal thresholds: optimal thresholds for $\nu = 12$ and $\rho = 0.3$ (upper-left), social welfare for different ν and $\rho = 0.3$ (upper-right), optimal thresholds for $\nu = 12$ and $\rho = 0.8$ (lower-left), social welfare for different ν and $\rho = 0.8$ (lower-right)

5.3. Optimal N

Proposition 5 provides the conditions for equilibrium with an active server. Proposition 6 shows that in this case, customers follow a threshold strategy where they join if and only if the number of customers in the system is at most $n_e - 1$ where $n_e = \lfloor \nu \rfloor$. The social problem is to compute an activation value N^* which maximizes

$$(R - \theta W)\Lambda(1 - p_{n_e}) - C(N),$$

where p_{n_e} is the probability for the state to be n_e.

Similar to the analysis in Case 1 of §4.1.2, we obtain:

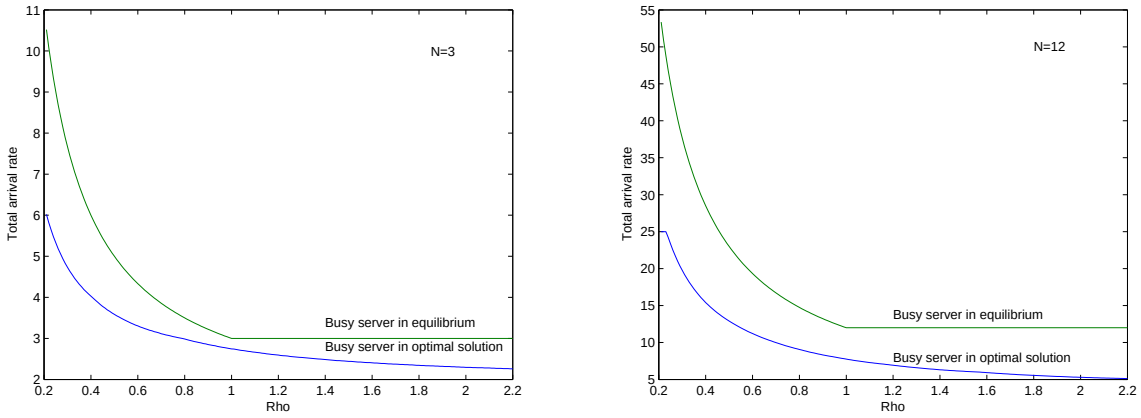

Figure 5 Equilibrium and optimal requirement for busy server: the boundary curve for $N=3$ (left) and for $N=12$ (right)

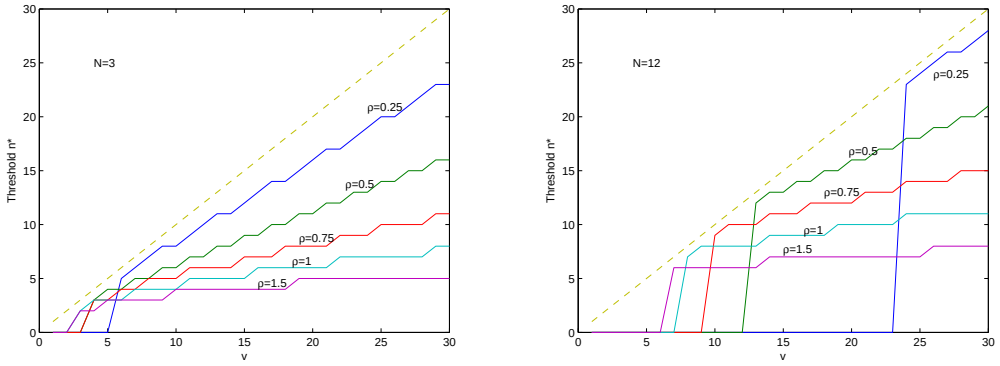

Figure 6 Equilibrium and optimal thresholds: the 45-degree line (dashed) represents the equilibrium threshold and other curves represent the optimal threshold.

$$p_0 = \frac{(1-\rho)^2}{N - N\rho - \rho^{n_e-N+2} + \rho^{n_e+2}}.$$

Thus,

$$P_{\text{idle}} = p_0 + p_{1^-} + \cdots + p_{(N-1)^-} = Np_0 = \frac{N(1-\rho)^2}{N - N\rho - \rho^{n_e-N+2} + \rho^{n_e+2}}.$$

We also obtain

$$p_{n_e} = \frac{\rho^{n_e-N+1}(1-\rho^N)}{1-\rho} p_0 = \frac{(1-\rho)(\rho^{n_e-N+1} - \rho^{n_e+1})}{N(1-\rho) - \rho^{n_e-N+2} + \rho^{n_e+2}}. \tag{9}$$

A comparison of expressions of P_{idle} and p_{n_e} yields the following equation:[5]

$$1 - P_{\text{idle}} = \Lambda \frac{1 - p_{n_e}}{\mu}.$$

[5] This formula can also be directly obtained with a sample path analysis. Denote by N_T the number of customers

Then, the social welfare in a system with N can be expressed as:

$$\Lambda(1-p_{n_e})\left\{R-\theta W-\frac{c_b}{\mu}\right\}. \tag{10}$$

We assume $R \geq \theta W + c_b/\mu$; otherwise, the server will not work at all.

We now provide a lemma on the monotonicity of p_{n_e} and W with respect to N.

LEMMA 1. *In an $M/M/1/n$ queue with N-policy, both p_n and W are increasing in N.*

Lemma 1 implies that the welfare function (10) is decreasing in N. Thus, the decision maker should make N the smallest; that is, $N^* = 1$ as summarized in the following proposition.

PROPOSITION 8. *When the queue is observable, $N^* = 1$.*

One may believe that there exists a trade off between the operation cost and customer welfare associated with the decision of N. Although setting a larger N can turn away some customers, it may save on operation costs for the server due to a lighter workload. However, when the utilization is reduced by a certain percentage and thus the cost is also reduced accordingly, the system has exactly the same percentage of welfare loss due to the balking of customers who would join with a smaller N. The amount of cost saving cannot cover the loss of welfare as long as it is optimal for the system to be active. Moreover, a larger N also reduces the welfare for customers who join the system, due to greater expected waiting time.

A comparison of Proposition 4 and Proposition 8 allows us to see that the decision of the server is different in the two cases when the traffic is heavy. In the unobservable case, customers always obtain utility 0 when the traffic is heavy enough. In that case, driving away some customers results in zero loss of customer welfare, but a decrease of the operation cost.

$N^* = 1$ means that the server starts service as long as there is one customer. This is similar to the base-stock policy in inventory management. However, if set-up and shut-down costs exist, one should expect a larger N^* as that will reduce the frequency of setting up and shutting down.

who joined the system in a time period T. Then,

$$1 - P_{\text{idle}} = \lim_{T\to\infty} \frac{N_T/\mu}{T} = \Lambda \frac{1-p_{n_e}}{\mu.}$$

6. Conclusions

In this paper, we studied the decentralized behavior of customers and social optimization in a queue with a threshold policy called the $N-$policy. We considered two scenarios based on the availability of information on delay; no information and full information.

With no information, the game among customers is a supermodular game when the effective arrival rate is below a certain level; that is, the joining behavior of a customer is encouraged by the joining decisions of others. We have shown that multiple equilibria exist, and among them, some are stable while others are not. We have demonstrated that a unique optimal arrival rate exists for the social optimization problem. In contrast with a regular queue where the optimal arrival rate is smaller than the equilibrium one, we show that the optimal arrival rate could be larger than some of the equilibrium arrival rates.

With full information, we have shown that customers use a threshold strategy for their joining decisions. We derived the conditions for the system to be active, that is, the effective arrival rate is positive when the server is on vacation. Closed-form expressions for the individually optimal thresholds are derived and the socially optimal thresholds are calculated.

Finally, we have studied the effect of controlling threshold N with the assumption that the customers are strategic. In assuming that start-up and shut-down costs are costless, we show that with no information, the optimal decision on N is either 1 or a large number, which depends on the utilization of the system. With information, the optimal N is always 1.

To transform our model into an amenable one, we make several simplified assumptions that need not be fully actualized in real situations. For example, we assume that the system's parameters, such as the arrival and service rates, are common knowledge, and that the service is memoryless so that the decision of the customer to join is not adaptive, but rather, irrevocable. A relaxation of these assumptions would lead to interesting developments, but the price would be a significant complication of the analysis. As is common in the scientific literature, we believe that our qualitative conclusions are of value under more general settings, but of course, one should take into consideration that they are obtained under simplified assumptions.

Acknowledgments

The authors thank Sunil Kumar and the anonymous referees for their helpful comments and constructive suggestions. Guo's research was supported by HK PolyU research foundation grant no. PB0W. Hassin's research was supported by the Israel Science Foundation grant no. 526/08.

References

Bertsekas, D. 1976. *Dynamic Programming and Stochastic Control.* Academic Press, New York.

Burnetas, A. and A. Economou. 2007. Equilibrium customer strategies in a single server Markovian queue with setup times. *Queueing Systems* **56** 213–228.

Doshi, B. 1986. Queueing system with vacations—a survey. Queueing Systems **1** 29-66.

Edelson, N. and K. Hildebrand. 1975. Congestion tolls for Poisson queueing processes. Econometrica **43** 81-92.

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

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

Hassin, R. and M. Haviv. 2003. *To Queue Or Not To Queue: Equilibrium Behavior in Queueing Systems*. Kluwer.

Johari, R. and S. Kumar. 2008. Externalities in services. Working paper, Graduate School of Business, Stanford University, Stanford, CA.

Knudsen, N. 1978. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica*. **40** 515-528.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37** 15–24.

Stidham, S. and R. Weber. 1989. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Oper. Res.* **87** 611-625.

Takagi, H. 2000 Analysis and application of polling models. In G. Haring, C. Lindemann, M. Reiser (Eds.), *Performance Evaluation: Origins and Directions*, Lecture Notes in Computer Science 1769, Springer, Berlin, 423-442.

Tian, N. and G. Zhang. 2006. *Vacation Queueing Models: Theory and Applications*. Springer.

Veeraraghavan, S. and L. Debo. 2009. *Joining longer queues: Information externalities in queue choice*. *Manufacturing & Service Operations Management* **11** 543–562.

Yadin, M. and P. Naor. 1963. Queueing systems with a removable service station. *Operations Research* **14** 393–405.

Yechiali, U. 1971. On optimal balking rules and tolls charges in the $GI/M/1$ queue, *Operations Research* **19** 349-370.

Zhang, G. 2009. Performance analysis of a queue with congestion-based staffing policy. *Management Science* **55** 240-251.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

EC.1. Social optimization with informed identical customers

Case 1: $n^* \geq N$**.** The balance equations are:

$$\begin{aligned}
p_0\Lambda &= p_{1+}\mu; \\
p_{m-} &= p_{(m+1)-}, && 0 \leq m \leq N-2; \\
p_0\Lambda + p_{m+}\Lambda &= p_{(m+1)+}\mu, && 0 < m < N; \quad (p_{N+} \equiv p_N) \\
p_m\Lambda &= p_{m+1}\mu, && N < m < n^*.
\end{aligned}$$

Then,

$$p_{m+} = \frac{\rho - \rho^{m+1}}{1-\rho} p_0, \qquad 0 < m \leq N; \tag{EC.1}$$

$$p_{m-} = p_0, \qquad 0 < m < N; \tag{EC.2}$$

$$p_{N+j} = \rho^j p_N, \qquad 1 \leq j \leq n^* - N. \tag{EC.3}$$

We also have the normalization condition:

$$p_0 + \sum_{m=1}^{N-1} p_{m-} + \sum_{m=1}^{N-1} p_{m+} + \sum_{j=0}^{n^*-N} \rho^j p_N = 1.$$

This can be simplified to be:

$$\frac{N - N\rho - \rho^{n^*-N+2} + \rho^{n^*+2}}{(1-\rho)^2} p_0 = 1. \tag{EC.4}$$

Note that:

$$\sum_{j=1}^{n} j\rho^j = \frac{\rho - (n+1)\rho^{n+1} + n\rho^{n+2}}{(1-\rho)^2}.$$

With that, we can obtain the expected number of customers in the system as:

$$\begin{aligned}
L &= \sum_{j=1}^{N-1} j\left(p_{j-} + p_{j+}\right) + \sum_{j=0}^{n^*-N} (j+N) p_{j+N} \\
&= \sum_{j=1}^{N-1} j\left(p_0 + \frac{\rho - \rho^{j+1}}{1-\rho} p_0\right) + \sum_{j=0}^{n^*-N} (j+N)\rho^j \frac{\rho - \rho^{N+1}}{1-\rho} p_0 \\
&= \sum_{j=1}^{N-1} j\frac{p_0}{1-\rho} - \frac{p_0\rho}{1-\rho}\sum_{j=1}^{N-1} j\rho^j + \frac{p_0(\rho - \rho^{N+1})}{1-\rho}\sum_{j=1}^{n^*-N} j\rho^j + N\frac{p_0(\rho - \rho^{N+1})}{1-\rho}\sum_{j=0}^{n^*-N} \rho^j
\end{aligned}$$

$$
\begin{aligned}
&= \frac{p_0}{1-\rho}\cdot\frac{(N-1)N}{2} - \frac{p_0\rho}{1-\rho}\cdot\frac{\rho - N\rho^N + (N-1)\rho^{N+1}}{(1-\rho)^2} \\
&\quad + \frac{p_0(\rho-\rho^{N+1})}{1-\rho}\cdot\frac{\rho-(n^*-N+1)\rho^{n^*-N+1}+(n^*-N)\rho^{n^*-N+2}}{(1-\rho)^2} \\
&\quad + N\frac{p_0(\rho-\rho^{N+1})}{1-\rho}\cdot\frac{1-\rho^{n^*-N+1}}{1-\rho} \\
&= \frac{p_0}{1-\rho}\cdot\frac{(N-1)N}{2} - \frac{p_0\rho}{1-\rho}\cdot\frac{\rho - N\rho^N + (N-1)\rho^{N+1}}{(1-\rho)^2} \\
&\quad + \frac{p_0(\rho-\rho^{N+1})}{1-\rho}\cdot\frac{N+(1-N)\rho-(n^*+1)\rho^{n^*-N+1}+n^*\rho^{n^*-N+2}}{(1-\rho)^2} \\
&= \frac{p_0}{1-\rho}\cdot\frac{(N-1)N}{2} + p_0\rho\frac{(1-\rho)N+\rho^{n^*+1}(1-\rho^{-N})(n^*(1-\rho)+1)}{(1-\rho)^3}.
\end{aligned}
$$

The expected waiting time is:

$$W = \frac{L}{\Lambda(1-p_{n^*})}. \tag{EC.5}$$

Case 2: $n^* \in \{1,\dots,N-1\}$.

The balance equations are (with $p_{N^+} \equiv p_N$)

$$
\begin{aligned}
p_0\Lambda &= p_{1^+}\mu; \\
p_{m^-} &= p_{(m+1)^-}, \qquad 0 \le m \le N-2; \\
p_0\Lambda + p_{m^+}\Lambda &= p_{(m+1)^+}\mu, \qquad 0 < m < n^*; \\
p_0\Lambda &= p_{m^+}\mu; \qquad n^* < m \le N.
\end{aligned}
$$

This can be simplified to:

$$p_{m^+} = \frac{\rho(1-\rho^m)}{1-\rho}p_0, \qquad 0 < m \le n^*; \tag{EC.6}$$

$$p_{m^-} = p_0, \qquad 0 < m < N; \tag{EC.7}$$

$$p_{m^+} = \rho p_0, \qquad n^* < m \le N. \tag{EC.8}$$

By the normalization condition, we obtain:

$$\frac{N - N\rho - (N-n^*+1)\rho^2 + (N-n^*)\rho^3 + \rho^{n^*+2}}{(1-\rho)^2}p_0 = 1. \tag{EC.9}$$

We can obtain the expected number of customers in the system:

$$L = \sum_{j=1}^{N-1} j\left(p_{j^-} + p_{j^+}\right) + Np_N$$

$$
\begin{aligned}
&= \sum_{j=1}^{N-1} jp_0 + \sum_{j=1}^{n^*} j \cdot \frac{\rho(1-\rho^j)}{1-\rho} p_0 + \sum_{j=n^*+1}^{N} j\rho p_0 \\
&= p_0 \cdot \frac{(N-1)N}{2} + \frac{p_0\rho}{1-\rho}\left(\frac{n^*(n^*+1)}{2} + \frac{-\rho + (n^*+1)\rho^{n^*+1} - n^*\rho^{n^*+2}}{(1-\rho)^2}\right) \\
&\quad + \rho p_0 \frac{(n^*+N+1)(N-n^*)}{2}.
\end{aligned}
$$

The expected waiting time is:

$$
W = \frac{L}{\Lambda(1 - p_{(n^*)+} - p_{(n^*+1)+} - \cdots - p_N)}.
$$

Proofs of Statements

EC.2. Proof of Proposition 2

Proof A direct verification is possible through a comparison of $\bar{\lambda}$ with $\tilde{\lambda}$ and λ_2. Alternatively, one can look at the derivative of SW at the boundary points.

$$
\begin{aligned}
SW'(\tilde{\lambda}) &= R - \theta W(\tilde{\lambda}) - \tilde{\lambda}\theta W'(\tilde{\lambda}) \\
&= R - \theta W(\tilde{\lambda}) > 0,
\end{aligned}
$$

and

$$
\begin{aligned}
SW'(\lambda_2) &= R - \theta W(\lambda_2) - \lambda_2 \theta W'(\lambda_2) \\
&= -\lambda_2 \theta W'(\lambda_2) < 0.
\end{aligned}
$$

Therefore, $\bar{\lambda}$ must be in the interval $(\tilde{\lambda}, \lambda_2)$. □

EC.3. Proof of Proposition 3

Proof Case (a) is easy. If $\frac{R\mu}{\theta} < \left(1 + \sqrt{(N-1)/2}\right)^2$, the utility is always negative for customers.

For Case (b), note that customers cannot obtain positive utility if $\frac{R\mu}{\theta} = \left(1 + \sqrt{(N-1)/2}\right)^2$. Hence, the optimal arrival rate is either 0 or $\tilde{\lambda}$.

The proof for Case (c) is trivial. □

EC.4. Proof of Proposition 4

Proof Note that the condition $\rho < 1 - \frac{1}{\sqrt{\nu}}$ is equivalent to $\Lambda < \tilde{\lambda}(\bar{N})$. Since we assume $N \le \bar{N}$, and since $\lambda_2(N)$ is decreasing with N, this condition implies that $\Lambda < \tilde{\lambda}(\bar{N}) \le \lambda_2(\bar{N}) \le \lambda_2(N)$, and therefore, $\lambda_e(N) = \min\{\lambda_2(N), \Lambda\} = \Lambda$. Thus the busy probability for the server equals ρ, which is independent of N. Therefore, the second term of the objective function (7), $C(N)$, is a constant, and independent of N. Consequently, maximizing the objective function (7) is equivalent to maximizing $SW(N)$, and this results in $N^* = 1$.

Observe that when $N = 1$, $\frac{\lambda_2}{\mu} = 1 - \frac{1}{\nu}$. Therefore, $\rho \ge 1 - \frac{1}{\nu}$ implies that $\lambda_e = \lambda_2 \le \Lambda$. In this case, all customers obtain the same utility as those who balk; namely, 0. Therefore, the first term of the objective function (7) becomes 0. Maximizing the objective function (7) is equivalent to minimizing $C(N)$, giving $N^* = \bar{N}$. $\square$

EC.5. Proof of Proposition 7

Proof We mainly follow the approach in Knudsen (1972). Consider two systems that are denoted as (n) and $(n+1)$ with thresholds n and $n+1$, respectively. To prove that SW is unimodal in n, we only need to demonstrate that $SW(n) - SW(n-1) \le 0$ implies $SW(n+1) - SW(n) \le 0$.

We first show that the following equation holds for all n

$$SW(n+1) = \frac{p_0(n+1)}{p_0(n)} SW(n) + p_n(n+1)\Lambda(R - \theta W_n). \tag{EC.10}$$

We consider three cases. First, consider $n \ge N$. Then $SW(n)$ and $SW(n+1)$ are $SW_1(n)$, and $SW_1(n+1)$, respectively. By the balance equations (EC.1)-(EC.3):

$$\frac{p_i(n+1)}{p_i(n)} = \frac{p_0(n+1)}{p_0(n)},$$

for all $i \in \Omega$ where $\Omega \equiv \{0, 1^+, \ldots, (N-1)^+, 1^-, \ldots, (N-1)^-, N, \ldots, n\}$. The social welfare function can be expressed as

$$SW_1(n+1) = \sum_{i \in \Omega} p_i(n+1)\Lambda(R - \theta W_i).$$

With this expression, we obtain:

$$\begin{aligned} SW_1(n+1) &= \sum_{i\in\Omega} p_i(n+1)\Lambda(R-\theta W_i) \\ &= \frac{p_0(n+1)}{p_0(n)} \sum_{i\in\Omega\setminus\{n\}} p_i(n)\Lambda(R-\theta W_i) + p_n(n+1)\Lambda(R-\theta W_n) \\ &= \frac{p_0(n+1)}{p_0(n)} SW_1(n) + p_n(n+1)\Lambda(R-\theta W_n). \end{aligned}$$

Second, we consider the situation $n < N-1$. $SW(n)$ and $SW(n+1)$ are $SW_2(n)$, and $SW_2(n+1)$, respectively. By the balance equations (EC.6)-(EC.8),

$$\frac{p_i(n+1)}{p_i(n)} = \frac{p_0(n+1)}{p_0(n)},$$

for all $i \in \{0, 1^-, \ldots, (N-1)^-, 1^+, \ldots, n^+\}$.

In this case, customers balk when they see state variable within $\{n^+, (n+1)^+, \ldots, N\}$. Hence, the social welfare function can be expressed as

$$SW_2(n) = \sum_{i\in\{0,1^-,\ldots,(N-1)^-,1^+,\ldots,(n-1)^+\}} p_i(n)\Lambda(R-\theta W_i).$$

Hence,

$$SW_2(n+1) = \frac{p_0(n+1)}{p_0(n)} SW_2(n) + p_{n^+}(n+1)\Lambda(R-\theta W_{n^+}). \tag{EC.11}$$

Third, we consider the situation $n = N-1$. $SW(n)$ becomes $SW_2(N-1)$ and $SW(n+1)$ becomes $SW_1(N)$. Still it holds that

$$\begin{aligned} SW_1(N) &= \sum_{i\in\{0,1^-,\ldots,(N-1)^-,1^+,\ldots,(N-1)^+\}} p_i(N)\Lambda(R-\theta W_i) \\ &= \frac{p_0(N)}{p_0(N-1)} \sum_{i\in\{0,\ldots,(N-2)^+\}} p_i(N-1)\Lambda(R-\theta W_i) + p_{(N-1)^+}(N)\Lambda(R-\theta W_{(N-1)^+}) \\ &= \frac{p_0(N)}{p_0(N-1)} SW_2(N-1) + p_{(N-1)^+}(N)\Lambda(R-\theta W_{(N-1)^+}), \end{aligned}$$

where the second equality holds by noting that probabilities expressions in (EC.1) and (EC.2) are in the same form as in (EC.6) and (EC.7).

Therefore (EC.10) always hold. Similarly,

$$SW(n)-SW(n-1)=\frac{p_0(n)-p_0(n-1)}{p_0(n-1)}SW(n-1)+p_{n-1}(n)(R-\theta W_{n-1}). \quad \text{(EC.12)}$$

Multiplying both sides of (EC.12) by a factor $\left(\frac{p_0(n+1)-p_0(n)}{p_0(n)}\right)\left(\frac{p_0(n-1)}{p_0(n)-p_0(n-1)}\right)$ yields

$$\left(\frac{p_0(n+1)-p_0(n)}{p_0(n)}\right)\left(\frac{p_0(n-1)}{p_0(n)-p_0(n-1)}\right)(SW(n)-SW(n-1))=\left(\frac{p_0(n+1)-p_0(n)}{p_0(n)}\right)SW(n-1)$$
$$+\left(\frac{p_0(n+1)-p_0(n)}{p_0(n)}\right)\left(\frac{p_0(n-1)}{p_0(n)-p_0(n-1)}\right)p_{n-1}(n)(R-\theta W_{n-1}). \quad \text{(EC.13)}$$

By (EC.1) - (EC.3) and (EC.6) - (EC.8), it can be shown that the following always holds

$$\left(\frac{p_0(n+1)-p_0(n)}{p_0(n)}\right)\left(\frac{p_0(n-1)}{p_0(n)-p_0(n-1)}\right)p_{n-1}(n)=p_n(n+1).$$

From the two equations (EC.10) and (EC.13) and, after some manipulation, one obtains:

$$SW(n+1)-SW(n)=\frac{p_0(n+1)-p_0(n)}{p_0(n)-p_0(n-1)}(SW(n)-SW(n-1))-p_n(n+1)\theta(W_n-W_{n-1}).$$

By (EC.4) and (EC.9), it can be shown that $p_0(n+1)<p_0(n)$ for all n. Hence, the first term on the right-hand side is negative since $SW(n)-SW(n-1)\le 0$ by the assumption; the second term is negative as $W_n>W_{n-1}$. Consequently, $SW(n+1)-SW(n)\le 0$. $\square$

EC.6. Proof of Lemma 1

Proof From (9), it follows that:

$$p_n=\frac{1}{\frac{N}{\rho^{n+1}(\rho^{-N}-1)}-\frac{\rho}{1-\rho}},$$

when $\rho\neq 1$. Since $\rho^N>1+N\ln\rho$ for all $\rho\neq 1$, the derivative of $N/(\rho^{-N}-1)$ with respect to N is negative and hence, the sign of the derivative of p_n with respect to N is positive. When $\rho=1$, $p_n=1/(n-N/2+1.5)$ which is increasing in N.

To prove that W is increasing in N, one can directly take the derivative of the expression of W in (EC.5) (with n^* replaced by n) with respect to N and show that it is positive, which is very tedious. Here, we provide a simple and direct argument that the average queue length is increasing in N, based on stochastic coupling.

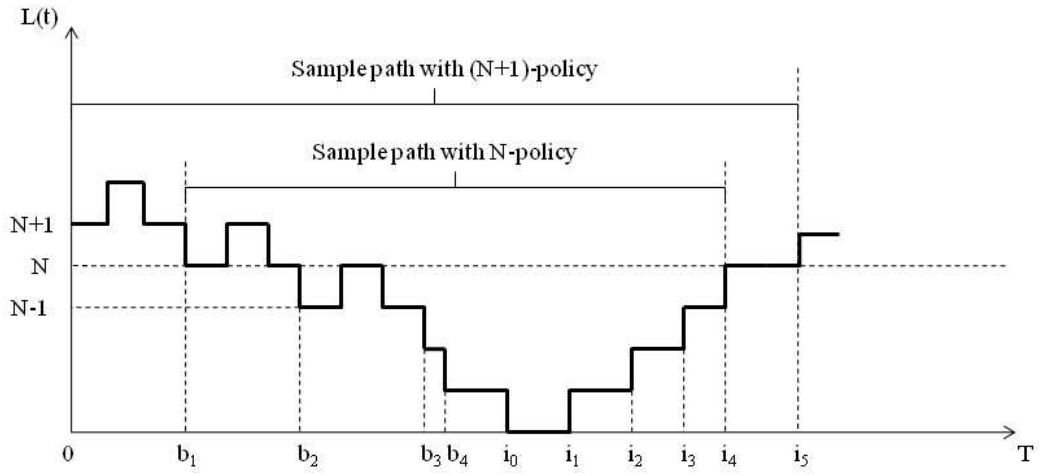

Figure EC.1 Sample path of $M/M/1/n_e$ queue with N-policy

Figure EC.1 shows a sample path of a queue length process (denoted by $L(t)$) of the $M/M/1/n$ queue with a threshold $N+1$; that is, the server starts service from idleness when the queue reaches $N+1$. Suppose that the process starts from $N+1$. $[0, i_0]$ is the busy period and $[i_0, i_5]$ is the idle period. b_1 is the first passage time for the queue to reach N starting from $N+1$ and i_4 is the first passage time for the queue length to increase from N to $N+1$ when the server is idle. Other time notations have similar meanings.

Graphically, the truncated sample path in $[b_1, i_4]$ can be exactly treated as the sample path for the same system but with a threshold N. That is, the $(N+1)$-policy queue has extra parts on $[0, b_1]$ and $[i_4, i_5]$, compared with the sample path of the N-policy queue. Now, if we cut the sample path in $[i_4, i_5]$ and connect it to the left side of the part in $[0, b_1]$, we obtain a sample path which represents a stochastic process that is stochastically larger than the one represented by the sample path in $[b_1, b_2]$. The former process represents a queue which stays at N for an exponential period with rate Λ, then jumps up to $N+1$ and continues until it reaches N again. The latter represents a process which is either identical to the former one, or a process which stays at N for an exponential period with rate μ and jumps down to $N-1$. Therefore, the mean queue length in the combined period $[0, b_1] \cup [i_4, i_5]$ is always larger than the mean queue length in period $[b_1, b_2]$, which in turn, can be similarly shown to be greater than the one in period $[b_2, b_3]$ and any other

period. Consequently, the addition of sample paths in the combined period $[0, b_1] \cup [i_4, i_5]$ makes the average queue length greater.

According to Little's formula, $W = \frac{L}{\Lambda(1-p_n)}$. Since both L and p_n are increasing in N, W is increasing in N. $\square$