Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On Socially Optimal Queue Length

Chia-Li Wang

To cite this article:
Chia-Li Wang (2016) On Socially Optimal Queue Length. Management Science 62(3):899-903. http://dx.doi.org/10.1287/mnsc.2014.2148

MANAGEMENT SCIENCE
Vol. 62, No. 3, March 2016, pp. 899–903
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2148

On Socially Optimal Queue Length

Chia-Li Wang
Department of Applied Mathematics, National Dong Hwa University, Hualien 97401, Taiwan, Republic of China,
cwang@mail.ndhu.edu.tw

Suppose customers arrive at an observable queueing system for service with a utility function of reward and waiting cost. The self- (customer) decision is whether to queue or balk, and the social (system administrator) goal is to maximize the profit of the whole system. Whereas the self-optimal policy is relatively easy to obtain, the socially optimal policy, which is of more practical importance, often requires a tedious and ad hoc analysis as a result of external effects. We will introduce a simple and general approach to determine the optimal admission policy. The main idea of this approach is to consider a special rule that admits an extra customer who is served only by the surplus capacity and bears all the increased waiting time and thus incurs no external cost. The approach applies in principle to queues with exponential service. In fact, for such queues, a marginal analysis based on this rule will explore the properties of the optimal social policy and lead to a general procedure for deriving the optimal threshold.

Keywords: multiserver queue; queueing control; threshold; externality; extra customer; priority; memoryless
History: Received April 2, 2014; accepted December 17, 2014, by Noah Gans, stochastic models and simulation.
Published online in *Articles in Advance* April 24, 2015.

1. Introduction

In a pioneering study on queueing control, Naor (1969) considered a single-server queueing system with Poisson arrivals and exponential service times, the $M/M/1$ first-come-first-serve (FCFS) queue. The control problem is associated with a utility function of fixed reward and linear cost. An individual decision maker seeking personal benefit first observes the number of customers in the system upon arrival and then decides to join for service or balk, depending on expected personal gain. On the other side, an administrator seeks to optimize the system for social benefit.

The individual customer's decision forms the optimal self-threshold, whereas the system administrator's goal is to maintain the socially optimal threshold. Naor (1969) showed that these two optimal thresholds are different—in particular, that the self-threshold is no smaller than the social threshold. This means that an individual customer exercising self-optimization does not optimize the social benefit. Naor demonstrated this difference with an explicit calculation of the expected conditional waiting time, the total time in the system for an entering customer given the number of customers in the system at arrival.

To clarify that this discrepancy is not associated with the risk of the individual's decision based on the *expected* waiting cost, a subsequent study (Adler and Naor 1969) considered the $M/D/1$ FCFS queue where service times are constant and exhaustive information (remaining service time, if any, and queue length) of the system state is available. An individual customer now faces a decision problem under certainty. It turns out that both optimal thresholds are real numbers on the waiting time. Again, the inequality between the self-threshold and the social threshold is shown to hold.

The reason for the difference is that the socially optimal policy balances the gain to the individual joining (or not joining) the queue with the loss to others as a result of this action, whereas the optimal self-threshold policy does not. This phenomenon is often referred to by economists as *external effects*, the consequences of an action by one individual or group on others.

Shortly thereafter, Yechiali (1972) and Simonovits (1976) extended the scope of the study to $GI/M/c$ FCFS queues. They showed that the socially optimal policy is of the following form: admit if and only if the queue length is smaller than some specific number, and the same inequality between the self-optimal and socially optimal thresholds must hold. In particular, a special admission rule was constructed in Simonovits to prove the monotonicity of conditional expected waiting time and admission rate. A customer who observes a critical number of customers in the system at his arrival is admitted as a "polite" customer unless another polite customer is waiting. This customer will be served only when no other customer is waiting for service and will be preempted if necessary; then service will be resumed. This is a clever and promising approach for the study of the socially optimal policy. Yet it is somewhat surprising that this

implication of the admission rule to the social optimality itself did not get fully explored.

Nevertheless, a similar approach was proposed by Xu and Shanthikumar (1992) for the admission control of an $M/M/c$ order-entry FCFS queue. They approached the problem by constructing a dual system: an $M/M/c$ order-entry preemptive last-come-first-serve (PL) queue that is subject to expulsion control, which has the same optimal control policy as the original system. The reason for working on the dual system is similar to Simonovits (1976) in that a present customer does not impose external effects on future customers. As a result, the customer's decision on when to renege is socially optimal. It is worth mentioning that this idea was originally proposed by Hassin (1985) for an $M/M/1$ PL queue. For the most complete and updated reference on this topic, see Hassin and Haviv (2003).

Generally speaking, whereas the interesting and challenging issue for self-policy is on its state of equilibrium in economics, of interesting and practical importance to social policy is its optimality in an operation that often requires a tedious and ad hoc analysis as a result of externality. So we aim to conduct a general and intuitive analysis of the optimal social policy in order to understand its properties and behavior.

To this end, the approach we take is based on the same admission rule as the one in Simonovits (1976), where the polite customer is called an *extra customer*. It applies in principle to queues with exponential service. The main results of this study come from the fact that the extra customer incurs all the externality; he has no effect on others. Consequently, the socially optimal threshold can be derived from his expected waiting time in the system.

2. The Extra Customer

Consider a c-server queue, $c \geq 1$, where customers arrive from a renewal arrival process at a rate of λ, service times are independent and exponentially distributed with a mean of $1/\mu$, there is a $GI/M/c$ queue, and the service discipline is nonidling; i.e., servers cannot be idle whenever there is at least one unattended customer. Assume that the number of present customers in the system is observable by arriving customers and known to the system administrator.

Suppose that the system limits the number of customers in it to $N-1$ ($\geq c$) by rejecting the arrivals when $N-1$ customers are already in it, called an $(N-1)$-limit system. We are interested in measuring the increase in the waiting time of customers when the limit is raised by 1 to N. To this end, we introduce a surplus-capacity admission rule: We define as *ordinary* customers those served at a c-server queue with queue limit $N-1$. Hence, ordinary customers are those who on arrival find $N-2$ or fewer customers in the system. Now, raise the queue limit to N. Arrivals finding $N-1$ customers in the system, all ordinary, are served. They are called *extra* customers.

Ordinary customers are served according to the service discipline among themselves. On arrival, they have preemptive priority over any extra customer found in service and on service completion have priority over any extra customer in the queue. Thus, ordinary customers are served in exactly the same manner that they were in the original $(N-1)$-limit system, and the extra customer is served only when there is no unattended ordinary customer in the system and there is an idle server—that is, when there is surplus service capacity. The interrupted extra customer resumes service at the point of interruption when a server is free and there are no ordinary customers in the queue.

Note that arrivals finding $N-1$ customers in the system, where fewer than $N-1$ are ordinary, are ordinary customers. Extra customers are those served only because the queue limit was raised; at all times, there is at most one extra customer in the system.

By design, the extra customer does not interfere with nor delay any ordinary customers; the ordinary customers enter, get served, and leave exactly at the same epochs of time as they do in the $(N-1)$-limit system for each sample path. So the extra customer is the worst-treated customer as far as waiting time is concerned.

Denote the expected waiting time of the extra customer by w_E. The next example demonstrates how we compute w_E.

Example 1. Consider an $M/M/1$ queue with arrival rate λ and service rate μ. Since the extra customer can be served only during the idle period of the embedded $M/M/1/N-1$ queue of ordinary customers, his waiting time is composed of three parts: the duration between the arrival and commencement of service, his own service time, and possible busy periods of the ordinary customers.

To compute w_E, we construct a birth and death process with state space $\{e, 0, 1, \ldots, N-1\}$, where state i, $0 \leq i \leq N-1$, means that there are i ordinary customers and one extra customer in the system and e means that the system is empty. Let T_i denote the time it takes to enter state $i-1$ from state i, $i \geq 1$, and let T_0 denote the time it takes to enter state e from 0 during which the extra customer is in service. Then, we can write

$$w_E = E(T_{N-1} + T_{N-2} + \cdots + T_0).$$

Since T_{N-1} is exponential with rate μ, we have $E(T_{N-1}) = 1/\mu$. For $0 \leq i < N-1$, we condition

whether the next transition takes the process into state $i-1$ or $i+1$ to get

$$E(T_i \mid \text{next state is } i-1) = \frac{1}{\mu+\lambda},$$

$$E(T_i \mid \text{next state is } i+1) = \frac{1}{\mu+\lambda} + E(T_{i+1}) + E(T_i).$$

Multiplying the above by associated probabilities and summing, we obtain the following recursive formula:

$$E(T_i) = \frac{1}{\mu(1+\rho)} + \frac{\rho}{1+\rho}[E(T_{i+1}) + E(T_i)] = \frac{1}{\mu} + \rho E(T_{i+1}),$$

where $\rho = \lambda/\mu$. Simple algebra yields

$$E(T_i) = \frac{1}{\mu}(1+\rho+\cdots+\rho^{N-i}) = \frac{1}{\mu}\frac{1-\rho^{N-i+1}}{1-\rho}, \quad i = 0, 1, \ldots, N-1,$$

and consequently,

$$w_E = \frac{1}{\mu(1-\rho)^2}[N(1-\rho) - \rho(1-\rho^N)].$$

Because exponential service has the memoryless property, the remaining service times of present customers are stochastically equivalent (i.e., governed by the same distribution), as is the number-of-customers-in-system process of the N-limit system with the extra customer stochastically equivalent to the same quantity of the *standard* N-limit system where all customers are served by the same service discipline. Thus, the system has work conservation, and some quantities, such as the arrival rate of served customers λ_N when the queue limit is N and the corresponding mean number-of-customers-in-system L_N, are unchanged by our "singling out" of extra customers.

Furthermore, when the system limit is raised from $N-1$ to N, the total waiting time increases but the average waiting time of ordinary customers does not. In other words, the extra customer alone bears the increase of the waiting time in the system resulting from the raise and causes no externality.

The lemma below contains a key identity that is obtained by Little's law.

Lemma 1. *For a GI/M/c queue,*

$$(\lambda_N - \lambda_{N-1})w_E = L_N - L_{N-1} \tag{1}$$

for all $N \geq 1$.

Proof. Because the arrival rate of extra customers is the increment in the rate of an $(N-1)$-limit system up to the rate of an N-limit system—that is, $\lambda_N - \lambda_{N-1}$—and by Little's law, the mean number of extra customers in the system is equal to $(\lambda_N - \lambda_{N-1})w_E$. Then, (1) follows from the fact that the mean number of ordinary customers in the system is L_{N-1}. □

Recall that in Example 1 the w_E is derived through the computation of the expected busy period formed by ordinary customers. When (1) holds, it is more efficient to compute w_E with available λ_N and L_N.

It should be noted that when the service time does not have the memoryless property, preemption may affect the mean number of customers in the system such that (1) does not hold and the arrival rate of the extra customer is not $\lambda_N - \lambda_{N-1}$.

Now, we incorporate the queueing system with the same utility function in Naor (1969); i.e., the customer earns a constant reward R upon service completion and incurs cost with a constant rate C while in the system. A system administrator decides whether to admit or to reject the customer in order to maximize the social benefit—namely, the long-run average of aggregate net gains of admitted customers per unit of time. This decision problem clearly satisfies the conditions of Lemma 1.5 in Ross (1968) as a denumerable-state Markovian decision process that has a deterministic optimal policy for an average cost criterion. Therefore, the social optimal policy is a threshold that, from Little's law, maximizes $\lambda_N R - CL_N$ with respect to N.

Since the cost is proportional to the average waiting time and customers have identical reward and cost rates, the arrival rate of admitted customers and average waiting time determine the social benefit. When these two quantities are constant, how the total waiting time is divided among the customers is irrelevant. This fact allows us to study the socially optimal threshold via the expected waiting time of the extra customer.

Indeed, (1) implies that

$$(\lambda_N - \lambda_{N-1})(R - Cw_E) = (\lambda_N - \lambda_{N-1})R - C(L_N - L_{N-1}). \tag{2}$$

In other words, the expected gain of an extra customer multiplied by the arrival rate of such a customer to the system is equivalent to the *marginal social benefit* of the system at threshold $N-1$. Thus, if the expected gain of the extra customer is positive, then the system limit should be raised by 1 to N.

3. A Priority Queue of Extra Customers

The extra-customer idea is worth exploring further. In fact, one can construct a further admission rule in which *every* arriving customer is admitted and assigned a different *priority* of service over any

present customers in the system, creating a priority queue of extra customers. Specifically,

• An admitted customer is an i-extra customer, where $i \geq 0$ is the *smallest* nonnegative integer such that there is no i-extra customer in the system. For example, a customer who is admitted when the system is empty is a 0-extra customer, and the next admitted customer while the 0-extra customer is still in the system is a 1-extra customer, or a 0-extra customer otherwise; an admitted customer who finds only a 0-extra and a 2-extra customer in the system is a 1-extra customer.

• When competing for a server, a lower-indexed extra customer has preemptive priority over any higher-indexed one. In particular, a j-extra customer is served only when there is (i) no lower-indexed extra customer, i.e., an i-extra customer with $i < j$, waiting for service and (ii) either an idle server or all servers are busy but at least one is serving a k-extra customer with $k > j$. In the latter case, the j-extra customer will preempt the highest-indexed extra customer in service. The service will be interrupted when a lower-indexed extra customer is admitted with no idle server and j is the highest-indexed one in service.

Similar to the previous surplus-capacity admission rule, for any i there is at most one i-extra customer at any time. When (1) holds, for fixed $j \geq 0$, any i-extra customer with $i < j$ is like an ordinary customer to the j-extra customer so that the j-extra customer is admitted and served as the previously described extra customer of the j-limit system. Thus, the respective mean waiting times of the j-extra customer, denoted as $w_E(j)$, and the extra customer are the same, and so are their respective arrival rates. However, there is no need to put a system limit under this admission rule.

An immediate consequence of this construction and (1) is as follows.

Theorem 1. *For a $GI/M/c$ queue, we have*

$$L_N = \sum_{i=1}^{N} (\lambda_i - \lambda_{i-1}) w_E(i) \tag{3}$$

for all $N \geq 1$.

Because $w_E(i)$ clearly increases in i, when (3) holds, (2) implies that the long-run average of the social benefit is unimodal in N. Furthermore, as the expected net gain of the i-extra customer decreases in i, it changes from positive to negative as i increases, and *the number at which this sign changes is the socially optimal threshold*. Hence, the self-threshold of the extra customers in the priority queue is the socially optimal threshold of the original queue: the largest N satisfying $R - Cw_E(N) \geq 0$. We conclude by stating the following result.

Theorem 2. *The socially optimal threshold has the form of*

$$n_o = \max\{N\colon w_E(N) \leq R/C\}. \tag{4}$$

Given that the extra customer is the worst-treated customer, the inequality between self-optimal and socially optimal thresholds follows from (4) directly. Moreover, when λ_i and L_i are available, the derivation of n_o becomes easier than differentiation, the usual method. The next example demonstrates the derivation.

Example 2. For an $M/M/c/K$ queue, the stationary probability of i customers in the system, $p_K(i)$, has a closed form:

$$p_K(0) = \left[\frac{(c\rho)^c}{c!} \frac{1-\rho^{K-c+1}}{1-\rho} + \sum_{i=0}^{c-1} \frac{(c\rho)^i}{i!} \right]^{-1} \quad \text{and}$$

$$p_K(i) = \begin{cases} \dfrac{c^i \rho^i}{i!} p_K(0), & i = 1, 2, \ldots, c; \\ \dfrac{c^c \rho^i}{c!} p_K(0), & i = c+1, 2, \ldots, K, \end{cases}$$

where $\rho = \lambda/(c\mu)$. Then, with

$$\lambda_K = \lambda(1 - p_K(K)) \quad \text{and}$$

$$L_K = \rho \sum_{i=c}^{K-1} (i - c + 1) p_K(i) + c\rho(1 - p_K(K)),$$

one readily gets from (1) that

$$w_E(N) = \frac{\sum_{i=c}^{N-1} (i+c-1) p_N(i) - \sum_{i=c}^{N-2} (i+c-1) p_{N-1}(i)}{c\mu[p_{N-1}(N-1) - p_N(N)]} + \frac{1}{\mu}.$$

Plugging the above into (4) yields

$$n_o = \max\left\{ N\colon (N-c+1)(1-\rho) - \frac{\rho(1-\rho^{N-c+1})(c\rho)^c/c!}{(c\rho)^c/c! + (1-\rho)\sum_{i=0}^{c-1}(c\rho)^i/i!} \leq c(\mu R/C - 1)(1-\rho)^2 \right\}.$$

Note that when a customer joins a standard $M/M/c/N$ FCFS queue with $N-1$ customers present, $N/(c\mu)$ is his expected wait and $w_E(N) - N/(c\mu)$ is the others' increased expected wait as a result of his

action. The latter multiplied by waiting cost rate is the associated external cost.

4. Concluding Remarks

The contribution of this study comes from exploring the implication of the extra-customer approach to the social optimality beyond that presented in Simonovits (1976). Under the assumption of exponential service, the expected waiting time of an extra customer is equal to the expected waiting time of a customer entering with the same condition upon arrival plus the external effect caused by him. This fact enables us to relate the marginal social benefit to the extra customer's personal gain. With the inclusion of a constructed priority queue of extra customers that is stochastically invariant to the original system, the marginal analysis shows that the self-threshold of extra customers is the socially optimal threshold. This finding explains how social optimization is reached and explains the monotonic and deterministic properties of the socially optimal policy, and it develops an efficient procedure for deriving the socially optimal threshold.

Acknowledgments

The author is grateful to Yi-Ching Yao for referring the work of András Simonovits and to Takashi Ishikida for many helpful comments to improve the presentation of this study.

References

Adler I, Naor P (1969) Social optimization versus self-optimization in waiting lines. Technical Report 126, Department of Operations Research, Stanford University, Stanford, CA.

Hassin R (1985) On the optimality of first come last served queues. *Econometrica* 53(1):201–202.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Kluwer Academic Publishers, Boston).

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

Ross SM (1968) Non-discounted denumerable Markovian decision models. *Ann. Math. Stat.* 39(2):412–423.

Simonovits A (1976) Self- and social optimization in queues. *Studia Scientiarum Mathematicarum Hungarica* 11:131–138.

Xu SH, Shanthikumar JG (1992) Optimal expulsion control—A dual approach to admission control of an ordered-entry system. *Oper. Res.* 41(6):1137–1152.

Yechiali U (1972) Customers' optimal joining rules for the $GI/M/s$ queues. *Management Sci.* 18(7):434–443.