MANAGEMENT SCIENCE
Vol. 55, No. 10, October 2009, pp. 1729–1742
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5510 | 1729

informs

DOI 10.1287/mnsc.1090.1041

Real-Time Delay Estimation in Overloaded Multiserver Queues with Abandonments

Rouba Ibrahim, Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
{rei2101@columbia.edu, ww2040@columbia.edu}

We use heavy-traffic limits and computer simulation to study the performance of alternative real-time delay estimators in the overloaded $GI/GI/s+GI$ multiserver queueing model, allowing customer abandonment. These delay estimates may be used to make delay announcements in call centers and related service systems. We characterize performance by the expected mean squared error in steady state. We exploit established approximations for performance measures with a nonexponential abandonment-time distribution to obtain new delay estimators that effectively cope with nonexponential abandonment-time distributions.

Key words: delay estimation; delay announcements; call centers; many-server queues; customer abandonment; simulation; heavy traffic
History: Received June 9, 2008; accepted March 24, 2009, by Michael Fu, stochastic models and simulation. Published online in *Articles in Advance* July 20, 2009.

1. Introduction

We investigate alternative ways to estimate, in real time, the delay (before entering service) of an arriving customer in a service system with customer abandonment. We do this primarily so that delay announcements can be made to arriving customers. Delay announcements can be helpful when queues are invisible to customers, as in call centers; see Gans et al. (2003) and Aksin et al. (2007) for background on call centers.

Comparing alternative delay estimators is complicated. Naturally, we would like to have a delay estimator that is effective. We quantify the effectiveness of a delay estimator by the mean squared error (MSE). Because the estimator typically depends on state information, we use the expected MSE, considering the steady-state distribution of the state information, which we estimate via simulation by computing the average squared error (ASE), averaging over a large number of customers in steady state. A lower expected MSE (or ASE) corresponds to a more effective delay estimator.

But, we would also like to have a simple delay estimator, which can be easily implemented in a real-life system, i.e., one that uses information that is readily available. Alternative delay estimators differ in the type and amount of information that their implementation requires. For example, this information may involve the model, the system state upon arrival, or the history of delays in the system. An important insight, which applies broadly, is that simplicity and ease of implementation are often obtained at the expense of statistical precision.

Our main contributions are (i) to propose new, effective, and simple ways to do better delay estimation in overloaded, many-server queues with customer abandonment; (ii) to establish heavy-traffic limits that generate approximations for the expected MSE of some delay estimators; and (iii) to describe results of simulation experiments evaluating alternative delay estimators. We obtain more effective delay estimators by exploiting approximations for performance measures in many-server queues with a nonexponential abandonment-time distribution, from Whitt (2005, 2006).

1.1. Queueing Model

We study the performance of alternative real-time delay estimators by considering the steady-state behavior of an overloaded $GI/GI/s + GI$ queueing model, allowing customer abandonment. This model has independent and identically distributed (i.i.d.) interarrival times with mean λ^{-1} and a general distribution. We only use the i.i.d. assumption for the interarrival times when simulating the model; it is not required for the implementation of our delay estimators. Service times are i.i.d. with mean μ^{-1} and a general distribution. Each arriving customer will abandon if he is unable to start service before a random time with mean α^{-1} and a general distribution. Abandonments times are i.i.d.; the arrival, service, and abandonment processes are all mutually independent. There is unlimited waiting space and arriving

customers are served in order of arrival; i.e., we use the first-come-first-served service discipline. The traffic intensity is $\rho \equiv \lambda/s\mu$.

We focus on overloaded scenarios, in which the arrival rate exceeds the maximum possible total service rate. Customer abandonment makes the system stable in this case. (That can be proved by bounding the model above by the $GI/GI/\infty$ model obtained by removing all servers; then the abandonment times can be thought of as service times. For more on the stability of the $GI/GI/\infty$ model, see p. 178 of Whitt 1982.) We consider overloaded systems because we are primarily interested in estimating delays when they are large. Many call centers are overloaded some of the time, especially service-oriented ones in which emphasis is placed on efficiency rather than on quality of service.

1.2. Queue-Length-Based and Delay-History-Based Delay Estimators

We consider both queue-length-based and delay-history-based delay estimators. Queue-length-based delay estimators exploit system-state information including the queue length (number of waiting customers) seen upon arrival. In contrast, delay-history-based estimators only exploit information about recent customer delay history in the system. These delay estimators are appealing because they are easy to interpret, and because they are simple and robust, applying to a broad range of models, without requiring knowledge of the model or its parameters; e.g., number of servers, mean service time, and arrival rate.

1.3. Customer Response to Delay Announcements

We envision our delay estimates being used to make delay announcements to arriving customers: Each delayed customer, upon arrival, is given a single-number delay estimate of that customer's delay until he can start service. Customers typically respond to delay announcements, and their response alters system performance. For example, some customers may elect to balk, upon arrival, in response to a delay announcement. As a result, the arrival rate to the system would become state dependent. Moreover, customers who decide to stay may have different abandonment behavior in response to the announcement. They may become increasingly impatient if they have to wait more than their announced delay. As a result, the abandonment distribution of customers in queue would depend on their elapsed waiting time. Changes in system performance alter, in turn, the delay estimates given. As discussed by Armony et al. (2009), studying customer responses to delay announcements requires an equilibrium analysis. However, it is not clear whether an equilibrium exists, or how to fully characterize it. There may even be multiple equilibria.

Here, we do not directly consider customer response. We think of our delay estimates being based on model information obtained after equilibrium has been reached (with the announcements being used). More generally, we regard our work as an essential first step toward studying the performance impact of delay announcements in the $GI/GI/s+GI$ model. It is not hard to see how the delay estimation methods of this paper can be applied to the more complicated setting involving customer response. Indeed, the delay-history-based estimators directly account for customer response because they depend on the history of delays in the system, which in turn is affected by customer response.

The queue-length-based estimators can also be extended to account for changes in customer behavior. For example, we could use an iterative simulation-based algorithm to develop approximations of the equilibrium steady-state performance of the $GI/GI/s+GI$ model with delay announcements. During each iteration, we would give real-time delay estimates to arriving customers, and model their response. We would then reestimate model parameters that are affected by customer response, and feed these new estimates into the subsequent iteration. The algorithm would continue until the observed difference between successive estimates of model parameters is negligible. It is significant that our proposed queue-length-based estimators apply directly to the successive iterations of this algorithm, because the queueing model in each iteration is a $GI/GI/s+GI$ model with a different set of parameters. There remains, however, to determine appropriate regularity conditions under which this algorithm terminates, i.e., under which there exists a unique equilibrium in the system.

1.4. Actual and Potential Waiting Times

As in Baccelli et al. (1984) and Garnett et al. (2002), we need to distinguish between the actual and potential waiting times of a given delayed customer in a queueing model with customer abandonment. A customer's actual waiting time is the amount of time that this customer spends in queue, until he either abandons or joins service, whichever comes first. A customer's potential waiting time is the delay he would experience, if he had infinite patience (quantified by his abandon time). For example, the potential waiting time of a delayed customer who finds n other customers waiting ahead in queue upon arrival, is the amount of time needed to have $n+1$ consecutive departures from the system (either service completions or abandonments from the queue). In this study,

we estimate the potential waiting times of delayed customers.

1.5. Quantifying Performance: Average Squared Error (ASE)

In our simulation experiments, we quantify the performance of a delay estimator by computing the ASE, defined by

$$\text{ASE} \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - e_i)^2, \tag{1}$$

where $p_i > 0$ is the potential waiting time of delayed customer i, e_i is the delay estimate given to customer i, and k is the number of customers in our sample. In our simulation experiments, we measure p_i for both served and abandoning customers. For abandoning customers, we compute the delay experienced, had the customer not abandoned, by keeping him "virtually" in queue until he would have begun service. Such a customer does not affect the waiting time of any other customer. The ASE should approximate the expected MSE in steady state.

1.6. Mean Squared Error (MSE)

Let $W_Q(n)$ represent a random variable with the conditional distribution of the potential delay of an arriving customer, given that this customer must wait before starting service, and given that the queue length at the time of his arrival, t, not counting himself, is $Q(t) = n$. (In this framework, the event "$Q(t) = 0$" corresponds to all servers being busy and our arriving customer being the first in queue.) Let $\theta_{\text{QL}}(n)$ be some given single-number delay estimate that is based on the queue length, n. Then, the MSE of the corresponding delay estimator is given by

$$\text{MSE} \equiv \text{MSE}(\theta_{\text{QL}}(n)) \equiv E[(W_Q(n) - \theta_{\text{QL}}(n))^2].$$

The MSE of a queue-length-based delay estimator is a function of n, the number of customers seen in queue upon arrival. By looking at the ASE, we are looking at the expected MSE averaging over all n, where the arrival must wait, in steady state. It is known that the conditional mean, $E[W_Q(n)]$, minimizes the MSE. Unfortunately, it is often difficult to find a closed-form expression for this mean, so we develop approximations of it.

1.7. Root Relative Squared Error

In addition to the ASE, we quantify the performance of a delay estimator by computing the root relative average squared error (RRASE), defined by

$$\text{RRASE} \equiv \frac{\sqrt{\text{ASE}}}{(1/k)\sum_{i=1}^{k} p_i}, \tag{2}$$

using the same notation as in (1). The denominator in (2) is the average potential waiting time of customers who must wait. For large samples, the RRASE should agree with the expected root relative mean squared error (RRMSE), in steady state. The RRASE and RRMSE are useful because they measure the effectiveness of an estimator relative to the mean, so that they are easy to interpret.

1.8. Related Literature

This paper is an extension of Ibrahim and Whitt (2009), which studies the performance of a wide range of alternative real-time delay estimators in the $GI/M/s$ queueing model (without abandonment), both analytically and numerically using computer simulation. That paper in turn builds on Whitt (1999) and Armony et al. (2009). Whitt (1999) discusses the possibility of making reliable delay estimations by exploiting information about the current state of the system. Armony et al. (2009) discuss the motivation for the last-to-enter-service delay estimator, whose performance we study here, and changes in customer behavior that result from such an announcement. Related literature is already discussed there.

Very recent papers on delay estimation include Jouini et al. (2007), Guo and Zipkin (2007), and Allon et al. (2007). For a review of the growing literature on delay estimation and delay announcements, see §2 of Jouini et al. (2007).

1.9. Organization of the Paper

The remainder of this paper is organized as follows: In §2, we describe a no-information (NI) delay estimator in the efficiency-driven many-server heavy-traffic limiting regime, which serves as a useful reference point. In §3, we define new queue-length-based delay estimators, and discuss relevant results. In §4, we briefly describe alternative delay-history-based delay estimators; a more complete description can be found in Ibrahim and Whitt (2009). In §5, we establish heavy-traffic limits for several delay estimators in the $G/M/s + M$ model. In §6, we present simulation results for the $M/M/s + GI$ model. In §7, we make concluding remarks and describe managerial insights. In §5, we postpone one long proof of a result (Theorem 4) to the e-companion[1], where we also present additional supporting material. Supplementary material is provided in the online supplement, which is available at http://www.columbia.edu/∼ww2040/.

2. A Theoretical Reference Point

An important theoretical reference is the many-server heavy-traffic limit for the number in the system in the Markovian $M/M/s + M$ queue with customer abandonment, in the efficiency-driven (ED) regime, as discussed in Garnett et al. (2002), Whitt (2004), and

[1] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Talreja and Whitt (2009). That limit describes how the model behaves as the arrival rate λ and number of servers s increase, whereas the individual service rate μ and individual abandonment rate α remain unchanged, with the traffic intensity held fixed at a value $\rho \equiv \lambda/s\mu > 1$. (There are also some results for the more general $GI/GI/s+GI$ model in the ED regime in Zeltyn and Mandelbaum 2005 and Whitt 2006.)

Let $W_s(\infty)$ represent the steady-state waiting time as a function of s in the ED regime, and let $\Rightarrow$ denote convergence in distribution. Whitt (2004) shows that

$$W_s(\infty) \;\Rightarrow\; w \equiv \frac{1}{\alpha}\ln(\rho) > 0 \quad \text{as } s \to \infty, \tag{3}$$

whereas Theorem 6.1 of Zeltyn and Mandelbaum (2005) (Theorem 5) and Theorem 6.4 of Talreja and Whitt (2009) show that

$$\sqrt{s}(W_s(\infty) - w) \;\Rightarrow\; N(0, 1/\alpha\mu) \quad \text{as } s \to \infty, \tag{4}$$

where $N(m, \sigma^2)$ denotes a normal random variable with mean m and variance σ^2. These limits lead to the deterministic fluid approximation $W_s(\infty) \approx w$ and the stochastic refinement $W_s(\infty) \approx N(w, 1/s\alpha\mu)$.

The deterministic fluid approximation w in (3) and the steady-state mean $E[W_s(\infty)]$ it approximates are candidate NI estimators, θ_{NI}, paralleling the NI estimator for the $GI/M/s$ model considered as a reference point in Ibrahim and Whitt (2009). In fact, the NI estimator is much more appealing now, because it is much more effective with customer abandonment than without. Based on the limits aforementioned (plus appropriate uniform integrability, which can also be established), we have

$$\mathrm{MSE}(\theta_{\mathrm{NI}}) \approx \mathrm{Var}(W_s(\infty)) \approx \frac{1}{s\alpha\mu} \to 0 \quad \text{as } s \to \infty. \tag{5}$$

Unlike in the $GI/M/s$ model, here the squared coefficient of variation (SCV, variance divided by the square of the mean), c^2_{NI}, is asymptotically negligible as well, because here $E[W_s(\infty)] \to w > 0$ as $s \to \infty$. (For the $GI/M/s$ model considered in Ibrahim and Whitt 2009, $c^2_{\mathrm{NI}} \to 1$ as $\rho \uparrow 1$ for all s.) The limit in (5) implies that any reasonable estimator should be effective in the ED regime as s gets larger. We will want to see that our proposed estimators outperform NI as well as become effective as s increases.

3. Queue-Length-Based Delay Estimators

In this section, we describe alternative estimators based on the queue length seen upon arrival to the system. The information needed for the implementation of each of these queue-length-based estimators is summarized in Table 1.

Table 1 **Summary of Information Required for Implementation of Each Queue-Length Based Delay Estimator**

	Information about the model
QL	$Q(t)$, s, μ
QL_r^m	$Q(t)$, s, μ, α
QL_r	$Q(t)$, s, μ, $F(x)$, λ
QL_m	$Q(t)$, s, μ, α
QL_{ap}	$Q(t)$, s, μ, $F(x)$, λ

3.1. Simple Queue-Length-Based Delay Estimator (QL)

For a system having s agents, each of whom on average completes one service request in μ^{-1} time units, we may predict that a customer, who finds n customers in queue upon arrival, will be able to begin service in $(n+1)/s\mu$ minutes. Let QL refer to this simple queue-length-based estimator, commonly used in practice. Let the estimator, as a function of n, be

$$\theta_{\mathrm{QL}}(n) \equiv (n+1)/s\mu. \tag{6}$$

The QL estimator is appealing because of its simplicity and ease of implementation: it uses information about the system that usually is readily available. In Ibrahim and Whitt (2009), the performance of QL is studied in the $GI/M/s$ model, where there is no customer abandonment. For that model, $W_{\mathrm{Q}}(n)$ is the time necessary to have exactly $n+1$ consecutive departures from service (service completions). But, the times between successive service completions, when all servers are busy, are i.i.d. random variables distributed as the minimum of s exponential random variables, each with mean μ^{-1}, which makes them i.i.d. exponential with mean $1/s\mu$. The optimal delay estimator, using the MSE criterion, is the one announcing the conditional mean, $E[W_{\mathrm{Q}}(n)]$. But, following the analysis above, $E[W_{\mathrm{Q}}(n)] = \theta_{\mathrm{QL}}(n)$ in (6). Hence, QL is optimal for the $GI/M/s$ model, under the MSE criterion. Extensive simulation experiments in Ibrahim and Whitt (2009) show the superiority of QL in that simple idealized setting.

When there is customer abandonment, the QL estimator overestimates the potential delay, because customers in queue may abandon before entering service, and QL fails to take that into account. That is confirmed by our simulation results in §6, but we now analytically quantify the effect for the Markovian $M/M/s+M$ model. To do so, we use the steady-state fluid approximations to the $M/M/s+M$ model in the ED regime discussed in §2. In the steady-state fluid limit, all served customers wait the same deterministic amount of time w in (3) and they all see the same number of customers, q, in queue upon arrival. From (2.26) of Whitt (2004),

$$q = \frac{s\mu}{\alpha}(\rho - 1). \tag{7}$$

In the fluid limit,

$$\theta_{\mathrm{QL}}(q) = \frac{q+1}{s\mu} \approx \frac{q}{s\mu} = \frac{1}{\alpha}(\rho - 1) > w = \frac{1}{\alpha}\ln(\rho).$$

Consistent with intuition, we see that QL overestimates w. Indeed,

$$\frac{\theta_{\mathrm{QL}}(q) - w}{w} = \frac{(\rho-1)/\alpha - \ln(\rho)/\alpha}{\ln(\rho)/\alpha}; \tag{8}$$

e.g., there is 10% relative error when $\rho = 1.2$, 19% relative error when $\rho = 1.4$, and much greater error when ρ is larger. (Exploiting the asymptotic expansion of the logarithm: $\ln(1+\delta) \approx \delta - \delta^2/2$ when δ is small, we can obtain the simple rough approximation to (8) of $(\rho-1)/(3-\rho) \approx (\rho-1)/2$ when ρ is slightly greater than 1.)

Motivated by the simple form of the QL delay estimate, $\theta_{\mathrm{QL}}(n)$ in (6), we now propose modified queue-length-based delay estimators that account for customer abandonment, and that are also easy to implement in practice.

3.2. Markovian Queue-Length-Based Delay Estimator (QL_m)

As in Whitt (1999), this estimator QL_m approximates the $GI/GI/s+GI$ model by the corresponding $GI/M/s+M$ model with the same service-time and abandon-time means. For the $GI/M/s+M$ model, we have the representation

$$W_{\mathrm{Q}}(n) \equiv \sum_{i=0}^{n} Y_i, \tag{9}$$

where the Y_i are independent random variables with Y_i being the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service) and i exponential random variables with rate α (corresponding to the abandonment times of the remaining customers waiting in line). That is, Y_i is exponential with rate $s\mu + i\alpha$. (Because $W_{\mathrm{Q}}(n)$ is the sum of independent exponential random variables, it has a hypoexponential distribution.) Therefore,

$$E[W_{\mathrm{Q}}(n)] = \sum_{i=0}^{n} E[Y_i] = \sum_{i=0}^{n} \frac{1}{s\mu + i\alpha}. \tag{10}$$

The QL_m estimator given to a customer who finds n customers in queue upon arrival is $\theta_{\mathrm{QL}_m}(n) \equiv E[W_{\mathrm{Q}}(n)]$. Under the MSE criterion, QL_m is the best possible in the $GI/M/s+M$ model, but we find that it is not always so good for the more general $GI/GI/s+GI$ model. Nonexponential service-time and abandonment-time distributions are commonly observed in practice; see Brown et al. (2005), and Mandelbaum and Zeltyn (2004, 2007). It is therefore important to propose other queue-length-based delay estimators that effectively cope with nonexponential distributions. Approximations are needed because direct mathematical analysis is difficult. Next, we propose two queue-length-based delay estimators, QL_r and QL_{ap}, exploiting approximations for performance measures in many-server queues with a nonexponential abandonment-time distribution, developed in Whitt (2005, 2006).

3.3. Simple-Refined Queue-Length-Based Delay Estimator (QL_r)

We now propose a simple refinement of QL by making use of the steady-state fluid approximations to the general $G/GI/s+GI$ model, in the ED limiting regime, as developed by Whitt (2006). For that purpose, let F be the cumulative distribution function (cdf) of the abandon-time distribution, and let F^c be the complementary cdf associated with F. (That is, $F^c(t) = 1 - F(t)$, for all t.) In this steady-state fluid limit, the deterministic waiting time w and the deterministic queue length q are given by Equations (3.6) and (3.7) of Whitt (2006), which we restate. Because "rate in" $\equiv \lambda F^c(w) = s\mu \equiv$ "rate out", we have

$$\rho F^c(w) = 1. \tag{11}$$

The associated equation for q is

$$q = \lambda \int_0^w F^c(x)\,dx = s\rho\mu \int_0^w F^c(x)\,dx. \tag{12}$$

In the fluid limit, QL estimates a customer's delay as the deterministic quantity:

$$\theta_{\mathrm{QL}}(q) = \frac{q+1}{s\mu} \approx \frac{q}{s\mu} = \rho \int_0^w F^c(x)\,dx.$$

For QL_r, we propose computing the ratio $\beta = w/(q/s\mu) = ws\mu/q$ (after solving numerically for w and q), and using it to refine the QL estimator. That is, the new delay estimate is

$$\theta_{\mathrm{QL}_r}(n) \equiv \beta \times \theta_{\mathrm{QL}}(n) = \beta(n+1)/s\mu.$$

The QL_r estimator is appealing because it is only a minor modification of the QL estimator, but performs much better in models with customer abandonment, as we show in §6. In particular, it is remarkably effective with nonexponential abandonment-time distributions. Note that in addition to s, n, and μ, we need to know ρ or, equivalently, λ, and the abandonment-time cdf F to implement QL_r.

3.4. Exponential Abandonment Case (QL_r^m)

We now propose a modification of QL_r which does not depend on ρ. It is based on assuming that the abandonment-time cdf F is exponential. Using the corresponding values of w and q for the $GI/M/s+M$ model, given respectively by (3) and (7), we obtain the

ratio $\beta = \ln(\rho)/(\rho - 1)$. From (7), we get $\rho = 1 + \alpha q/s\mu$, yielding

$$\beta = \frac{\ln(1 + \alpha q/s\mu)}{\alpha q/s\mu}.$$

The corresponding delay estimate, as a function of n, is given by

$$\theta_{\mathrm{QL}_r^m}(n) = \beta \times \theta_{\mathrm{QL}}(n) = \frac{\ln(1 + \alpha n/s\mu)}{\alpha n/s\mu} \times \frac{n+1}{s\mu}.$$

Thus, the implementation of QL_r^m requires knowledge of n, s, μ, and α, but not of ρ or, equivalently, λ. It approximates the abandonment-time distribution by the exponential distribution. We will see that QL_r^m performs nearly the same as QL_m, which is good when the abandonment is nearly exponential, but not necessarily otherwise.

3.5. Approximation-Based Queue-Length Delay Estimator (QL_{ap})

Our most promising estimator QL_{ap} draws on the approximations in Whitt (2005): it approximates the $GI/GI/s + GI$ model by the corresponding $GI/M/s + M(n)$ model, with state-dependent Markovian abandonment rates.

We begin by describing the Markovian approximation for abandonments, as in §3 of Whitt (2005). As an approximation, we assume that a customer who is jth from the end of the queue has an exponential abandonment time with rate α_j, where α_j is given by

$$\alpha_j \equiv h(j/\lambda), \quad 1 \le j \le k; \tag{13}$$

k is the current queue length, and h is the abandonment-time hazard-rate function, defined as $h(t) \equiv f(t)/F^c(t)$, $t \ge 0$, where f is the corresponding density function (assumed to exist). Having α_j depend on h instead of F is convenient, because it is natural to estimate F via h; e.g., see Brown et al. (2005). From (13), we see that the estimator QL_{ap} depends on the abandonment distribution having a relatively smooth density. We assume that is the case.

We now explain the derivation of (13). If we knew that a given customer had been waiting for time t, then the rate of abandonment for that customer, at that time, would be $h(t)$. The goal is to produce, as an approximation, abandonment rates that depend on a customer's position in queue, and on the length of that queue. We therefore need to estimate the elapsed waiting time of that customer, given the available state information. To that end, assume that the queue length at an arbitrary time is k, and consider the customer, C_j, who is jth from the end of the line, $1 \le j \le k$. If there were no abandonments, then there would have been exactly $j-1$ arrival events since C_j arrived. Assuming that abandonments are relatively rare compared to service completions, a reasonable estimate is that there have been j arrival events since C_j arrived. Because a simple rough estimate for the time between successive arrival events is the reciprocal of the arrival rate, $1/\lambda$, the elapsed waiting time of C_j is approximated by j/λ and his abandonment rate by (13). The associated total abandonment rate from the queue in that system state is $\delta_k = \sum_{j=1}^{k} \alpha_j = \sum_{j=1}^{k} h(j/\lambda)$, $k \ge 1$, and $\delta_0 \equiv 0$.

For the $GI/M/s + M(n)$ model, we need to make further approximations to describe the potential waiting time of a customer who finds n other customers waiting in line, upon arrival. We have the approximate representation:

$$W_Q(n) \approx \sum_{i=0}^{n} X_i, \tag{14}$$

where X_{n-i} is the time between the ith and $(i+1)$st departure events. There is no difficulty for the first departure: X_n is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and n exponential random variables with rates α_j, $1 \le j \le n$, (corresponding to the abandonment times of the remaining customers waiting in line). That is, X_n has an exponential distribution with rate $s\mu + \sum_{j=1}^{n} \alpha_j = s\mu + \delta_n$.

The distribution of the remaining X_is is more complicated. Because individual customers have different abandonment rates that, in our framework, depend on how long these customers have been waiting in line, we need to consider the dynamics of the system over time to determine, after each departure, who are the remaining customers and what are their individual abandonment rates (to compute the resulting total abandonment rate). To simplify matters, we propose a further approximation, which is a slight modification of the argument in §7 of Whitt (2005).

The following describes our process. As a further approximation, we assume that successive departure events are either service completions, or abandonments from the head of the line. We also assume that an estimate of the time between successive departures is $1/\lambda$. As a result of these extra assumptions, we approximate the X_is in (14) by exponential random variables. Let X_{n-l}, which is the time between the lth and $(l+1)$st departure events, have an exponential distribution with rate $s\mu + \delta_n - \delta_l$. This is appropriate because it is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and $n-l$ exponential random variables with rates α_i, $l+1 \le i \le n$ (corresponding to the abandonment times of the customers waiting in line).

The QL_{ap} delay estimator given to a customer who finds n customers in queue upon arrival is

$$\theta_{\mathrm{QL}_{ap}}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}}. \tag{15}$$

Because QL_{ap} coincides with QL_m in the $GI/GI/s+M$ model, it is the optimal delay estimator in the $GI/M/s+M$ model, under the MSE criterion. But, in contrast to QL_m, this new queue-length-based estimator also performs remarkably well in the general $GI/GI/s+GI$ model. The simulation experiments of §6 suggest that QL_{ap} is uniformly superior to all other delay estimators, in all models considered.

We emphasize that all queue-length-based estimators apply equally well to steady-state and transient settings. They differ in the amount of information that their implementation requires. It is significant that QL, QL_m, and QL_r^m are all independent of the arrival process. For these three estimators, the arrival process can be arbitrary, even nonstationary. The QL_r and QL_{ap} estimators require knowledge of the arrival rate λ, which requires some degree of stationarity. (There should not be too much variation over time.)

4. Candidate Delay-History-Based Delay Estimators

In this section, we briefly describe alternative delay estimators based on recent customer delay history in the system. For a more detailed description, including performance approximations and refinements, see Ibrahim and Whitt (2009). We emphasize that delay-history-based estimators apply directly to more complex settings, such as models including customer response to delay announcements.

4.1. Last-To-Enter-Service (LES) Delay Estimator

As in Armony et al. (2009), a candidate delay estimator based on recent customer delay history is the delay of the last customer to have entered service, prior to our customer's arrival. That is, letting w be the delay of the last customer to have entered service, the corresponding LES delay estimate is $\theta_{\mathrm{LES}}(w) \equiv w$. As discussed in Whitt (1999), the possibility of making reliable delay estimations is enhanced by exploiting information about the current state of the system. Thus, we anticipate that queue-length-based estimators should be more effective than LES. Nevertheless, simulation experiments in §6 show that LES is relatively accurate in all models considered.

4.2. Other Delay-History-Based Delay Estimators

We can consider alternative delay-history-based estimators, in addition to LES. Closely related is the elapsed waiting time of the customer at the head of the line (HOL), assuming that there is at least one customer waiting at the new arrival epoch.

Another alternative delay estimator is the delay of the last customer to have completed service (LCS). We naturally would want to consider this alternative estimator if we only learn customer delay experience after service is completed. That might be the case for customers and outside observers. Under some circumstances, the LCS and LES estimators will be similar, but they typically are very different when s is large, because the last customer to complete service may have experienced his waiting time much before the last customer to enter service, since customers need not depart in order of arrival.

Thus, we are led to propose other candidate delay estimators based on the delay experience of customers that have already completed service. RCS is the delay experienced by the customer that arrived most recently (and thus entered service most recently) among those customers who have already completed service. We found that RCS is far superior to LCS when s is large.

Through analysis and extensive simulation experiments, we conclude that the LES and HOL estimators are very similar, with both being slightly more accurate than RCS and much more accurate than LCS. Here, we only discuss LES.

5. Heavy-Traffic Limits for Several Estimators in $G/M/s+M$

Because we are considering overloaded systems with $\rho > 1$, it is natural to develop analytical approximations for the mean-squared errors of our estimators by considering stochastic-process limits in the ED many-server heavy-traffic limiting regime, as specified in §2. As before, we add a subscript s to indicate the dependence upon s and then let $s \to \infty$.

In this section, we establish several limits for the $G/M/s+M$ model in the ED regime. Throughout this section we assume that the arrival process satisfies a functional central limit theorem (FCLT). Let $A_s(t)$ count the number of arrivals in the interval $[0, t]$ in model s. We assume that $A_s(t) \equiv A(st)$ for some given arrival process A with arrival rate λ. Let $\bar{A}_s(t) = A_s(t)/s \equiv A(st)/s$ for $t \geq 0$. Let $D \equiv D([0,\infty), \mathbb{R})$ be the function space of all right continuous real-valued functions with left limits, endowed with the usual Skorohod (J_1) topology; e.g., see Whitt (2002). We assume that A satisfies a functional weak law of large numbers and an FCLT refinement:

$$\begin{aligned} &\bar{A}_s(t) \;\Rightarrow\; \lambda t \quad \text{in } D \quad \text{and} \\ &\sqrt{s}(\bar{A}_s(t) - \lambda t) \;\Rightarrow\; \sqrt{\lambda c_a^2} B(t) \quad \text{in } D \text{ as } s \to \infty, \end{aligned} \tag{16}$$

where B is a standard Brownian motion. That condition will be satisfied if A is a renewal process with

an interarrival-time distribution having finite first and second moments. As usual, the arrival process affects the limits for the other random quantities (the estimators) only via the two normalization constants λ and c_a^2. When A is a renewal counting process, c_a^2 is the SCV of an interarrival time.

We start by considering the Markovian estimator QL_m, which is the best possible estimator for the $G/M/s+M$ model, under the MSE criterion. It does not depend on the arrival process. Recall that the waiting time for an arrival that finds n customers in queue upon arrival is given by (9). We will apply the following lemma, which is Lemma 6.1 of Talreja and Whitt (2009).

Lemma 1. *For the $G/M/s+M$ model in the ED many-server heavy-traffic regime,*

$$E[W_{Q,s}(\lfloor st\rfloor)] \to c(t), \quad s\mathrm{Var}(W_{Q,s}(\lfloor st\rfloor) \to d(t) \tag{17}$$

and

$$\widehat{W}_{Q,s}(t) \equiv \sqrt{s}(W_{Q,s}(\lfloor st\rfloor) - c(t)) \;\Rightarrow\; B(d(t)) \quad \text{in } D \text{ as } s\to\infty, \tag{18}$$

where B is a standard Brownian motion, and c and d are the deterministic real-valued functions:

$$c(t) \equiv \frac{1}{\alpha}\ln\left(1+\frac{\alpha t}{\mu}\right) \quad \text{and} \quad d(t) \equiv \frac{t}{\mu(\mu+\alpha t)}. \tag{19}$$

As a consequence of the stochastic-process limit in (18), we obtain the one-dimensional limit

$$\sqrt{s}(W_{Q,s}(\lfloor st\rfloor) - c(t)) \;\Rightarrow\; N(0,d(t)) \quad \text{in } \mathbb{R} \text{ as } s\to\infty \text{ for each } t. \tag{20}$$

As a further consequence, we obtain the following result for the best-possible estimators $\theta_{\mathrm{QL}_{m,s}}(n)$. We use a random time change by the fluid limit

$$\overline{Q}_s(\infty) \equiv \frac{Q_s(\infty)}{s} \;\Rightarrow\; q \equiv \frac{\lambda-\mu}{\alpha} \quad \text{as } s\to\infty, \tag{21}$$

from Theorem 2.3 of Whitt (2004) or Theorem 6.1 of Talreja and Whitt (2009).

Theorem 1. *For the $G/M/s+M$ model in the ED many-server heavy-traffic regime,*

$$s\mathrm{MSE}(\theta_{\mathrm{QL}_{m,s}}(\lfloor st\rfloor)) \equiv s\mathrm{Var}(W_{Q,s}(\lfloor st\rfloor)) \to d(t) \quad \text{as } s\to\infty \tag{22}$$

for each $t>0$, where $d(t)$ is given in (19) *and*

$$s\mathrm{MSE}(\theta_{\mathrm{QL}_{m,s}}(Q_s(\infty))) \equiv s\mathrm{Var}(W_{Q,s}(Q_s(\infty))) \;\Rightarrow\; d(q) \equiv \frac{q}{\lambda\mu} \equiv \frac{\lambda-\mu}{\lambda\mu\alpha} \quad \text{as } s\to\infty. \tag{23}$$

As a consequence (after establishing appropriate uniform integrability to get convergence of moments from convergence in distribution, which is not difficult at this point), we get associated convergence of moments from the convergence in distribution in (23), i.e.,

$$sE[\mathrm{MSE}(\theta_{\mathrm{QL}_{m,s}}(Q_s(\infty)))] \to d(q) \quad \text{as } s\to\infty. \tag{24}$$

From either (23) or (24), we get the approximation

$$E[\mathrm{MSE}(\theta_{\mathrm{QL}_{m,s}}(Q_s(\infty)))] \approx \frac{\lambda-\mu}{s\lambda\mu\alpha}. \tag{25}$$

Note that the FCLT normalization constant c_a^2 does not appear in (23)–(25). Other estimators that do not exploit knowledge of the queue length will fare worse, largely according to c_a^2. First, we can apply an extension of Theorem 6.4 of Talreja and Whitt (2009) to describe the asymptotic behavior of the no-information estimator $W_s(\infty)$. We extend the result for the $M/M/s+M$ model to the $G/M/s+M$ model, which is not difficult, reasoning as in §7.3 of Pang et al. (2007). First, we can extend Theorem 6.1 of Talreja and Whitt (2009) in that way to get an ED stochastic-process limit for the queue-length process in the $G/M/s+M$ model, getting an Ornstein-Uhlenbeck diffusion-process limit with infinitesimal mean $\mu(x)=-\alpha x$ and an infinitesimal variance $\sigma^2(x)=\lambda(c_a^2+1)$, which in turn leads to a limit for the steady-state queue lengths. We then apply that result to get a generalization of the limit for the steady-state waiting time in Theorem 6.4 of Talreja and Whitt (2009).

Theorem 2. *For the $G/M/s+M$ model in the ED many-server heavy-traffic regime,*

$$\widehat{Q}_s(\infty) \equiv \sqrt{s}(\overline{Q}_s(\infty)-q) \;\Rightarrow\; N\left(0,\frac{\lambda(c_a^2+1)}{2\alpha}\right) \quad \text{as } s\to\infty \tag{26}$$

and

$$\widehat{W}_s(\infty) \equiv \sqrt{s}(W_s(\infty)-w) \;\Rightarrow\; N(0,\sigma_w^2) \quad \text{as } s\to\infty, \tag{27}$$

where $\sigma_w^2 \equiv 1/\alpha\mu + (c_a^2-1)/2\lambda\alpha$, with w in (3) *and q in* (21).

Note that the variance terms in Theorem 2 simplify when $c_a^2=1$. We immediately obtain the limit for the MSE of the NI estimator, assuming appropriate uniform integrability. The NI estimator can be either the mean steady-state waiting time $E[W_s(\infty)]$ or the fluid limit w, because of the fluid limit in (3).

Corollary 1. *In the setting of Theorem* 2, *assuming necessary uniform integrability,*

$$s\mathrm{MSE}(\theta_{\mathrm{NI},s}) \equiv s\mathrm{Var}(W_s(\infty)) \to \frac{1}{\alpha\mu} + \frac{c_a^2-1}{2\lambda\alpha} \quad \text{as } s\to\infty. \tag{28}$$

Combining the limits in (23) and (28), we obtain the following:

COROLLARY 2. *In the setting of Theorem* 2, *assuming necessary uniform integrability,*

$$\frac{\mathrm{MSE}(\theta_{\mathrm{NI},\,s})}{E[\mathrm{MSE}(\theta_{\mathrm{QL}_{m,\,s}}(Q_s(\infty)))]} \to \frac{2\lambda+\mu(c_a^2-1)}{2(\lambda-\mu)} > 1 \quad \text{as } s\to\infty. \tag{29}$$

We now establish corresponding results for the delay-history-based estimator LES. We exploit the fact that we can represent $W_{\mathrm{LES}}(w)$ in terms of the random variable $W_{\mathrm{QL}_{m,\,s}}(n)$ in (9) and a net-input process $N_s \equiv \{N_s(t)\colon t\ge 0\}$ over the interval $[0, w]$, i.e.,

$$W_{\mathrm{LES},\,s}(w) \approx W_{Q,\,s}(N_s(w)) \equiv \sum_{i=0}^{N_s(w)} X_{s,\,i}, \tag{30}$$

where $N_s(w)$ counts the number of arrivals in the interval $[0, w]$ who do not abandon, in system s. Formula (30) is not an exact relation because it does not account for the state change since the last customer entered service, but that change is clearly asymptotically negligible in the ED many-server limiting regime.

It is significant that the net-input stochastic process N_s has the structure of the number in system in a $G/M/\infty$ infinite-server system, starting out empty, with arrival rate $\lambda_s \equiv \lambda s$ and individual service rate equal to our abandonment rate α. The Markovian $M/M/\infty$ special case is very well studied; e.g., see Eick et al. (1993). In particular, it is well known that $N_s(t)$ has a Poisson distribution for each s and t with

$$E[N_s(t)] = \frac{s\lambda}{\alpha}(1-e^{-\alpha t}), \quad t\ge 0. \tag{31}$$

The heavy-traffic limit for more general infinite-server models, starting out empty, was established by Borovkov (1967), as reviewed on p. 176 of Whitt (1982).

THEOREM 3 (BOROVKOV 1967). *For the $G/M/\infty$ models under consideration, with arrival rate $\lambda_s = \lambda s$ and service rate α,*

$$\bar{N}_s(t) \equiv \frac{N_s(t)}{s} \;\Rightarrow\; a(t) \equiv \frac{\lambda}{\alpha}(1-e^{-\alpha t}) \quad \text{in } D \text{ as } s\to\infty \tag{32}$$

and

$$\widehat{N}_s(t) \equiv \sqrt{s}(\bar{N}_s(t) - a(t)) \;\Rightarrow\; \widehat{G}(t) \quad \text{in } D \text{ as } s\to\infty, \tag{33}$$

where $\widehat{G} \equiv \{\widehat{G}(t)\colon t\ge 0\}$ is a Gaussian stochastic process with

$$\widehat{G}(t) \overset{\mathrm{d}}{=} N(0,\sigma_n^2(t)), \quad \text{where} \quad \sigma_n^2(t) \equiv a(t) + \frac{\lambda(c_a^2-1)}{2\alpha}(1-e^{-2\alpha t}), \tag{34}$$

for $a(t)$ defined in (32) *and c_a^2 in* (16).

We apply Theorem 3 to establish the following results for LES. To go beyond the $M/M/s+M$ model to treat the more general $G/M/s+M$ model, we add an extra assumption here. We assume that the limits for $\widehat{N}_s$ in (33) and for $\widehat{W}_s(\infty)$ in (27) hold jointly with independent limits. That holds automatically if the arrival process has independent increments (which is covered by the M case), because the evolution of N_s occurs after the arrival of the customer with the observed LES waiting time $W_s(\infty)$. For renewal processes, that joint convergence with independent limits should also hold because the interarrival times are i.i.d. and the arrivals are very fast. We add this condition to the general FCLT assumed in (16). We prove the following result in the e-companion.

THEOREM 4. *For the $G/M/s+M$ model in the ED many-server limiting regime* (*assuming the extra assumption immediately above and the necessary uniform integrability for the moment convergence*), *as $s\to\infty$,*

$$\theta_{\mathrm{LES},\,s}(W_s(\infty)) \equiv W_s(\infty) \;\Rightarrow\; w \equiv \frac{1}{\alpha}\left(\ln\left(\frac{\lambda}{\mu}\right)\right),$$

$$\widehat{W}_{\mathrm{LES},\,s}(W_s(\infty)) \equiv \sqrt{s}(W_{\mathrm{LES},\,s}(W_s(\infty)) - W_s(\infty)) \;\Rightarrow\; N(0,\sigma_{\mathrm{LES}}^2), \tag{35}$$

$$sE[\mathrm{MSE}(\theta_{\mathrm{LES},\,s}(W_s(\infty)))] \to \sigma_{\mathrm{LES}}^2, \tag{36}$$

where

$$\sigma_{\mathrm{LES}}^2 \equiv d(a(w)) + \frac{\sigma_n^2(w)}{\lambda^2} + \left(\frac{\lambda-\mu}{\lambda}\right)^2\sigma_w^2 = 2d(q) + \frac{(c_a^2-1)(\lambda-\mu)}{\alpha\lambda^2}, \tag{37}$$

for σ_w^2 in Theorem 2, *$\sigma_n^2(t)$ in* (34), *$a(w) = q$ and $d(q) = q/\lambda\mu$.*

COROLLARY 3. *Consider the setting of Theorem* 4. *For the $M/M/s+M$ model,*

$$\frac{E[\mathrm{MSE}(\theta_{\mathrm{LES},\,s}(W_s(\infty)))]}{E[\mathrm{MSE}(\theta_{\mathrm{QL}_{m,\,s}}(Q_s(\infty)))]} \to 2 \quad \text{as } s\to\infty. \tag{38}$$

For the $D/M/s+M$ model,

$$\frac{E[\mathrm{MSE}(\theta_{\mathrm{LES},\,s}(W_s(\infty)))]}{E[\mathrm{MSE}(\theta_{\mathrm{QL}_{m,\,s}}(Q_s(\infty)))]} \to (2-\rho^{-1}) > 1 \quad \text{as } s\to\infty. \tag{39}$$

For the more general $G/M/s+M$ model,

$$\frac{E[\mathrm{MSE}(\theta_{\mathrm{LES},\,s}(W_s(\infty)))]}{E[\mathrm{MSE}(\theta_{\mathrm{QL}_{m,\,s}}(Q_s(\infty)))]} \to r(\mathrm{LES},\mathrm{QL}_m) \quad \text{as } s\to\infty, \tag{40}$$

where

$$r(\mathrm{LES},\mathrm{QL}_m) = 2 \;(\ge 2 \text{ or } \le 2) \quad \text{if and only if} \quad c_a^2 = 1 (\ge 1 \text{ or } \le 1).$$

From (39), we see that QL_m is only slightly better than LES in the $D/M/s+M$ model when $\rho \equiv \lambda/\mu$ is only slightly greater than 1. Combining the MSE ratio limits in Theorems 1 and 4, we obtain the following:

COROLLARY 4. *For the $M/M/s+M$ model in the ED many-server limiting regime,*

$$\frac{E[\mathrm{MSE}(\theta_{\mathrm{LES},s}(W_s(\infty)))]}{\mathrm{MSE}(\theta_{\mathrm{NI},s})} \to \frac{2(\rho-1)}{\rho}, \qquad (41)$$

so that LES is asymptotically more (*less*) *efficient than NI if $\rho < 2$ ($\rho > 2$).*

We conclude this section by stating a CLT for the steady-state waiting time, and thus the NI delay estimator, in the $M/M/s+GI$ model in the ED regime, which is Theorem 6.1 (e) of Zeltyn and Mandelbaum (2005).

THEOREM 5 (ZELTYN AND MANDELBAUM 2005). *For the $M/M/s+GI$ model in the ED regime, where the abandonment-time cdf F has density f, $W_s(\infty) \Rightarrow w$ for w in* (11) *and*

$$\sqrt{s}(W_s(\infty)-w) \;\Rightarrow\; N(0, 1/\lambda f(w)) \quad \text{as } s \to \infty. \qquad (42)$$

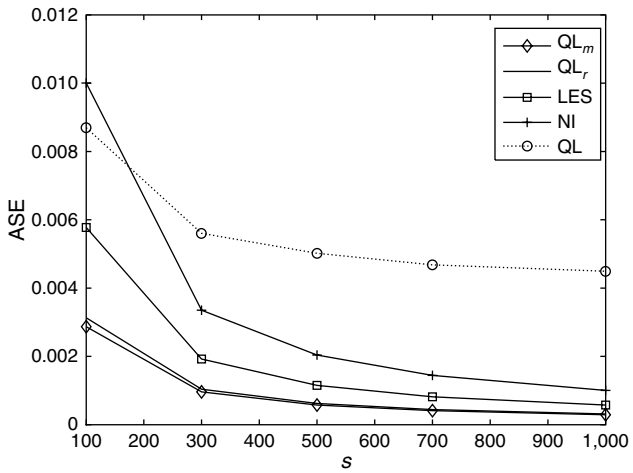
6. Simulation Results for the $M/M/s+GI$ Model

In this section, we present simulation results quantifying the performance of the alternative queue-length-based delay estimators of §3, and of the LES delay estimator, with exponential and nonexponential abandonment-time distributions; i.e., we consider the $M/M/s+GI$ model. For the abandonment-time distribution, we consider M (exponential) and E_{10} (Erlang, sum of 10 exponentials) distributions. In the e-companion, we also consider the H_2 distribution (hyperexponential with SCV equal to 4 and balanced means). We use a Poisson arrival process because this model is commonly used in practice. We briefly discuss other models in §6.5.

6.1. Description of the Experiments

We vary the number of servers, s, but consider only relatively large values ($s \geq 100$), because we are interested in large service systems. We let the service rate, μ, be equal to 1. We do this without loss of generality, because we are free to choose the time units in our system, and this assumption amounts to measuring time in units of mean service time. We also let the abandonment rate, α, be equal to 1 because that seems to be a representative value. We also consider $\alpha = 0.2$ and $\alpha = 5.0$ in the e-companion. We vary λ to get a fixed value of ρ, for alternative values of s. We let $\rho = 1.4$ in all models. This value is chosen to let our systems be significantly overloaded. Because of abandonment, the congestion is not extraordinarily high. For example, with $s = 100$ servers and exponential abandonments, the mean queue length is about $q \approx (\rho-1)s/\alpha \approx 40$, whereas the average potential waiting time is about $w \approx q/s\mu \approx 0.4/\mu$ (less than half a mean service time).

Figure 1 **ASE in the $M/M/s+M$ Model with $\rho = 1.4$**



Our simulations are steady-state simulations. The simulation results are based on 10 independent replications of five million events each, where an event is either a service completion, an arrival event, or an abandonment from the system. In this section, we show plots of simulation estimates. Corresponding tables with 95% confidence intervals appear in the e-companion.

6.2. Results for the $M/M/s+M$ Model

In this model, QL_{ap} coincides with QL_m. Therefore, we do not include separate results for QL_{ap}. Consistent with the theory in §3, Figure 1 shows that QL_m is the best possible, under the MSE criterion. The RRASE for QL_m ranges from about 14% for $s = 100$ to about 4% when $s = 1{,}000$. We see that the accuracy of this estimator improves as the number of servers increases. Note that all estimators are relatively accurate for this model, with the exception of QL. For example, the RRASE of LES ranges from about 22% for $s = 100$ to about 7% for $s = 1{,}000$. Figure 2 shows that $s \times \mathrm{ASE}(QL_m)$, the ASE of QL_m multiplied by the number of servers s, is nearly constant for all values of s considered. In particular, Figure 2 shows that $s \times \mathrm{ASE}(QL_m) \approx (\lambda-\mu)/(\lambda\mu\alpha)$, as in (25) of §5. The relative error between the simulation estimates for $\mathrm{ASE}(QL_m)$ and the numerical value given by (25) is less than 1% throughout.

The QL_r^m estimator is nearly identical to QL_m. This can be easily explained: When the number seen in queue upon arrival, n, is large, $\theta_{QL_m}(n)$ can be approximated by an integral (limit of the Riemann sum)

$$\theta_{QL_m}(n) \approx \int_0^n \frac{1}{s\mu+\alpha x}\,dx = \ln(s\mu+\alpha n) - \ln(s\mu)$$
$$= \frac{1}{\alpha}\ln(1+\alpha n/s\mu).$$

Figure 2 $s \times$ **ASE in the** $M/M/s+M$ **Model with** $\rho = 1.4$

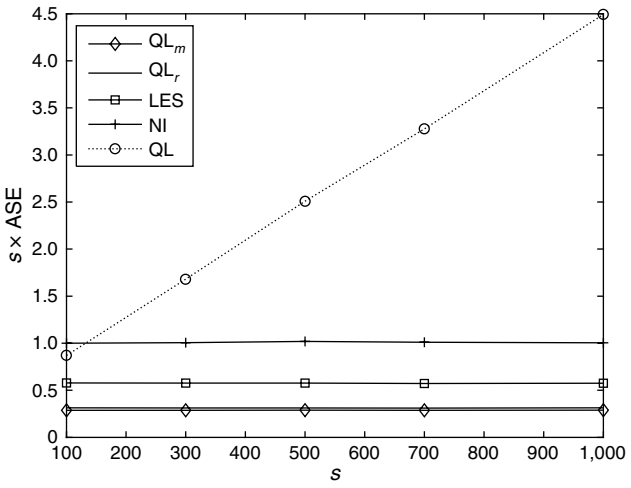
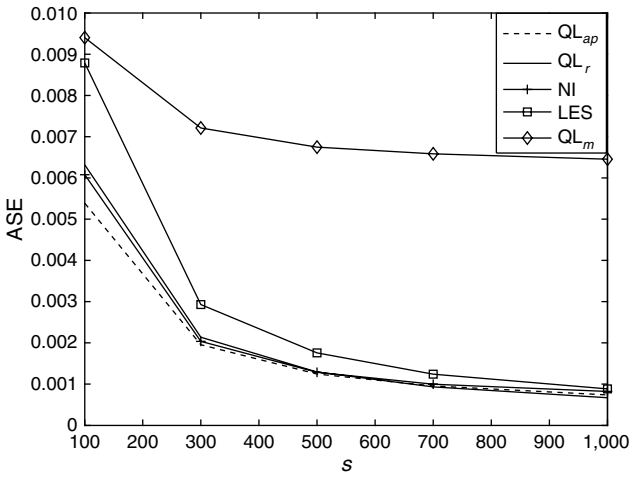


Figure 3 **ASE in the** $M/M/s+E_{10}$ **Model with** $\rho = 1.4$



On the other hand, we have that

$$\theta_{\mathrm{QL}_r^m}(n) \equiv \left[\ln\left(\frac{\alpha n}{s\mu}+1\right)\Big/\left(\frac{\alpha n}{s\mu}\right)\right]\times\frac{n+1}{s\mu}$$
$$\approx \frac{1}{\alpha}\ln(1+\alpha n/s\mu).$$

So that, for large n, the two estimators QL_m and QL_r^m should perform nearly the same.

The LES estimator performs worse than QL_m and QL_r. The ratio ASE(LES)/ASE(QL_m) is close to 2 for all values of s, which provides support to (38). This is consistent with the results in Ibrahim and Whitt (2009) for the $GI/M/s$ model, without customer abandonment. Figure 2 shows that $s \times$ ASE(LES) $\approx \sigma^2_{\mathrm{LES}}$, consistent with (36). Indeed, the relative error between the simulation estimates and the numerical value given by (37) is less than 1% throughout.

The NI estimator performs worse than LES: The ratio ASE(NI)/ASE(LES) is close to 1.75 throughout. The relative error between the simulation estimates for ASE(NI)/ASE(LES) and the numerical value given by (41) is less than 2% throughout. Figure 2 shows that $s \times$ ASE(NI) $\approx 1/\alpha\mu$, as in (28), with $c_a^2 = 1$. The relative error between the simulation estimates for ASE(NI) and the numerical value given by (28) is less than 2% throughout.

The QL estimator performs significantly worse than the other three estimators and its performance gets worse as s increases. The ratio ASE(QL)/ASE(QL_m) ranges from about 3 when $s = 100$ to about 16 when $s = 1{,}000$. Figure 2 shows that $s \times$ ASE(QL) is monotone increasing in s. This shows the need to go beyond QL when customer abandonment is included.

6.3. Results for the $M/M/s+E_{10}$ Model

Figure 3 shows that QL_{ap} is the best possible delay estimator for this model, except when s is very large (e.g., $s = 700$ or $s = 1{,}000$). The corresponding RRASE ranges from about 10% when $s = 100$ to about 3% when $s = 1{,}000$. The QL_r estimator performs worse than QL_{ap} for smaller values of s, but slightly outperforms QL_{ap} for larger values of s. The ratio ASE(QL_r)/ASE(QL_{ap}) ranges from about 2 when $s = 100$ to about 0.9 when $s = 1{,}000$.

In contrast to previous cases, NI is the second or third most effective delay estimator here, depending on the number of servers. It performs nearly as well as QL_{ap}, particularly when s is large. This confirms that NI can be a competitive delay estimator, with customer abandonment. The NI estimator is especially appealing because it does not use any information beyond the model.

The LES estimator also fares well. The corresponding RRASE ranges from about 14% when $s = 100$ to about 3% when $s = 1{,}000$. Figure 4 shows that $s \times$ ASE(LES) equals a constant, for all values of s. It is significant that LES is the only estimator with this property here, unlike the previous two models.

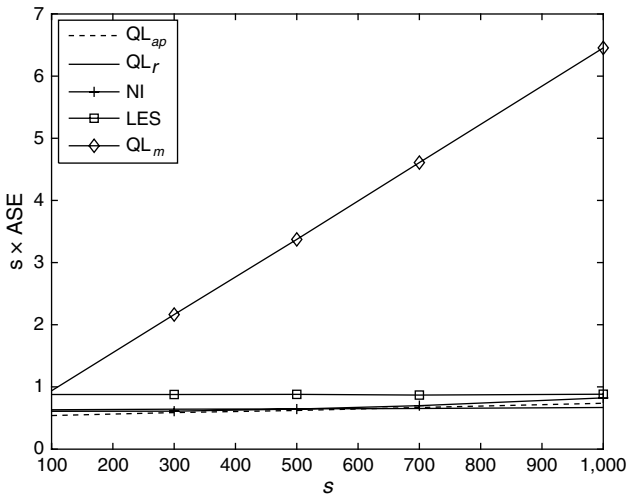
The QL_m estimator, which was nearly identical to QL_{ap} before, now performs worse: the corresponding RRASE ranges from about 14% when $s = 100$ to about 10% when $s = 1{,}000$.

QL_m is relatively effective when $s = 100$ but becomes significantly worse than QL_{ap} when $s = 1{,}000$ (in that case, the ratio of respective ASE's is close to 9). The QL estimator is consistently the least effective delay estimator in this model too: the ratio ASE(QL)/ASE(QL_{ap}) ranges from about 15 when $s = 100$ to about 95 when $s = 1{,}000$. That is why the corresponding ASE curve is not even included in Figures 3 and 4.

6.4. Results for Other Models

We consider more general interarrival-time and service-time distributions in the e-companion and

Figure 4 $s \times$ **ASE in the** $M/M/s + E_{10}$ **Model with** $\rho = 1.4$



the online supplement (http://www.columbia.edu/~ww2040/). For the interarrival-time distribution, we consider M, D, and H_2; i.e., we consider the $GI/M/s+M$ model. Our simulation results for this model substantiate the heavy-traffic limits of §5, which quantify the performances of some delay estimators in the $GI/M/s+M$ model; e.g., see Theorems 1, 2, and 4.

For the service-time distribution, we consider H_2, D, E_{10}, E_4, and E_2 (sum of 4, and 2 exponentials, respectively). We also consider $LN(1,1)$ (lognormal with mean and variance equal to 1), because there is empirical evidence suggesting a good fit of the service-time distribution to the lognormal distribution; see Brown et al. (2005). These additional simulation results are consistent with those aforementioned, with one notable exception. There is a significant increase in ASE for all estimators with deterministic (constant) service times, with performance tending to be independent of s. In fact, the NI estimator is best here. That indicates a need for new methods for this special case. However, even very low variability in the service times, e.g., the E_{10} distribution with SCV equal to 0.1, is enough for our delay estimators to be relatively accurate; see the e-companion.

We also consider different combinations of service-time and abandon-time distributions. We do not consider D abandonment times because our QL$_{ap}$ estimator requires a density; see (13). Constant service times cause a problem in all cases, but otherwise the estimators perform well; e.g., there is no difficulty when both the service times and abandonments are E_{10}.

7. Conclusions

In this paper, we studied the performance of alternative real-time delay estimators in the overloaded $GI/GI/s+GI$ queueing model, allowing customer abandonment. We considered both queue-length-based and delay-history-based delay estimators. Queue-length-based estimators exploit system-state information, including the queue length seen upon arrival. In contrast, delay-history-based estimators have the advantage of not relying on any model or system-state information: their implementation only exploits customer delay history in the system. We also considered the NI delay estimator, exploiting no information beyond the model. We established heavy-traffic limits for the expected MSE's of QL$_m$, LES, and NI in the $G/M/s+M$ model, in the ED many-server heavy-traffic limiting regime. For nonexponential service-time and abandonment-time distributions, we used computer simulation to study the performance of the candidate delay estimators.

7.1. Managerial Insights

Our starting point is the notion that it is often desirable to make delay announcements to arriving customers in service systems. We consider how reliable real-time delay estimates can be made.

As a frame of reference, we considered the classical delay estimator based on the queue length, QL, which multiplies the queue length plus one times the mean interval between successive service completions. The QL estimator is straightforward, and is commonly used in practice. We observed that it is the most accurate delay estimator, under the MSE criterion, in the $GI/M/s$ queueing model, without customer abandonment. Whenever the actual service system is well modeled by a $GI/M/s$ queueing model and the system state is known accurately at each time, there is little motivation for considering other delay estimators besides the standard QL estimator.

Intuitively, we should expect that, when there is significant customer abandonment while waiting in line, the QL estimator will overestimate the delay because many customers in queue may abandon before entering service, and QL fails to take that into account. Consistent with intuition, we showed in §6 that QL makes consistent estimation error when there is significant customer abandonment; e.g., see Figures 1 and 2. Motivated by the simple form of the QL delay estimate, $\theta_{\mathrm{QL}}(n)$ in (6), we proposed modified queue-length-based delay estimators that account for customer abandonment and are also easy to implement in practice.

The Markovian queue-length-based estimator, QL$_m$, is a variant of QL that accounts for customer abandonment by assuming that waiting customers have i.i.d. exponential abandonment times with rate α. It also assumes that service times are i.i.d. with an exponential distribution. In §3, we showed that QL$_m$ is the most accurate estimator, under the MSE criterion, in the $GI/M/s+M$ model. We established

heavy-traffic limits that generated an approximation for the expected MSE of QL_m in the $GI/M/s+M$ model in §5.

In practice, the QL_m estimator is effective whenever the abandonment-time and service-time distributions in the actual service system are well modeled by an exponential distribution. The abandonment rate α, which is required for the implementation of QL_m, can be estimated from system data as the ratio of the proportion of abandoning customers to the average waiting time in the queue; see §5 of Garnett et al. (2002). The average waiting time in the queue and the proportion of abandoning customers are fairly standard system data outputs. In the context of call centers, for example, they can be easily obtained from the automatic call distributor's data.

However, QL_m is not always so good for the more general $GI/GI/s+GI$ model. In Figures 3 and 4, we showed that it can be inferior to all other estimators (except QL) with a nonexponential abandonment-time distribution. Because nonexponential service-time and abandonment-time distributions are commonly observed in practice, it is important to propose other queue-length-based delay estimators that effectively cope with nonexponential distributions. Approximations are needed because direct mathematical analysis is difficult.

We proposed the simple-refined QL_r estimator, which multiplies the QL estimate by a model-dependent constant, based on fluid approximations in the ED heavy-traffic limiting regime. Simulation results in §6 and the e-companion show that QL_r performs remarkably well. The QL_r estimator is competitive whenever the actual service system is large and overloaded, i.e., whenever the fluid approximations are appropriate. The QL_r estimator performs significantly better than QL_m (and QL) when the abandonment-time distribution is not nearly exponential; e.g., see Tables EC.2 and EC.3 in the e-companion.

Our most promising delay estimator is the new approximation-based estimator, QL_{ap}. Simulation results in §6 and the e-companion show that QL_{ap} is consistently the most effective estimator (with the exception of D service; see the e-companion). It is a variant of QL_m that assumes abandonment times are independent, exponential, with state-dependent abandonment rates. The QL_{ap} estimator coincides with QL_m in the setting of the $GI/M/s+M$ model, and is thus the most accurate for that model, under the MSE criterion. It also performs remarkably well for nonexponential abandonment-time distributions.

The QL_{ap} delay estimate, $\theta_{QL_{ap}}(n)$ in (15), requires knowledge of the abandonment-time hazard-rate function, h. That is convenient from a practical point of view, because it is relatively easy to estimate hazard rates from system data; see Brown et al. (2005).

It is significant that QL_r and QL_{ap} require knowledge of the arrival rate, λ, which requires some degree of stationarity. These estimators should be effective whenever the arrival rate in the actual service system does not vary too rapidly.

Unlike without abandonments, the NI estimator, announcing the deterministic heavy-traffic fluid limit w of the waiting time, is an effective estimator in the overloaded $GI/GI/s+GI$ model. It is best possible for D service, but not otherwise. Nevertheless, it is remarkably effective, especially when the abandonment-time distribution has low variability. The NI estimator is a competitive estimator whenever the actual service system is large and overloaded, and the service and abandonment times are not highly variable.

Finally, we considered the LES estimator, which is appealing because it only depends on the history of delays in the system. Intuitively, we should expect that LES will perform worse than queue-length-based estimators when the queue length and model parameters are known, because it does not exploit information about system state. Simulation shows that this is usually true. Nevertheless, the LES estimator is quite effective in all models considered. In §5, we showed that the expected MSE of LES in the $GI/M/s+M$ model increases with the squared coefficient of variation of the interarrival times, c_a^2. The practical significance of this result is that reliability of LES increases as the variability in the arrival process decreases.

In practice, LES has the advantage of robustness: it responds automatically to changes in system parameters, because it does not depend on those parameters. That is important because real-life systems are often quite complicated. For one example, there may be multiple customer classes and multiple service pools with some form of skill-based routing; see Gans et al. (2003). For a second example, the number of servers and mean service times may be time varying, in part because the servers are humans who serve in different shifts and may well have different service-time distributions. Delay-history-based estimators may be preferred in such scenarios.

7.2. Future Research Directions

In ongoing work, we have begun studying the delay estimation problem with both customer abandonment and time-varying arrival rates. A natural model for capturing time-dependent arrivals is the nonhomogeneous Poisson process. Experience indicates that it is appropriate for most real service systems. Such a process is completely characterized by its arrival-rate function. We are developing delay estimators that effectively cope with time-varying arrivals, and studying the performance of these estimators by using computer simulation.

As discussed in §6.4, it remains to carefully examine the case of deterministic service times. As discussed in §1.3, it remains to carefully consider the effect of customer response to delay announcements.

8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

The reported research was supported by NSF Grant DMI-0457095.

References

Aksin, O. Z., M. Armony, V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Allon, G., A. Bassamboo, I. Gurvich. 2007. We will be right with you: Managing customers with vague promises. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* **57**(1) 66–81.

Baccelli, Boyer, Hebuterne. 1984. Single-server queues with impatient customers. *Adv. Appl. Probab.* **16** 887–905.

Borovkov, A. A. 1967. On limit laws for service processes in multi-channel systems. *Siberian Math.* **8** 746–763.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Eick, S. G., W. A. Massey, W. Whitt. 1993. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41**(4) 731–742.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Sci.* **53**(6) 962–970.

Ibrahim, R., W. Whitt. 2009. Real-time delay estimation based on delay history. *Manufacturing Service Oper. Management* **11**(3) 397–415.

Jouini, O., Y. Dallery, Z. Aksin. 2007. Modeling call centers with delay information. Working paper, Koç University, Sariyer-Istanbul, Turkey.

Mandelbaum, A., S. Zeltyn. 2004. The impact of customers patience on delay and abandonment: Some empirically-driven experiments with the M/M/N+G Queue. *OR Spectrum* **26**(3) 377–411.

Mandelbaum, A., S. Zeltyn. 2007. Service engineering in action: The palm/erlang-a queue, with applications to call centers. D. Spath, K.-P. Fähnrich, eds. *Advances in Services Innovations*. Springer-Verlag, 17–48.

Pang, G., R. Talreja, W. Whitt. 2007. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* **4** 193–267.

Talreja, R., W. Whitt. 2009. Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Probab.* Forthcoming.

Whitt, W. 1982. On the heavy-traffic limit theorem for GI/G/infinity queues. *Adv. Appl. Probab.* **14** 171–190.

Whitt, W. 1999. Predicting queueing delays. *Management Sci.* **45**(6) 870–888.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer-Verlag, New York.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Sci.* **51**(2) 221–235.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems* **51**(3–4) 361–402.