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, NY 10027-6699
{rei2101, 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 non-exponential abandonment-time distribution to obtain new delay estimators that effectively cope with non-exponential abandonment-time distributions.

Key words: delay estimation; delay announcements; call centers; many-server queues; customer abandonment; simulation; heavy traffic.

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 so 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). Since 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 non-exponential abandonment-time distribution, from Whitt (2005, 2006).

1.1. The 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 \( \lambda^{-1} \) 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 \( \mu^{-1} \) and a general distribution. Each arriving customer will abandon if he is unable to start service before a random time with mean \( \alpha^{-1} \) 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 (FCFS) 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. (2008), 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 as above. We would then re-estimate model parameters which 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 *average squared error* (ASE), defined by:

\[ \text{ASE} = \frac{1}{k} \sum_{i=1}^{k} (p_i - e_i)^2, \]

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 \( \theta_{QL}(n) \) be some given single-number delay estimate which is based on the queue length, \( n \). Then, the MSE of the corresponding delay estimator is given by:

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

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:

\[
RRASE \equiv \sqrt{\frac{ASE}{\sum_{i=1}^{k} P_i}},
\]
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 (2007), 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. (2008). Whitt (1999) discusses the possibility of making reliable delay estimations by exploiting information about the current state of the system. Armony et al. (2008) 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 Section 2 of Jouini et al. (2007).

1.9. Organization of the paper

The rest of this paper is organized as follows: In §2, we describe a no-information delay estimator (NI) 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. We briefly describe alternative delay-history-based delay estimators in §4; a more complete description can be found in Ibrahim and Whitt (2007). We establish heavy-traffic limits for several delay estimators in the $G/M/s + M$ model in §5, and present simulation results for the $M/M/s + GI$ model in §6. We make concluding remarks and describe managerial insights in §7. We postpone one long proof of a result in §5 (Theorem 4) to the e-companion, where we also present additional supporting material. More appears in an online supplement, Ibrahim and Whitt (2008).

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 Talreja and Whitt (2008). That limit describes how the model behaves as the arrival rate $\lambda$ and number of servers $s$ increase, while the individual service rate $\mu$ and individual abandonment rate $\alpha$ 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 $G/GI/s + GI$ 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

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

while Theorem 6.1 of Zeltyn and Mandelbaum (2005) (Theorem 5 below) and Theorem 6.4 of Talreja and Whitt (2008) show that

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

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

Table 1  Summary of the information required for the implementation of each queue-length delay estimator.

<table>
<thead>
<tr>
<th>Information About the Model</th>
<th>QL</th>
<th>Q(t), s, µ</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL&lt;sub&gt;r&lt;/sub&gt;</td>
<td>Q(t), s, µ, α</td>
<td></td>
</tr>
<tr>
<td>QL&lt;sub&gt;r&lt;/sub&gt;</td>
<td>Q(t), s, µ, F(x), λ</td>
<td></td>
</tr>
<tr>
<td>QL&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Q(t), s, µ, α</td>
<td></td>
</tr>
<tr>
<td>QL&lt;sub&gt;ap&lt;/sub&gt;</td>
<td>Q(t), s, µ, F(x), λ</td>
<td></td>
</tr>
</tbody>
</table>

The deterministic fluid approximation $w$ in (3) and the steady-state mean $E[W_s(\infty)]$ it approximates, are candidate no-information (NI) estimators, $\theta_{NI}$, paralleling the NI estimator for the GI/M/s model considered as a reference point in Ibrahim and Whitt (2007). 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 above (plus appropriate uniform integrability, which can also be established), we have

$$MSE(\theta_{NI}) \approx Var(W_s(\infty)) \approx \frac{1}{s\alpha\mu} \to 0 \quad \text{as} \quad s \to \infty.$$  (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 (2007), $c^2_{NI} \to 1$ as $\rho \uparrow 1$ for all $s$.) The limit in (5) implies that any reasonable estimator ought to 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.

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

For a system having $s$ agents, each of whom on average completes one service request in $\mu^{-1}$ 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:

\[
\theta_{QL}(n) \equiv (n + 1)/s\mu .
\] (6)

The QL estimator is appealing due to its simplicity and ease of implementation: It uses information about the system that usually is readily available. In Ibrahim and Whitt (2007), the performance of QL is studied in the \(GI/M/s\) model, where there is no customer abandonment. For that model, \(W_Q(n)\) 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 \(\mu^{-1}\), 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, \(E[W_Q(n)]\). But, following the analysis above, \(E[W_Q(n)] = \theta_{QL}(n)\) in (6). Hence, QL is optimal for the \(GI/M/s\) model, under the MSE criterion. Extensive simulation experiments in Ibrahim and Whitt (2007) 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),

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

In the fluid limit,
Consistent with intuition, we see that QL overestimates \( w \). Indeed,

\[
\frac{\theta_{QL}(q) - w}{w} = \frac{(\rho - 1)/\alpha - \ln(\rho)/\alpha}{\ln(\rho)/\alpha}.
\] (8)

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

Motivated by the simple form of the QL delay estimate, \( \theta_{QL}(n) \) 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. The 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:

\[
W_Q(n) \equiv \sum_{i=0}^{n} Y_i,
\] (9)

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

\[
E[W_Q(n)] = \sum_{i=0}^{n} E[Y_i] = \sum_{i=0}^{n} \frac{1}{s\mu + i\alpha}.
\] (10)
The QL$_m$ estimator given to a customer who finds $n$ customers in queue upon arrival is $\theta_{QL_m}(n) \equiv E[W_Q(n)]$. 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. Non-exponential 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 non-exponential 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 non-exponential abandonment-time distribution, developed in Whitt (2005, 2006).

3.3. The 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 (ccdf) 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. Since “rate in” $\equiv \lambda F^c(w) = s\mu$ “rate out”, we have:

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

The associated equation for $q$ is

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

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

$$\theta_{QL}(q) = \frac{q + 1}{s\mu} \approx \frac{q}{s\mu} = \rho \int_0^w F^c(x) dx .$$
For QL_r, we propose computing the ratio $\beta = w/(q/s\mu) = w\mu/q$ (after solving numerically for $w$ and $q$), and using it to refine the QL estimator. That is, the new delay estimate is:

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

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 non-exponential abandonment-time distributions. Note that in addition to $s$, $n$ and $\mu$, we need to know $\rho$ or, equivalently, $\lambda$, and the abandonment-time cdf $F$ in order to implement QL_r.

3.4. The Exponential Abandonment Case (QL_r^m)

We now propose a modification of QL_r which does not depend on $\rho$. 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

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

Thus, the implementation of QL_r^m requires knowledge of $n$, $s$, $\mu$, and $\alpha$, but not of $\rho$ or, equivalently, $\lambda$. 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. The Approximation-Based Queue-Length Delay Estimator (QL\textsubscript{ap})

Our most promising estimator QL\textsubscript{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 \(j\)th from the end of the queue has an exponential abandonment time with rate \(\alpha_j\), where \(\alpha_j\) is given by

\[
\alpha_j \equiv h\left(\frac{j}{\lambda}\right), \quad 1 \leq j \leq k;
\]

(13)

\(k\) is the current queue length, and \(h\) is the abandonment-time hazard-rate function, defined as \(h(t) \equiv f(t)/F(t)\), \(t \geq 0\), where \(f\) is the corresponding density function (assumed to exist). Having \(\alpha_j\) 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\textsubscript{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 \(j\)th from the end of the line, \(1 \leq j \leq k\). 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. Since 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 \(j/\lambda\) and his abandonment rate by (13).

The associated total abandonment rate from the queue in that system state is \(\delta_k = \sum_{j=1}^{k} \alpha_j = \sum_{j=1}^{k} h(j/\lambda)\), \(k \geq 1\), and \(\delta_0 \equiv 0\).
For the GI/M/s + M(n) model, we need to make further approximations in order to describe the potential waiting time of a customer who finds \( n \) other customers waiting in line, upon arrival. We have the approximate representation:

\[
W_Q(n) \approx \sum_{i=0}^{n} X_i ,
\]

where \( X_{n-i} \) is the time between the \( i \)th 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 \( \mu \) (corresponding to the remaining service times of customers in service), and \( n \) exponential random variables with rates \( \alpha_j, 1 \leq j \leq n \), (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 \( X_i \)'s is more complicated. Since individual customers have different abandonment rates which, 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 (in order 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).

Here is what we do: 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 \( X_i \)'s in (14) by exponential random variables. Let \( X_{n-l} \), which is the time between the \( l \)th 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 \( \mu \) (corresponding to the remaining service times of customers in service), and \( n - l \) exponential random variables with rates \( \alpha_i, l + 1 \leq i \leq n \) (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
\[ \theta_{QL_{ap}}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_{n} - \delta_{n-i}}. \]  

(15)

Since 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_{m}^{r} are all independent of the arrival process: For these three estimators, the arrival process can be arbitrary, even non-stationary. The QL_{r} and QL_{ap} estimators require knowledge of the arrival rate \( \lambda \), 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 (2007). We emphasize that delay-history-based estimators apply directly to more complex settings, such as models including customer response to delay announcements.

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

As in Armony et al. (2008), 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: \( \theta_{LES}(w) \equiv w \). 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 $G/M/s+M$

Since 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 \( \lambda \). 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 (FWLLN) and a FCLT refinement:

\[ \bar{A}_s(t) \Rightarrow \lambda t \quad \text{in} \quad D \quad \text{and} \quad \sqrt{s}(\bar{A}_s(t) - \lambda t) \Rightarrow \sqrt{s}c_a^2 B(t) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \quad (16) \]

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 \( \lambda \) 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 (2008).

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

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

and

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

where \( B \) is a standard Brownian motion, while \( c \) and \( d \) are the deterministic real-valued functions

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

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

\[ \sqrt{s}(W_{Q,s}(\lfloor st \rfloor) - c(t)) \Rightarrow N(0,d(t)) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad s \to \infty \quad \text{for each} \quad t. \quad (20) \]
As a further consequence, we obtain the following result for the best-possible estimators \( \theta_{QL,m,s}(n) \). We use a random time change by the fluid limit

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


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

\[
sMSE(\theta_{QL,m,s}([st])) \equiv s\text{Var}(W_{Q,s}([st])) \to d(t) \quad \text{as} \quad s \to \infty \quad (22)
\]

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

\[
sMSE(\theta_{QL,m,s}(Q_s(\infty))) \equiv s\text{Var}(W_{Q,s}(Q_s(\infty))) \Rightarrow d(q) \equiv \frac{q}{\lambda \mu} \equiv \frac{\lambda - \mu}{\lambda \mu \alpha} \quad \text{as} \quad s \to \infty. \tag{23}
\]

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.,

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

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

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

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 (2008) 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 (2008) 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 (2008).

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

\[
\hat{Q}_s(\infty) \equiv \sqrt{s}(\bar{Q}_s(\infty) - q) \Rightarrow N \left(0, \frac{\lambda(c_a^2 + 1)}{2\alpha}\right) \quad \text{as} \quad s \to \infty
\]

and

\[
\hat{W}_s(\infty) \equiv \sqrt{s}(W_s(\infty) - w) \Rightarrow N \left(0, \sigma_w^2\right) \quad \text{as} \quad s \to \infty,
\]

where \(\sigma_w^2 = \frac{1/\alpha \mu + (c_a^2 - 1)/2\lambda \alpha}{\alpha \mu + (c_a^2 - 1)/2\lambda \alpha}\), 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 no-information (NI) estimator, assuming appropriate uniform integrability. The no-information 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,

\[
sMSE(\theta_{NI,s}) \equiv s\text{Var}(W_s(\infty)) \to \frac{1}{\alpha \mu} + \frac{c_a^2 - 1}{2\lambda \alpha} \quad \text{as} \quad s \to \infty.
\]

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

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

\[
\frac{MSE(\theta_{NI,s})}{E[MSE(\theta_{QLm,s}(Q_s(\infty)))]} \to \frac{2\lambda + \mu(c_a^2 - 1)}{2(\lambda - \mu)} > 1 \quad \text{as} \quad s \to \infty.
\]

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

\[
W_{LES,s}(w) \approx W_{QL,s}(N_s(w)) \equiv \sum_{i=0}^{N_s(w)} X_{s,i},
\]
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 \( \alpha \). 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

\[
E[N_s(t)] = \frac{s \lambda}{\alpha} \left(1 - e^{-\alpha t}\right), \quad t \geq 0.
\]

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 \( \alpha \),

\[
\bar{N}_s(t) \equiv \frac{N_s(t)}{s} \Rightarrow a(t) \equiv \frac{\lambda}{\alpha} \left(1 - e^{-\alpha t}\right) \quad \text{in } D \quad \text{as } s \to \infty
\]

and

\[
\hat{N}_s(t) \equiv \sqrt{s}(\bar{N}_s(t) - a(t)) \Rightarrow \hat{G}(t) \quad \text{in } D \quad \text{as } s \to \infty,
\]

where \( \hat{G} \equiv \{\hat{G}(t) : t \geq 0\} \) is a Gaussian stochastic process with

\[
\hat{G}(t) \overset{d}{=} N(0, \sigma^2_n(t)) \quad \text{where} \quad \sigma^2_n(t) \equiv a(t) + \frac{\lambda \left(c^2_n - 1\right)}{2\alpha} \left(1 - e^{-2\alpha t}\right),
\]

for \( a(t) \) defined in (32) and \( c^2_n \) 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 \( \hat{N}_s \) in (33) and for \( \hat{W}_s(\infty) \) 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 $s \to \infty$,

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

$$\hat{W}_{LES,s}(W_s(\infty)) \equiv \sqrt{s}(W_{LES,s}(W_s(\infty)) - W_s(\infty)) \Rightarrow N(0, \sigma_{LES}^2),$$

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

where

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

for $\sigma_w^2$ 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,

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

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

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

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

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

where

$$r(LES, QL_m) = 2 \quad (\geq 2 \quad \text{or} \quad \leq 2) \quad \text{if and only if} \quad c_u^2 = 1 \quad (\geq 1 \quad \text{or} \quad \leq 1).$$
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

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

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

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

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

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 non-exponential 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, $\mu$, be equal to 1. We do this without loss of generality, since 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, $\alpha$, 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 $\lambda$ to get a fixed value of $\rho$, 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 $q \approx (\rho - 1)s/\alpha \approx 40$, while the average potential waiting time is about $w \approx q/s\mu \approx 0.4/\mu$ (less than half a mean service time).

Our simulation are steady-state simulations. The simulation results are based on 10 independent replications of 5 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 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 $s = 1000$. 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 $s = 1000$. Figure 2 shows that $s \times \text{ASE}(QL_m)$, 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 $s \times \text{ASE}(QL_m) \approx (\lambda - \mu)/(\lambda \mu \alpha)$, as in (25) of §5. The relative error between the simulation estimates for $\text{ASE}(QL_m)$ and the numerical value given by (25) is less than 1% throughout.

The $QL_m^n$ 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)

$$\theta_{QL_m}(n) \approx \int_0^n \frac{1}{s\mu + \alpha x} dx = \ln (s\mu + \alpha n) - \ln (s\mu) = \frac{1}{\alpha} \ln (1 + \alpha n/s\mu) .$$
On the other hand, we have that

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

So that, for large $n$, the two estimators $QL_m$ and $QL^m$ 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 (2007) for the $GI/M/s$ model, without customer abandonment. Figure 2 shows that $s \times ASE(LES) \approx \sigma_{LES}^2$, 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 nearly 16 when \( s = 1000 \). Figure 2 shows that \( s \times \text{ASE}(\text{QL}) \) 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\(_{\text{ap}}\) is the best possible delay estimator, for this model, except when \( s \) is very large (e.g., \( s = 700 \) or \( s = 1000 \)). The corresponding RRASE ranges from about 10% when \( s = 100 \) to about 3% when \( s = 1000 \). The QL\(_r\) estimator performs worse than QL\(_{\text{ap}}\) for smaller values of \( s \), but slightly outperforms QL\(_{\text{ap}}\) for larger values of \( s \). The ratio \( \text{ASE}(\text{QL}_r)/\text{ASE}(\text{QL}_{\text{ap}}) \) ranges from nearly 2 when \( s = 100 \) to nearly 0.9 when \( s = 1000 \).

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\(_{\text{ap}}\), 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 \( s = 1000 \). Figure 6 shows that \( s \times \text{ASE}(\text{LES}) \) 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\(_{\text{ap}}\) before, now performs worse: The corresponding RRASE ranges from about 14% when \( s = 100 \) to about 10% when \( s = 1000 \).

QL\(_m\) is relatively effective when \( s = 100 \) but becomes significantly worse than QL\(_{\text{ap}}\) when \( s = 1000 \) (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 \( \text{ASE}(\text{QL})/\text{ASE}(\text{QL}_{\text{ap}}) \) ranges from about 15 when \( s = 100 \) to nearly 95 when \( s = 1000 \). 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 the online supplement, Ibrahim and Whitt (2008). 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 reported above, 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 $G_1/G_1/s + G_1$ 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 no-information delay estimator (NI), 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 non-exponential 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 $G_1/M/s$ queueing model, without customer abandonment. Whenever the actual service system is well modeled by a $G_1/M/s$ queueing model and the system state is known accurately at each time, then 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, $\theta_{QL}(n)$ in (6), we proposed modified queue-length-based delay estimators that account for customer abandonment, and that 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 $\alpha$. 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 $\alpha$, 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 (ACD) data.

But, $QL_m$ 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 non-exponential abandonment-time distribution. Since non-exponential 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 non-exponential 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
QLₘ (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, QLₐₚ. Simulation results in §6 and the e-companion show that QLₐₚ is consistently the most effective estimator (with the exception of D service; see the e-companion). It is a variant of QLₘ which assumes that abandonment times are independent, exponential, with state-dependent abandonment rates. The QLₐₚ estimator coincides with QLₘ 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 non-exponential abandonment-time distributions.

The QLₐₚ delay estimate, θ₉ₐₚ(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 QLᵥ and QLₐₚ 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, $c_2^2$. 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 (SBR); 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 on-going 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.

**Acknowledgments**

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

**References**


Available at http://columbia.edu/~ww2040.


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.
**e-Companion**

**EC.1. Introduction**

We present additional material in this e-companion. In §EC.2, we give the proof of Theorem 4. In §EC.3, we present detailed simulation results for the $M/M/s + GI$ model. In §EC.4, we present additional experimental results for non-exponential service-time distributions. In §EC.5, we present simulation results substantiating the heavy-traffic limits of §5, for the $GI/M/s + M$ model, with alternative interarrival-time distributions and alternative values of the abandonment rate $\alpha$. We present additional experimental results in the supplement to the main paper, available on the authors’ webpages, Ibrahim and Whitt (2008).

**EC.2. Proof of Theorem 4**

We now prove the convergence in distribution in (35). The proof follows the general approach used to prove Theorem 6.4 of Talreja and Whitt (2008), exploiting stochastic-process limits in order to obtain the desired one-dimensional limit in $\mathbb{R}$. As in (6.37) of Talreja and Whitt (2008), we use the continuous mapping theorem with the composition map to treat random time changes. We start with the joint convergence

$$ (\hat{W}_{Q,s}(t), \hat{N}_s(t), \hat{W}_s(\infty)) \Rightarrow (B(d(t)), \hat{G}(t), N(0, \sigma_w^2)) \quad \text{in} \quad D^2 \times \mathbb{R} \quad \text{(EC.1)} $$

for the processes defined in (18), (33) and (27), where the limits are mutually independent.

For the $M/M/s + M$ model, we can obtain the joint convergence from the individual limits established above, because we can regard the component processes on the left as mutually independent. That requires some comment, however. First in time we have the waiting time for the last customer to enter service, which is distributed asymptotically the same as $W_s(\infty)$. Then we have the buildup of the queue behind this customer until this customer starts service, given by $\hat{N}_s(t)$. Finally, we have the remaining times between successive departures after the new arrival enters the system, as given by $\hat{W}_{Q,s}(t)$, which involves independent exponential random variables. These are mutually independent with reference to the designated arrival at one fixed time, for whom we are doing the
estimation. The processes are well defined as independent random elements of $D$, but they only
correctly apply to describe our system at a single time, as stated in the final one-dimensional limit
in (35). (In the case of the $G/M/s + M$ model, we assume that the joint limit of $(\tilde{N}_s(t), \tilde{W}_s(\infty))$ is
the same as if these were independent.)

Assuming the limit in (EC.1), since $\tilde{N}_s$ converges to a deterministic limit, we can append the
limit for $\tilde{N}_s$ to get

$$(\hat{W}_{Q,s}(t), \tilde{N}_s(t), \tilde{W}_s(\infty)) \Rightarrow (B(d(t)), \hat{G}(t), a(t), N(0, \sigma^2_w)) \quad \text{in} \quad D^3 \times \mathbb{R}. \quad \text{(EC.2)}$$

We can now apply the continuous mapping theorem with composition to perform a random time
change with $\tilde{N}_s$ to obtain the limit

$$\hat{W}_{Q,s}(\tilde{N}_s(t)) \equiv \sqrt{s} (W_{Q,s}(N_s(t)) - c(\tilde{N}_s(t))) \Rightarrow B(d(a(t))) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \quad \text{(EC.3)}$$

jointly with the limit in (EC.2), where $B$ is the given standard Brownian motion and $a(t)$ is defined
in (32). We can now apply a random-time-change argument one more time with $W_s(\infty)$ to obtain
the limit

$$\hat{Z}_s \equiv \sqrt{s} (W_{Q,s}(N_s(\infty))) \Rightarrow B(d(a(w))) \quad \text{in} \quad \mathbb{R} \quad \text{(EC.4)}$$
as $s \to \infty$, again jointly with the limit in (EC.2), where again the limit involves the same Brownian
motion $B$.

We obtain the desired limit in (35) by writing

$$\hat{W}_{LES,s}(W_s(\infty)) \equiv \sqrt{s} (W_{LES,s}(W_s(\infty)) - W_s(\infty)) \equiv \hat{Z}_s + \hat{Y}_s \quad \text{(EC.5)}$$

for $\hat{Z}_s$ in (EC.4) and

$$\hat{Y}_s \equiv \sqrt{s} (c(\tilde{N}_s(W_s(\infty))) - W_s(\infty)) \quad \text{(EC.6)}$$

and establishing a limit for $\hat{Y}_s$ in (EC.6) within the framework of the initial limits in (EC.2). In
order to make a connection to the given limits for $(\tilde{N}_s(t), \tilde{W}_s(\infty))$ in (EC.2), we exploit a Taylor
series expansion for the functions $c(t)$ and $a(t)$ in (19) and (32). Note that

$$c(q) = w \equiv \frac{1}{\alpha} \ln (\rho), \quad a(w) = q \equiv \frac{\lambda - \mu}{\alpha} \quad \text{and} \quad d(q) = \frac{q}{\lambda \mu}. \quad \text{(EC.7)}$$
Hence, \( d(a(w)) = d(q) = q/(\lambda \mu) \).

We write

\[
\hat{Y}_s \equiv \sqrt{s} \left( c(\bar{N}_s(W_s(\infty))) - W_s(\infty) \right) \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3},
\]

where

\[
\begin{align*}
\hat{Y}_{s,1} &\equiv \sqrt{s} \left( c(\bar{N}_s(W_s(\infty))) - c(a(W_s(\infty))) \right), \\
\hat{Y}_{s,2} &\equiv \sqrt{s} \left( c(a(W_s(\infty))) - c(a(w)) \right), \\
\hat{Y}_{s,3} &\equiv \sqrt{s} \left( c(a(w)) - W_s(\infty) \right),
\end{align*}
\]

(EC.8)

Using a Taylor series expansion of \( c \), we see that

\[
\hat{Y}_{s,1} - c'(a(w)) \sqrt{s} \left( \bar{N}_s(W_s(\infty)) - a(W_s(\infty)) \right) \Rightarrow 0,
\]

(EC.10)

where \( c'(a(w)) = 1/\lambda \). By Theorem 3,

\[
\hat{Y}_{s,1} \Rightarrow \frac{1}{\lambda} \hat{G}(w) \equiv N(0, \sigma_w^2(w)/\lambda^2) \quad \text{as} \quad s \to \infty.
\]

(EC.11)

Using a Taylor series expansion of \( c \circ a \), noting that \( a'(w) = \mu \), we get

\[
\hat{Y}_{s,2} - c'(a(w))a'(w) \sqrt{s} \left( W_s(\infty) - w \right) \Rightarrow 0,
\]

(EC.12)

so that, by Theorem 2,

\[
\hat{Y}_{s,2} \Rightarrow \frac{\mu}{\lambda} N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty.
\]

(EC.13)

Similarly, using the relation \( c(a(w)) = c(q) = w \) and replacing \( c(a(w)) \) by \( w \), we get

\[
\hat{Y}_{s,3} - \sqrt{s} \left( w - W_s(\infty) \right) \Rightarrow 0,
\]

(EC.14)

so that, by Theorem 2 again,

\[
\hat{Y}_{s,3} \Rightarrow N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty,
\]

(EC.15)

where the limiting random variables \( N(0, \sigma_w^2) \) in (EC.13) and (EC.15) are identical. By these constructions, we obtain convergence of the vector \( (\hat{Y}_{s,1}, \hat{Y}_{s,2}, \hat{Y}_{s,3}) \) jointly with the initial limits.
in (EC.2) and thus also jointly with \( \hat{Z}_s \) in (EC.4). The processes \( \hat{Y}_{s,i} \) are each asymptotically equivalent to processes that are simple functions of the processes in the original limit (EC.2).

Hence,

\[
\hat{Y}_s \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3} \Rightarrow N \left( 0, \frac{\sigma^2(w)}{\lambda^2} + \frac{(\lambda - \mu)^2 \sigma^2_w}{\lambda^2} \right). \tag{EC.16}
\]

We can thus obtain the limit from (EC.4)–(EC.6), (EC.8), (EC.9) and (EC.16) by adding the normal components.

**EC.3. Simulation Results for the \( M/M/s + GI \) Model**

In this section, we present tables of simulation results (point and 95% confidence interval estimates) quantifying the performance of the alternative delay estimators in the \( M/M/s + GI \) model. The corresponding plots are shown and discussed in §6. For the abandonment-time distribution, we consider \( M, H_2, \) and \( E_{10} \) distributions. We consider alternative values of \( s \geq 100, \) and vary the arrival rate, \( \lambda, \) to keep the traffic intensity, \( \rho, \) fixed for alternative values of \( s (\rho = 1.4). \) We let the abandonment rate, \( \alpha, \) be equal to 1.

With exponential abandonments, Table EC.1 shows that, consistent with theory, \( QL_m \) is the best possible delay estimator, under the MSE criterion. The \( QL_r^m \) and \( QL_r \) estimators are nearly identical, with \( QL_r^m \) slightly outperforming \( QL_r. \) They are both nearly as efficient as \( QL_m. \) Consistent with (38), the LES estimator performs worse than \( QL_m, \) but not greatly so: The relative error between the simulation estimates for \( ASE(LES)/ASE(QL_m) \) and the numerical value, 2, given by (38) is less than 1% throughout. Consistent with (29), the NI estimator is less efficient than \( QL_m: \) The relative error between the simulation estimates for \( ASE(NI)/ASE(QL_m) \) and the numerical value, 3.5, given by (29) is less than 1% throughout. The QL estimator performs significantly worse than all the other estimators, particularly for large values of \( s. \) The ratio \( ASE(QL)/ASE(QL_m) \) ranges from about 3 when \( s = 100 \) to nearly 16 when \( s = 1000. \)

With hyperexponential abandonments, Table EC.2 shows that \( QL_{ap} \) is the best delay estimator. The \( QL_r \) estimator performs nearly the same as \( QL_{ap} \) and is only slightly outperformed. The \( QL_m \) estimator, which is optimal for the \( GI/M/s + M \) model, is now outperformed by \( QL_r, \) particularly
Efficiency of the estimators in the $M/M/s + M$ model with $\rho = 1.4$ and $\alpha = 1.0$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$\text{ASE}[\theta_{QL_m}]$</th>
<th>$\text{ASE}[\theta_{QL_p}]$</th>
<th>$\text{ASE}[\theta_{QL}]$</th>
<th>$\text{ASE}[\theta_{LES}]$</th>
<th>$\text{ASE}[\theta_{NI}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$2.867 \times 10^{-3}$</td>
<td>$2.869 \times 10^{-3}$</td>
<td>$3.130 \times 10^{-3}$</td>
<td>$8.693 \times 10^{-3}$</td>
<td>$5.772 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.76 \times 10^{-5}$</td>
<td>$\pm 1.78 \times 10^{-5}$</td>
<td>$\pm 1.89 \times 10^{-5}$</td>
<td>$\pm 3.20 \times 10^{-5}$</td>
<td>$\pm 2.79 \times 10^{-5}$</td>
</tr>
<tr>
<td>300</td>
<td>$9.587 \times 10^{-4}$</td>
<td>$9.601 \times 10^{-4}$</td>
<td>$1.039 \times 10^{-3}$</td>
<td>$5.602 \times 10^{-3}$</td>
<td>$1.922 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 6.86 \times 10^{-6}$</td>
<td>$\pm 6.92 \times 10^{-6}$</td>
<td>$\pm 6.41 \times 10^{-6}$</td>
<td>$\pm 2.64 \times 10^{-5}$</td>
<td>$\pm 1.50 \times 10^{-5}$</td>
</tr>
<tr>
<td>500</td>
<td>$5.761 \times 10^{-4}$</td>
<td>$5.661 \times 10^{-4}$</td>
<td>$6.224 \times 10^{-4}$</td>
<td>$5.017 \times 10^{-3}$</td>
<td>$1.153 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.94 \times 10^{-6}$</td>
<td>$\pm 3.86 \times 10^{-6}$</td>
<td>$\pm 2.94 \times 10^{-6}$</td>
<td>$\pm 2.41 \times 10^{-5}$</td>
<td>$\pm 9.99 \times 10^{-6}$</td>
</tr>
<tr>
<td>700</td>
<td>$4.104 \times 10^{-4}$</td>
<td>$4.201 \times 10^{-4}$</td>
<td>$4.440 \times 10^{-4}$</td>
<td>$4.682 \times 10^{-3}$</td>
<td>$8.166 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.82 \times 10^{-6}$</td>
<td>$\pm 2.839 \times 10^{-6}$</td>
<td>$\pm 2.71 \times 10^{-6}$</td>
<td>$\pm 2.40 \times 10^{-5}$</td>
<td>$\pm 5.78 \times 10^{-6}$</td>
</tr>
<tr>
<td>1000</td>
<td>$2.892 \times 10^{-4}$</td>
<td>$2.839 \times 10^{-4}$</td>
<td>$3.136 \times 10^{-4}$</td>
<td>$4.492 \times 10^{-3}$</td>
<td>$5.752 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 3.48 \times 10^{-6}$</td>
<td>$\pm 3.86 \times 10^{-6}$</td>
<td>$\pm 3.09 \times 10^{-6}$</td>
<td>$\pm 1.54 \times 10^{-5}$</td>
<td>$\pm 6.91 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

Table EC.1 Point and confidence interval estimates of the ASEs - average square errors - of the estimators

when $s$ is large (e.g., $\text{ASE}(QL_m)/\text{ASE}(QL_p)$ is close to 2 when $s = 1000$). The LES estimator performs worse than $QL_m$ when $s = 100$, but is nearly identical to $QL_m$ when $s = 1000$. The NI estimator performs worse than LES, but not as bad as QL. Once more, QL is the least efficient delay estimator: The ratio $\text{ASE}(QL)/\text{ASE}(QL_p)$ ranges from about 2 when $s = 100$ to about 10 when $s = 1000$.

With Erlang abandonments, Table EC.3 shows that $QL_p$ is, once more, the best possible delay estimator, except when $s$ is very large (e.g., $s = 700$ or $s = 1000$). The $QL_r$ estimator performs worse than $QL_p$ for relatively small values of $s$, but slightly outperforms $QL_p$ for relatively large values of $s$. The NI estimator is more competitive in this model, than in the previous two models. It is nearly as efficient as $QL_p$, particularly when $s$ is large. The LES estimator also fares well, but is slightly outperformed by $QL_p$, $QL_r$ and NI. The $QL_m$ estimator performs significantly worse than $QL_p$ when $s$ is large (e.g., when $s = 1000$, $\text{ASE}(QL_m)/\text{ASE}(QL_p) \approx 9$). Finally, QL is yet again the least effective estimator for this model. The ratio $\text{ASE}(QL)/\text{ASE}(QL_p)$ ranges from about 15 when $s = 100$ to about 93 when $s = 1000$. 
Efficiency of the estimators in the $M/M/s + H_2$ model with $\rho = 1.4$ and $\alpha = 1.0$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$ASE[\theta_{QL_{ap}}]$</th>
<th>$ASE[\theta_{QL_{m}}]$</th>
<th>$ASE[\theta_{QL_r}]$</th>
<th>$ASE[\theta_{QL}]$</th>
<th>$ASE[\theta_{LES}]$</th>
<th>$ASE[\theta_{NI}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$1.859 \times 10^{-3}$</td>
<td>$2.100 \times 10^{-3}$</td>
<td>$2.032 \times 10^{-3}$</td>
<td>$4.662 \times 10^{-3}$</td>
<td>$3.866 \times 10^{-3}$</td>
<td>$6.503 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 6.52 \times 10^{-6}$</td>
<td>$\pm 5.54 \times 10^{-6}$</td>
<td>$\pm 6.31 \times 10^{-6}$</td>
<td>$\pm 1.83 \times 10^{-5}$</td>
<td>$\pm 8.10 \times 10^{-6}$</td>
<td>$\pm 3.85 \times 10^{-5}$</td>
</tr>
<tr>
<td>300</td>
<td>$6.116 \times 10^{-4}$</td>
<td>$7.933 \times 10^{-4}$</td>
<td>$6.599 \times 10^{-4}$</td>
<td>$2.593 \times 10^{-3}$</td>
<td>$1.236 \times 10^{-3}$</td>
<td>$2.165 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 4.64 \times 10^{-6}$</td>
<td>$\pm 7.62 \times 10^{-6}$</td>
<td>$\pm 8.82 \times 10^{-6}$</td>
<td>$\pm 2.25 \times 10^{-5}$</td>
<td>$\pm 1.76 \times 10^{-5}$</td>
<td>$\pm 2.09 \times 10^{-5}$</td>
</tr>
<tr>
<td>500</td>
<td>$3.695 \times 10^{-4}$</td>
<td>$5.367 \times 10^{-4}$</td>
<td>$3.921 \times 10^{-4}$</td>
<td>$2.205 \times 10^{-3}$</td>
<td>$7.331 \times 10^{-4}$</td>
<td>$1.311 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.19 \times 10^{-6}$</td>
<td>$\pm 2.12 \times 10^{-6}$</td>
<td>$\pm 2.47 \times 10^{-6}$</td>
<td>$\pm 9.97 \times 10^{-6}$</td>
<td>$\pm 5.41 \times 10^{-6}$</td>
<td>$\pm 1.03 \times 10^{-5}$</td>
</tr>
<tr>
<td>700</td>
<td>$2.630 \times 10^{-4}$</td>
<td>$4.257 \times 10^{-4}$</td>
<td>$2.802 \times 10^{-4}$</td>
<td>$2.024 \times 10^{-3}$</td>
<td>$5.250 \times 10^{-4}$</td>
<td>$9.378 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.43 \times 10^{-6}$</td>
<td>$\pm 1.89 \times 10^{-6}$</td>
<td>$\pm 1.00 \times 10^{-5}$</td>
<td>$\pm 2.35 \times 10^{-5}$</td>
<td>$\pm 2.52 \times 10^{-5}$</td>
<td>$\pm 1.07 \times 10^{-5}$</td>
</tr>
<tr>
<td>1000</td>
<td>$1.833 \times 10^{-4}$</td>
<td>$3.474 \times 10^{-4}$</td>
<td>$1.978 \times 10^{-4}$</td>
<td>$1.900 \times 10^{-3}$</td>
<td>$3.691 \times 10^{-4}$</td>
<td>$6.533 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.55 \times 10^{-6}$</td>
<td>$\pm 1.43 \times 10^{-6}$</td>
<td>$\pm 6.90 \times 10^{-7}$</td>
<td>$\pm 5.93 \times 10^{-6}$</td>
<td>$\pm 3.00 \times 10^{-6}$</td>
<td>$\pm 1.14 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Table EC.2  Point and confidence interval estimates of the ASES - average square errors - of the estimators

Efficiency of the estimators in the $M/M/s + E_{10}$ model with $\rho = 1.4$ and $\alpha = 1.0$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$ASE[\theta_{QL_{ap}}]$</th>
<th>$ASE[\theta_{QL_{m}}]$</th>
<th>$ASE[\theta_{QL_r}]$</th>
<th>$ASE[\theta_{QL}]$</th>
<th>$ASE[\theta_{LES}]$</th>
<th>$ASE[\theta_{NI}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$5.388 \times 10^{-4}$</td>
<td>$9.409 \times 10^{-4}$</td>
<td>$6.317 \times 10^{-3}$</td>
<td>$8.097 \times 10^{-2}$</td>
<td>$8.810 \times 10^{-3}$</td>
<td>$6.077 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.54 \times 10^{-5}$</td>
<td>$\pm 3.48 \times 10^{-5}$</td>
<td>$\pm 4.51 \times 10^{-5}$</td>
<td>$\pm 2.47 \times 10^{-4}$</td>
<td>$\pm 3.91 \times 10^{-5}$</td>
<td>$\pm 2.63 \times 10^{-5}$</td>
</tr>
<tr>
<td>300</td>
<td>$1.955 \times 10^{-3}$</td>
<td>$7.211 \times 10^{-3}$</td>
<td>$2.139 \times 10^{-3}$</td>
<td>$7.211 \times 10^{-2}$</td>
<td>$2.933 \times 10^{-3}$</td>
<td>$2.040 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 5.13 \times 10^{-6}$</td>
<td>$\pm 3.86 \times 10^{-5}$</td>
<td>$\pm 1.83 \times 10^{-5}$</td>
<td>$\pm 3.301 \times 10^{-4}$</td>
<td>$\pm 3.22 \times 10^{-5}$</td>
<td>$\pm 2.23 \times 10^{-5}$</td>
</tr>
<tr>
<td>500</td>
<td>$1.244 \times 10^{-3}$</td>
<td>$6.746 \times 10^{-3}$</td>
<td>$1.293 \times 10^{-3}$</td>
<td>$7.049 \times 10^{-2}$</td>
<td>$1.760 \times 10^{-3}$</td>
<td>$1.288 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.54 \times 10^{-5}$</td>
<td>$\pm 2.68 \times 10^{-5}$</td>
<td>$\pm 1.35 \times 10^{-5}$</td>
<td>$\pm 2.48 \times 10^{-4}$</td>
<td>$\pm 2.44 \times 10^{-5}$</td>
<td>$\pm 2.61 \times 10^{-5}$</td>
</tr>
<tr>
<td>700</td>
<td>$9.572 \times 10^{-4}$</td>
<td>$6.584 \times 10^{-3}$</td>
<td>$9.319 \times 10^{-4}$</td>
<td>$6.975 \times 10^{-2}$</td>
<td>$1.241 \times 10^{-3}$</td>
<td>$9.966 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 8.31 \times 10^{-6}$</td>
<td>$\pm 1.43 \times 10^{-6}$</td>
<td>$\pm 1.00 \times 10^{-5}$</td>
<td>$\pm 1.00 \times 10^{-5}$</td>
<td>$\pm 2.35 \times 10^{-6}$</td>
<td>$\pm 1.30 \times 10^{-5}$</td>
</tr>
<tr>
<td>1000</td>
<td>$7.369 \times 10^{-4}$</td>
<td>$6.454 \times 10^{-3}$</td>
<td>$6.694 \times 10^{-4}$</td>
<td>$6.902 \times 10^{-2}$</td>
<td>$8.830 \times 10^{-4}$</td>
<td>$8.242 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.96 \times 10^{-5}$</td>
<td>$\pm 1.70 \times 10^{-5}$</td>
<td>$\pm 1.13 \times 10^{-5}$</td>
<td>$\pm 1.68 \times 10^{-4}$</td>
<td>$\pm 1.28 \times 10^{-5}$</td>
<td>$\pm 1.17 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

Table EC.3  Point and confidence interval estimates of the ASES - average square errors - of the estimators

EC.4.  Simulation Results for the $M/GI/s + M$ Model

In this section we present simulation results quantifying the performance of the alternative delay estimators with non-exponential service-time distributions; i.e., we consider the $M/GI/s + M$ model. In this model, $QL_{ap}$ coincides with $QL_{m}$, so we do not include separate results for it. For the service-time distribution, we consider $D$, $E_{10}$, and $LN(1,1)$ (lognormal with mean and variance
equal to 1) distributions. We let $\mu = \alpha = 1.0$, and vary $\lambda$, for alternative values of $s$, to keep $\rho = 1.4$. Corresponding tables with estimates of the 95% confidence intervals, and additional simulation results for the $M/GI/s + M$ model, are presented in the supplement, Ibrahim and Whitt (2008).

**EC.4.1. Results for the $M/D/s + M$ model**

Figures EC.1 and EC.2 show that all delay estimators do not perform well in this model. The NI estimator, which uses no information at all beyond the model, is the most effective delay estimator, when $s \geq 300$. (For $s = 100$, QL$_m$ slightly outperforms NI.) But even the NI estimator is not very accurate: The RRASE for NI is roughly equal to 25% for all values of $s$ considered. This suggests that our procedures for estimating delays perform relatively poorly when the service times are deterministic. The ASE’s for QL$_m$, QL$_r$, QL, and LES do not vary much in this model; e.g., ASE(QL$_m$) varies little about 0.01, for all values of $s$ considered. Figure EC.2 shows that, unlike previous models, the accuracy of the estimators does not improve as the number of servers increases. Alternative delay estimation procedures, appropriate for deterministic service times, remain to be investigated.

**EC.4.2. Results for the $M/E_{10}/s + M$ model**

Simulation results with an $E_{10}$ distribution (SCV = 0.1) for the service times, suggest that the proposed delay estimators remain effective, even with very low variability in the service times.
Figures EC.3 and EC.4 show that QL

is the most effective delay estimator for this model. The QL

estimator is nearly identical to QL

, particularly when s is large enough (s ≥ 300). Once more, the relative accuracy of the delay estimators improves as s increases. The RRASE for QL

ranges from approximately 13% when s = 100 to approximately 4% when s = 1000. The LES estimator is relatively accurate as well: The RRASE of LES ranges from approximately 21% when s = 100 to approximately 7% when s = 1000. The NI estimator does not perform as well as LES, nor as bad as QL. The QL estimator is the least efficient estimator: The ratio ASE(QL)/ASE(QL

) ranges from approximately 4 when s = 100 to approximately 22 when s = 1000. Consistent with §5, Figure EC.4 shows that all estimators, except QL, have an ASE which is inversely proportional to the number of servers, but mathematical support for the estimators has yet to be provided with non-exponential service-time distributions.

EC.4.3. Results for the M/LN(1,1)/s + M model

We consider the lognormal distribution for the service times because there is empirical evidence suggesting a remarkable fit of the service-time distribution to the lognormal distribution; e.g., see Brown et al. (2005). Table EC.4 shows that QL

is the most effective delay estimator for this model. The RRASE for QL

ranges from approximately 14% when s = 100 to approximately 5% when
heavily overloaded when α changing by Formulas (3) and (7), the queue length and delay tend to be inversely proportional to the model. For the interarrival-time distribution, we consider D_\text{estimators with non-exponential interarrival-time distributions}; i.e., we consider the GI/M/s model. In this section, we present simulation results quantifying the performance of the alternative delay estimators with non-exponential interarrival-time distributions; i.e., we consider the GI/M/s model.

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
s & ASE[\theta_{QL_m}] & ASE[\theta_{QL}] & ASE[\theta_{LES}] & ASE[\theta_{NI}] \\
\hline
100 & $2.359 \times 10^{-3}$ & $2.596 \times 10^{-3}$ & $8.207 \times 10^{-3}$ & $5.248 \times 10^{-3}$ & $9.089 \times 10^{-3}$ \\
& $\pm 7.00 \times 10^{-6}$ & $\pm 9.02 \times 10^{-6}$ & $\pm 4.45 \times 10^{-5}$ & $\pm 2.37 \times 10^{-5}$ & $\pm 4.80 \times 10^{-5}$ \\
300 & $7.810 \times 10^{-4}$ & $8.506 \times 10^{-4}$ & $5.394 \times 10^{-3}$ & $1.716 \times 10^{-3}$ & $3.032 \times 10^{-3}$ \\
& $\pm 5.14 \times 10^{-6}$ & $\pm 5.68 \times 10^{-6}$ & $\pm 3.36 \times 10^{-5}$ & $\pm 1.25 \times 10^{-5}$ & $\pm 5.30 \times 10^{-5}$ \\
500 & $4.663 \times 10^{-4}$ & $5.085 \times 10^{-4}$ & $4.836 \times 10^{-3}$ & $1.029 \times 10^{-3}$ & $1.826 \times 10^{-3}$ \\
& $\pm 2.04 \times 10^{-6}$ & $\pm 2.12 \times 10^{-6}$ & $\pm 2.085 \times 10^{-5}$ & $\pm 7.29 \times 10^{-6}$ & $\pm 8.10 \times 10^{-6}$ \\
700 & $3.346 \times 10^{-4}$ & $3.635 \times 10^{-4}$ & $4.615 \times 10^{-3}$ & $7.438 \times 10^{-4}$ & $1.290 \times 10^{-3}$ \\
& $\pm 2.71 \times 10^{-6}$ & $\pm 3.37 \times 10^{-6}$ & $\pm 1.77 \times 10^{-5}$ & $\pm 6.47 \times 10^{-6}$ & $\pm 1.12 \times 10^{-5}$ \\
1000 & $2.340 \times 10^{-4}$ & $2.548 \times 10^{-4}$ & $4.443 \times 10^{-3}$ & $5.290 \times 10^{-4}$ & $8.942 \times 10^{-4}$ \\
& $\pm 1.84 \times 10^{-6}$ & $\pm 2.81 \times 10^{-6}$ & $\pm 2.54 \times 10^{-5}$ & $\pm 5.90 \times 10^{-6}$ & $\pm 2.46 \times 10^{-5}$ \\
\hline
\end{tabular}
\end{center}

\textbf{Table EC.4} \hspace{1cm} \text{Point and confidence interval estimates of the ASEs - average square errors - of the estimators}

$s = 1000$. The QL_r estimator is slightly less efficient than QL_m: The ratio ASE(QL_r)/ASE(QL_m) ranges from approximately 1.1 when $s = 100$ to approximately 1.08 when $s = 1000$. The LES estimator is relatively accurate as well: The RRASE of LES ranges from approximately 26% when $s = 100$ to approximately 7% when $s = 1000$. The NI estimator does not perform as well as LES, nor as bad as QL. The QL estimator is the least efficient estimator: the ratio ASE(QL)/ASE(QL_m) ranges from approximately 4 when $s = 100$ to approximately 19 when $s = 1000$.

\textbf{EC.5. Simulations Results for the GI/M/s + M Model}

In this section, we present simulation results quantifying the performance of the alternative delay estimators with non-exponential interarrival-time distributions; i.e., we consider the GI/M/s + M model. For the interarrival-time distribution, we consider $D$ and $H_2$ distributions.

We also consider different abandonment rates; specifically we let $\alpha = 0.2$ and $\alpha = 5.0$. As indicated by Formulas (3) and (7), the queue length and delay tend to be inversely proportional to $\alpha$. Thus, changing $\alpha$ from 1.0 to 0.2 or 5.0 tends to change congestion by a factor of 5. The system is very heavily overloaded when $\alpha = 0.2$, but relatively lightly loaded when $\alpha = 5.0$.

We consider the same values of $s$ as before and we let $\mu = 1$. We vary $\lambda$ to get a fixed value of $\rho$.
(ρ = 1.4), for alternative values of s. Additional simulation results for the GI/M/s + M model are presented in the supplement, Ibrahim and Whitt (2008).

**EC.5.1. Results for the D/M/s + M model with α = 0.2**

Table EC.5 compares the efficiencies of the alternative delay estimators in the D/M/s + M model with α = 0.2. Consistent with theory, QLm is the optimal delay estimator for this model, under the MSE criterion. The RRASE of QLm ranges from approximately 35% when s = 100 to approximately 11% when s = 1000. The QLr estimator is slightly less efficient than QLm: ASE(QLr)/ASE(QLm) is less than 1.05 for all values of s considered. The LES estimator is slightly less accurate, with an RRASE ranging from approximately 40% when s = 100 to approximately 13% when s = 1000. The NI estimator is less accurate than LES, but not as bad as QL. The QL estimator is, once more, the least effective estimator: The ratio ASE(QL)/ASE(QLm) ranges from approximately 8 when s = 100 to approximately 71 when s = 1000.

Tables EC.6 and EC.7 substantiate (39) and (29) of §5, that compare the performances of QLm, LES and NI in the D/M/s + M model. Consistent with (39), Table EC.6 shows that the performance of LES is close to that of QLm, when the arrival process is deterministic. The simulation estimates of ASE(LES)/ASE(QLm), for alternative values of s, are remarkably close to the numerical value, approximately 1.286, predicted by (39); the relative error (RE) observed is less than 1% for all values of s considered. Consistent with (29), Table EC.7 shows that the performance of NI is worse than that of LES and QLm. The simulation estimates of ASE(NI)/ASE(QLm) are also remarkably close to the numerical value, 2.25, predicted by (29); the RE observed is less than 4% for all values of s considered.

**EC.5.2. Results for the H₂/M/s + M model**

Table EC.8 compares the efficiencies of the alternative delay estimators in the H₂/M/s + M model with α = 5.0, which makes the model more lightly loaded. Consistent with theory, QLm is the optimal delay estimator for this model, under the MSE criterion. The RRASE of QLm ranges from approximately 8% when s = 100 to approximately 2% when s = 1000.
Efficiency of the estimators in the \( D/M/s + M \) model with \( \rho = 1.4 \) and \( \alpha = 0.2 \)

<table>
<thead>
<tr>
<th>( s )</th>
<th>( \text{ASE}[\theta_{QL_m}] )</th>
<th>( \text{ASE}[\theta_{QL_L}] )</th>
<th>( \text{ASE}[\theta_{QL}] )</th>
<th>( \text{ASE}[\theta_{LES}] )</th>
<th>( \text{ASE}[\theta_{NI}] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>( 1.436 \times 10^{-2} )</td>
<td>( 1.492 \times 10^{-2} )</td>
<td>( 1.192 \times 10^{-1} )</td>
<td>( 1.863 \times 10^{-2} )</td>
<td>( 3.266 \times 10^{-2} )</td>
</tr>
<tr>
<td></td>
<td>( \pm 9.78 \times 10^{-5} )</td>
<td>( \pm 9.40 \times 10^{-5} )</td>
<td>( \pm 1.57 \times 10^{-4} )</td>
<td>( \pm 1.64 \times 10^{-4} )</td>
<td>( \pm 5.33 \times 10^{-4} )</td>
</tr>
<tr>
<td>300</td>
<td>( 4.798 \times 10^{-3} )</td>
<td>( 5.005 \times 10^{-3} )</td>
<td>( 1.071 \times 10^{-1} )</td>
<td>( 6.172 \times 10^{-3} )</td>
<td>( 1.056 \times 10^{-2} )</td>
</tr>
<tr>
<td></td>
<td>( \pm 5.99 \times 10^{-5} )</td>
<td>( \pm 6.08 \times 10^{-5} )</td>
<td>( \pm 1.41 \times 10^{-4} )</td>
<td>( \pm 7.45 \times 10^{-5} )</td>
<td>( \pm 1.92 \times 10^{-4} )</td>
</tr>
<tr>
<td>500</td>
<td>( 2.865 \times 10^{-3} )</td>
<td>( 2.966 \times 10^{-3} )</td>
<td>( 1.044 \times 10^{-1} )</td>
<td>( 3.672 \times 10^{-3} )</td>
<td>( 6.641 \times 10^{-3} )</td>
</tr>
<tr>
<td></td>
<td>( \pm 5.43 \times 10^{-5} )</td>
<td>( \pm 5.24 \times 10^{-5} )</td>
<td>( \pm 1.071 \times 10^{-4} )</td>
<td>( \pm 6.67 \times 10^{-5} )</td>
<td>( \pm 2.933 \times 10^{-4} )</td>
</tr>
<tr>
<td>700</td>
<td>( 2.091 \times 10^{-3} )</td>
<td>( 2.170 \times 10^{-3} )</td>
<td>( 1.033 \times 10^{-1} )</td>
<td>( 2.691 \times 10^{-3} )</td>
<td>( 4.802 \times 10^{-3} )</td>
</tr>
<tr>
<td></td>
<td>( \pm 2.39 \times 10^{-5} )</td>
<td>( \pm 1.90 \times 10^{-5} )</td>
<td>( \pm 1.53803 \times 10^{-4} )</td>
<td>( \pm 3.23 \times 10^{-5} )</td>
<td>( \pm 2.26 \times 10^{-4} )</td>
</tr>
<tr>
<td>1000</td>
<td>( 1.435 \times 10^{-3} )</td>
<td>( 1.507 \times 10^{-3} )</td>
<td>( 1.026 \times 10^{-1} )</td>
<td>( 1.859 \times 10^{-3} )</td>
<td>( 3.030 \times 10^{-3} )</td>
</tr>
<tr>
<td></td>
<td>( \pm 1.15 \times 10^{-5} )</td>
<td>( \pm 1.52 \times 10^{-5} )</td>
<td>( \pm 1.20 \times 10^{-4} )</td>
<td>( \pm 2.06 \times 10^{-5} )</td>
<td>( \pm 1.05 \times 10^{-4} )</td>
</tr>
</tbody>
</table>

Table EC.5  Point and confidence interval estimates of the ASEs - average square errors - of the estimators

Comparison of the efficiency of LES and QL\(_m\) in the \( D/M/s + M \) model with \( \rho = 1.4 \) and \( \alpha = 0.2 \)

<table>
<thead>
<tr>
<th>( s )</th>
<th>( \text{ASE}[\theta_{QL_m}] )</th>
<th>( \text{ASE}[\theta_{LES}] )</th>
<th>( \text{ASE}[\theta_{LES}]/\text{ASE}[\theta_{QL_m}] )</th>
<th>Predicted ratio by (38)</th>
<th>RE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>( 1.436 \times 10^{-2} )</td>
<td>( 1.863 \times 10^{-2} )</td>
<td>1.297</td>
<td>1.286</td>
<td>0.885</td>
</tr>
<tr>
<td></td>
<td>( \pm 9.78 \times 10^{-5} )</td>
<td>( \pm 1.642 \times 10^{-4} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>( 4.798 \times 10^{-3} )</td>
<td>( 6.172 \times 10^{-3} )</td>
<td>1.286</td>
<td>1.286</td>
<td>0.0421</td>
</tr>
<tr>
<td></td>
<td>( \pm 5.99 \times 10^{-5} )</td>
<td>( \pm 7.45 \times 10^{-5} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>( 2.865 \times 10^{-3} )</td>
<td>( 3.672 \times 10^{-3} )</td>
<td>1.281</td>
<td>1.286</td>
<td>-0.329</td>
</tr>
<tr>
<td></td>
<td>( \pm 5.43 \times 10^{-5} )</td>
<td>( \pm 6.67 \times 10^{-5} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>( 2.091 \times 10^{-3} )</td>
<td>( 2.691 \times 10^{-3} )</td>
<td>1.287</td>
<td>1.286</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>( \pm 2.39 \times 10^{-5} )</td>
<td>( \pm 3.23 \times 10^{-5} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>( 1.435 \times 10^{-3} )</td>
<td>( 1.859 \times 10^{-3} )</td>
<td>1.296</td>
<td>1.286</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>( \pm 1.15 \times 10^{-5} )</td>
<td>( \pm 2.06 \times 10^{-5} )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table EC.6

In this more lightly loaded setting, the ASE’s of all the estimators are relatively low, being smaller than for the \( M/M/s + M \) model with \( \alpha = 1.0 \) in Table EC.1 by a factor of about 4, despite having \( c_a^2 = 4.0 \) instead of \( c_a^2 = 1.0 \). However, the lighter loading makes the ED heavy-traffic approximations less appropriate.

The QL\(_r\) estimator is less efficient than QL\(_m\): \( \text{ASE}(\text{QL}_r)/\text{ASE}(\text{QL}_m) \) ranges from approximately
Comparison of the efficiency of NI and QL\textsubscript{m} in the $D/M/s+M$ model with $\rho = 1.4$ and $\alpha = 0.2$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$ ASE[\theta_{QL\textsubscript{m}}]$</th>
<th>$ ASE[\theta_{NI}]$</th>
<th>$ ASE[\theta_{NI}]/ASE[\theta_{QL\textsubscript{m}}]$</th>
<th>Predicted ratio by (28)</th>
<th>RE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$1.436 \times 10^{-2}$</td>
<td>$3.266 \times 10^{-2}$</td>
<td>2.275</td>
<td>2.25</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>$\pm 9.78 \times 10^{-5}$</td>
<td>$\pm 5.33 \times 10^{-4}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>$4.798 \times 10^{-3}$</td>
<td>$1.056 \times 10^{-2}$</td>
<td>2.201</td>
<td>2.25</td>
<td>-2.18</td>
</tr>
<tr>
<td></td>
<td>$\pm 5.99 \times 10^{-5}$</td>
<td>$\pm 1.92 \times 10^{-4}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>$2.865 \times 10^{-3}$</td>
<td>$6.641 \times 10^{-3}$</td>
<td>2.318</td>
<td>2.25</td>
<td>3.01</td>
</tr>
<tr>
<td></td>
<td>$\pm 5.43 \times 10^{-5}$</td>
<td>$\pm 2.933 \times 10^{-4}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>$2.091 \times 10^{-3}$</td>
<td>$4.802 \times 10^{-3}$</td>
<td>2.297</td>
<td>2.25</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.39 \times 10^{-5}$</td>
<td>$\pm 2.26 \times 10^{-4}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>$1.435 \times 10^{-3}$</td>
<td>$3.130 \times 10^{-3}$</td>
<td>2.111</td>
<td>2.25</td>
<td>-3.08</td>
</tr>
<tr>
<td></td>
<td>$\pm 1.15 \times 10^{-5}$</td>
<td>$\pm 1.05 \times 10^{-4}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table EC.8  
Point and confidence interval estimates of the ASES - average square errors - of the estimators

Efficiency of the estimators in the $H_2/M/s+M$ model with $\rho = 1.4$ and $\alpha = 5.0$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$ ASE[\theta_{QL\textsubscript{m}}]$</th>
<th>$ ASE[\theta_{QL}]/ASE[\theta_{LES}]/ASE[\theta_{NI}]$</th>
<th>Predicted ratio by (28)</th>
<th>RE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$7.193 \times 10^{-4}$</td>
<td>$1.059 \times 10^{-4}$/2.217/2.393/3.101</td>
<td>3.01</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.63 \times 10^{-6}$</td>
<td>$\pm 4.47 \times 10^{-6}/\pm 1.01 \times 10^{-5}/\pm 6.72 \times 10^{-6}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>$2.008 \times 10^{-4}$</td>
<td>$2.675 \times 10^{-4}/7.240/7.569/1.169$</td>
<td>1.69</td>
<td>5.82</td>
</tr>
<tr>
<td></td>
<td>$\pm 7.85 \times 10^{-7}$</td>
<td>$\pm 1.28 \times 10^{-6}/\pm 2.63 \times 10^{-6}/\pm 2.70 \times 10^{-6}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>$1.167 \times 10^{-4}$</td>
<td>$1.495 \times 10^{-4}/4.792/4.540/7.624$</td>
<td>7.62</td>
<td>6.07</td>
</tr>
<tr>
<td></td>
<td>$\pm 7.05 \times 10^{-7}$</td>
<td>$\pm 8.78 \times 10^{-7}/\pm 2.68 \times 10^{-6}/\pm 1.71 \times 10^{-6}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>$8.277 \times 10^{-5}$</td>
<td>$1.042 \times 10^{-4}/3.856/3.280/5.714$</td>
<td>5.71</td>
<td>4.72</td>
</tr>
<tr>
<td></td>
<td>$\pm 4.12 \times 10^{-7}$</td>
<td>$\pm 6.52 \times 10^{-7}/\pm 2.50 \times 10^{-6}/\pm 1.27 \times 10^{-6}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>$5.733 \times 10^{-5}$</td>
<td>$7.141 \times 10^{-5}/3.184/2.302/4.095$</td>
<td>4.09</td>
<td>4.15</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.48 \times 10^{-7}$</td>
<td>$\pm 2.44 \times 10^{-7}/\pm 1.34 \times 10^{-6}/\pm 1.19 \times 10^{-6}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table EC.8  
Point and confidence interval estimates of the ASES - average square errors - of the estimators

1.5 when $s = 100$ to approximately 1.25 when $s = 1000$. The LES estimator is less accurate, with an RRASE ranging from approximately 14% when $s = 100$ to approximately 4% when $s = 1000$. The QL estimator performs slightly worse than LES: The ratio $\text{ASE}(QL)/\text{ASE}(QL_m)$ ranges from about 3 when $s = 100$ to about 5 when $s = 1000$. The NI estimator is the least efficient estimator for this model.
Comparison of the efficiency of LES and QLm in the $H_2/M/s + M$ model with $\rho = 1.4$ and $\alpha = 5.0$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$\text{ASE}[\theta_{QLm}]$</th>
<th>$\text{ASE}[\theta_{LES}]$</th>
<th>$\text{ASE}[\theta_{LES}]/\text{ASE}[\theta_{QLm}]$</th>
<th>Predicted by (39)</th>
<th>RE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$7.193 \times 10^{-4}$</td>
<td>$2.393 \times 10^{-3}$</td>
<td>3.326</td>
<td>4.143</td>
<td>-19.7</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.63 \times 10^{-6}$</td>
<td>$\pm 6.72 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>$2.008 \times 10^{-4}$</td>
<td>$7.569 \times 10^{-4}$</td>
<td>3.769</td>
<td>4.143</td>
<td>-9.03</td>
</tr>
<tr>
<td></td>
<td>$\pm 7.85 \times 10^{-7}$</td>
<td>$\pm 2.70 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>$1.167 \times 10^{-4}$</td>
<td>$4.540 \times 10^{-4}$</td>
<td>3.891</td>
<td>4.143</td>
<td>-6.09</td>
</tr>
<tr>
<td></td>
<td>$\pm 7.05 \times 10^{-7}$</td>
<td>$\pm 1.71 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>$8.277 \times 10^{-5}$</td>
<td>$3.280 \times 10^{-4}$</td>
<td>3.962</td>
<td>4.143</td>
<td>-4.36</td>
</tr>
<tr>
<td></td>
<td>$\pm 4.12 \times 10^{-7}$</td>
<td>$\pm 1.27 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>$5.733 \times 10^{-5}$</td>
<td>$2.302 \times 10^{-4}$</td>
<td>4.014</td>
<td>4.143</td>
<td>-3.10</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.48 \times 10^{-7}$</td>
<td>$\pm 1.19 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table EC.9

Comparison of the efficiency of NI and QLm in the $H_2/M/s + M$ model with $\rho = 1.4$ and $\alpha = 5.0$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$\text{ASE}[\theta_{QLm}]$</th>
<th>$\text{ASE}[\theta_{NI}]$</th>
<th>$\text{ASE}[\theta_{NI}]/\text{ASE}[\theta_{QLm}]$</th>
<th>Predicted ratio by (28)</th>
<th>RE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>$7.193 \times 10^{-4}$</td>
<td>$3.101 \times 10^{-3}$</td>
<td>4.310</td>
<td>7.25</td>
<td>-40.5</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.63 \times 10^{-6}$</td>
<td>$\pm 1.42 \times 10^{-5}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>$2.008 \times 10^{-4}$</td>
<td>$1.169 \times 10^{-3}$</td>
<td>5.821</td>
<td>7.25</td>
<td>-19.7</td>
</tr>
<tr>
<td></td>
<td>$\pm 7.85 \times 10^{-7}$</td>
<td>$\pm 5.82 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>$1.167 \times 10^{-4}$</td>
<td>$7.624 \times 10^{-4}$</td>
<td>6.533</td>
<td>7.25</td>
<td>-9.89</td>
</tr>
<tr>
<td></td>
<td>$\pm 7.05 \times 10^{-7}$</td>
<td>$\pm 6.07 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>$8.277 \times 10^{-5}$</td>
<td>$5.714 \times 10^{-4}$</td>
<td>6.904</td>
<td>7.25</td>
<td>-4.78</td>
</tr>
<tr>
<td></td>
<td>$\pm 4.12 \times 10^{-7}$</td>
<td>$\pm 4.72 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>$5.733 \times 10^{-5}$</td>
<td>$4.0951 \times 10^{-4}$</td>
<td>7.143</td>
<td>7.25</td>
<td>-1.48</td>
</tr>
<tr>
<td></td>
<td>$\pm 2.48 \times 10^{-7}$</td>
<td>$\pm 4.15 \times 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table EC.10

Tables EC.9 and EC.10 substantiate (40) and (29) of §5, that compare the performances of QLm, LES and NI in the $H_2/M/s + M$ model. Consistent with (40), Table EC.9 shows that the performance of LES is significantly worse than that of QLm, when the arrival process is highly variable. The simulation estimates of $\text{ASE}(\text{LES})/\text{ASE}(\text{QLm})$, for alternative values of $s$, are close to the numerical value, approximately 4.143, predicted by (40), especially for large values of $s$; the
RE observed ranges from approximately $-20\%$ for $s=100$ to approximately $-3\%$ when $s=1000$. We observe a relatively poor performance of the approximation in (40) when the number of servers is small. That is understandable because the system is not very heavily loaded when $\alpha = 5.0$. Consistent with (29), Table EC.10 shows that the performance of NI is much worse than that of $\text{QL}_m$, when the arrival process is highly variable. The approximation in (29) performs poorly when $s=100$ (RE $\approx -40\%$) but becomes remarkably accurate when $s=1000$ (RE $\approx -1.5\%$).

**References**
