Submitted to *Management Science*
manuscript MS-0001-2008.xxx

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; many-server queues; call centers; 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. These delay estimates may be used to make delay announcements to arriving customers. Delay announcements can be helpful in call centers, where queues are invisible, so that customers are unable to estimate their own delay; 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 a wide range of 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, developed in Whitt (2005, 2006).

1.1. The Delay Estimation Framework

We study the performance of alternative real-time delay estimators by considering the steady-state behavior of an overloaded $GI/GI/s+GI$ queueing model, allowing customer abandonment. This model has independent and identically distributed (i.i.d.) interarrival times with mean λ^{-1} and a general distribution. We only use the i.i.d. assumption for the interarrival times when simulating the model; it is not required for the implementation of our delay estimators. Service times are i.i.d. with mean μ^{-1} and a general distribution. Associated with each arriving customer is a random variable quantifying this customer's patience. Abandonment times are i.i.d. with mean α^{-1} and a general distribution. The arrival, service and abandonment processes are all assumed to be mutually independent. This model has unlimited waiting space. Arriving customers are served in order of arrival; i.e., we use the first-come-first-served (FCFS) service discipline. The traffic intensity, ρ, is given by $\rho \equiv \lambda/s\mu$.

We focus on the overloaded scenario, in which the arrival rate to the system exceeds the maximum total service rate in the system. Customer abandonment makes the system stable in this case. 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.

To each delayed customer, upon arrival, we give a single-number delay estimate of that customer's delay until he starts service. In this work, we assume that these delay estimates have no impact on customer behavior. For other work that does focus on customer response, see Armony et al. (2008). Even though we do not directly model customer response, our estimators indirectly account for it, because they depend on the system state or history, which in turn is affected by customer response.

1.2. Actual and Potential Waiting Times

As in 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 (his patience is 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. In this study, we estimate the potential waiting times of delayed customers.

1.3. 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 have the advantage of not relying on any system-state or model information: Their implementation only exploits information about recent customer delay history in the system. Delay-history-based estimators are especially useful when the queue length is not directly observable. That is nicely illustrated by the ticket queues studied by Xu et al. (2007). Upon arriving at a ticket queue, each customer is issued a numbered ticket. The number currently being served is displayed. The queue length is not known

to ticket-holding customers or even to system managers, because they do not observe customer abandonments.

As discussed in Whitt (1999a), 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 delay-history-based estimators. Our simulation results in §6, suggest that this is usually, but not always, true. Nevertheless, the delay-history-based estimators are quite effective.

1.4. 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:

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

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

1.5. 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 the new arrival, 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:

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

Note that the MSE of a queue-length-based delay estimator is a function of n, the number of customers seen in queue upon arrival, assuming steady-state conditions. 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. By looking at the ASE, we are looking at the expected MSE averaging over all n, where the arrival must wait, in steady state.

1.6. Root Relative Squared Error

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

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

using the same notation as in (1). The denominator in (2) is the average potential waiting time of customers who must wait. For large samples, the RRASE should agree with the expected *root relative mean squared error* (RRMSE), in steady state. The RRASE (RRMSE) is useful because it measures the effectiveness of an estimator relative to the mean steady-state potential waiting time, given that the customer must wait. It is thus easy to interpret.

1.7. Related Literature

This paper is an extension of Ibrahim and Whitt (2007a), 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.

The body of literature related directly to delay estimations in service systems is large and growing. Jouini et al. (2007) investigate delay estimation in multi-class Markovian models with and without reneging. 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. Nakibly (2002) analyzes methods to estimate delays, both exactly and approximately, in queueing models with two servers, two service types, and a priority discipline,

assuming exponential service times and ignoring customer abandonment. Guo (2007) studies the effect of giving delay information on the performance of the system, by considering first the single server all-exponential queue with no abandonment, and then extending his analysis to phase-type service times. Armony and Maglaras (2003) analyze a system that offers two modes of service: real-time and postponed with a delay guarantee. They propose a delay estimation scheme that is asymptotically correct based on multiserver limits.

There is also a stream of literature on the psychology of waiting. For example, see Maister (1984), Hui and Tse (1996), Taylor (1994), and references therein.

1.8. Organization of the paper

The rest of this paper is organized as follows: In §2, we describe the 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 (2007a). 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 in §7. We postpone one long proof of a result in §5 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

Since we will be considering the overloaded $G/GI/s+GI$ model, an important theoretical reference point for this work is the literature on many-server heavy-traffic limits for the Markovian $M/M/s+M$ queue with customer abandonment, sometimes called the Palm model or Erlang-A model, in the efficiency-driven (ED) regime, as discussed in Garnett et al. (2002), Whitt (2004) and Talreja and Whitt (2008). This theory describes how the model behaves as the arrival rate λ and number of servers s increase, while the individual service rate μ and individual abandonment rate α remain unchanged, with the traffic intensity $\rho \equiv \lambda/s\mu$ held fixed at a value $\rho > 1$. (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

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

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

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

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

The fluid approximation suggests that we might simply use the deterministic fluid approximation w itself, or the steady-state mean $E[W(\infty)]$ it approximates, as a *no-information* (NI) estimator, θ_{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 for our model with customer abandonment than it was for the $GI/M/s$ model considered previously. Based on the limits above (plus appropriate uniform integrability, which can also be established), we have

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

Unlike in the $GI/M/s$ model, here the squared coefficient of variation (SCV, variance divided by the square of the mean), c^2_{NI}, is asymptotically negligible as well, because here $E[W_s(\infty)] \to w > 0$ as $s \to \infty$. (For the $GI/M/s$ model considered in Ibrahim and Whitt (2007a), $c^2_{NI} \to 1$ as $\rho \uparrow 1$ for all s.) The asymptotic behavior of $MSE(\theta_{NI})$ here means that any reasonable estimator

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

Table 1 Summary of the information required for the implementation of each queue-length delay 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. Even though the NI estimator has reasonable MSE, it has a defect: By not depending upon system state or history, it cannot react to model deficiencies, such as customer balking in response to delay announcements.

3. Queue-Length-Based Delay Estimators

In this section, we describe alternative estimators based on knowing the queue length 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 m minutes, we may predict that a customer, who finds n customers in queue upon arrival, will be able to begin service in $(n+1)m/s$ minutes. Let QL refer to this simple queue-length-based estimator, commonly used in practice. Let the estimator, as a function of n, be:

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

The QL estimator is appealing due to its simplicity and its ease of implementation: It uses information about the system that usually is readily available. In Ibrahim and Whitt (2007a), the performance of QL is studied in the $GI/M/s$ model, where there is an exponential service-time distribution and no customer abandonment. For this 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 μ^{-1}, which makes them i.i.d. exponential with mean $1/s\mu$. The optimal delay estimator, using the MSE criterion, is the one announcing the conditional mean, $E[W_Q(n)]$. But, following the analysis above, $E[W_Q(n)] = \theta_{QL}(n)$ in (6). Hence, the optimality of QL in the $GI/M/s$ model, under the MSE criterion, is mathematically demonstrated. Extensive simulation experiments in Ibrahim and Whitt (2007a) support this result, and show the superiority of QL in that simple idealized setting.

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

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

In the fluid limit,

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

Consistent with intuition, we see that QL overestimates w. Exploiting the asymptotic expansion of the logarithm: $\ln(1+\delta) \approx \delta - \delta^2/2$ when δ is small, we can quantify the approximate relative error resulting from the QL estimation. Indeed,

$$\frac{\theta_{QL}(q) - w}{w} = \frac{(\rho-1)/\alpha - \ln(\rho)/\alpha}{\ln(\rho)/\alpha} \approx \frac{(\rho-1) - (\rho-1) + (\rho-1)^2/2}{(\rho-1) - (\rho-1)^2/2} = \frac{(\rho-1)/2}{1-(\rho-1)/2};$$

e.g., there is 11% relative error when $\rho = 1.2$, and 25% relative error when $\rho = 1.4$, and of course much greater error when ρ is larger.

3.2. The Markovian Queue-Length-Based Delay Estimator ($\mathbf{QL}_m$)

As in Whitt (1999a), 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:

$$W_Q(n) \equiv \sum_{i=0}^{n} Y_i \ , \tag{8}$$

where the Y_i are independent random variables with Y_i being the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service) and i exponential random variables with rate α (corresponding to the abandonment times of the remaining customers waiting in line). That is, Y_i is exponential with rate $s\mu + i\alpha$. Therefore,

$$E[W_Q(n)] = \sum_{i=0}^{n} E[Y_i] = \sum_{i=0}^{n} \frac{1}{s\mu + i\alpha} \ .$$

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.

3.3. The Simple-Refined Queue-Length-Based Delay Estimator ($\mathbf{QL}_r$)

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. 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" $\equiv \lambda F^c(w) = s\mu \equiv$ "rate out", we have:

$$\rho F^c(w) = 1 \ . \tag{9}$$

The associated equation for q is

$$q = \lambda \int_0^w F^c(x)dx = s\rho\mu \int_0^w F^c(x)dx \ . \tag{10}$$

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

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

The QL_r estimator is appealing because it is a minor modification of the QL estimator, but performs much better in models with customer abandonment, as we show in §6. Note that in addition to s, n and μ, we need to know ρ or, equivalently, λ, and the abandonment-time distribution in order to implement QL_r. With i.i.d. exponential abandonment times, we propose an adjustment to QL_r which does not depend on ρ, as we show next.

3.4. The Exponential Abandonment Case (QL_r^m)

We now consider QL_r in the overloaded $G/GI/s+M$ queueing model. Using the corresponding values of w and q for that model, given respectively by (3) and (7), yields a ratio $\beta = \ln(\rho)/(\rho-1)$. From (7), we have that $\rho = 1 + \alpha q/s\mu$, which when plugged in yields

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

For the $G/GI/s+M$ model, we propose a simple adjustment to QL_r, denoted by QL_r^m (where m stands for Markovian), which does not depend on ρ. 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, μ, and α, but not of ρ or, equivalently, λ. It approximates the abandonment-time distribution by the exponential distribution. We will see that QL_r^m performs nearly the same as QL_m, which is good when the abandonment is nearly exponential, but not necessarily otherwise.

3.5. The Approximation-Based Queue-Length Delay Estimator ($\mathbf{QL}_{ap}$)

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

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

$$\alpha_j \equiv h(j/\lambda), \quad 1 \le j \le k \ ; \tag{11}$$

k is the current queue length, and h is the abandonment-time hazard-rate function, defined as $h(t) \equiv f(t)/F^c(t)$, $t \ge 0$, where $F^c(t) = 1 - F(t)$ is the complementary cdf (ccdf) associated with F, and f is the corresponding density function (assumed to exist). Having α_j depend on h instead of F is convenient, because it is natural to estimate F via h; e.g., see Brown et al. (2005). From (11), we see that the estimator QL_{ap} depends on the abandonment distribution having a relatively smooth density. We assume that is the case.

We now explain the derivation of (11). If we knew that a given customer had been waiting for time t, then the rate of abandonment for that customer, at that time, would be $h(t)$. The goal is to produce, as an approximation, abandonment rates that depend on a customer's position in queue, and on the length of that queue. We therefore need to estimate the elapsed waiting time of that customer, given the available state information. To that end, assume that the queue length at an arbitrary time is k, and consider the customer, C_j, who is jth from the end of the line, $1 \le j \le k$. If there were no abandonments, then there would have been exactly $j-1$ arrival events

since C_j arrived. Assuming that abandonments are relatively rare compared to service completions, a reasonable estimate is that there have been j arrival events since C_j arrived. 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/λ and his abandonment rate by (11). 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:

$$W_Q(n) \approx \sum_{i=0}^{n} X_i \ , \tag{12}$$

where X_{n-i} is the time between the ith and $(i+1)$st departure events. There is no difficulty for the first departure: X_n is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and n exponential random variables with rates α_j, $1 \leq j \leq n$, (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 (12) by exponential random variables. Let X_{n-l}, which is the time between

the lth and $(l+1)$st departure events, have an exponential distribution with rate $s\mu+\delta_n-\delta_l$. This is appropriate because it is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and $n-l$ exponential random variables with rates α_i, $l+1 \le i \le n$ (corresponding to the abandonment times of the customers waiting in line).

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

$$\theta_{QL_{ap}}(n) = \sum_{i=0}^{n} \frac{1}{s\mu+\delta_n-\delta_{n-i}} \ .$$

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_r^m 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 λ, which requires some degree of stationarity.

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 (2007a).

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

As in Armony et al. (2006), 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: $\theta_{LES}(w) \equiv w$.

The LES estimator is appealing because it does not depend on the model and uses very little information about the system. It is robust because it responds automatically to changes in system parameters (e.g., number of servers, mean service time, and arrival rate). 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 they complete service. That might be the case for customers and outside observers. Under some circumstances, the LCS and LES estimators will be similar, but they actually can be 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 λ. Let $\bar{A}_s(t) = A_s(t)/s \equiv A(st)/s$ for $t \geq 0$. Let $D \equiv D([0,\infty),\mathbb{R})$ be the function space of all right continuous real-valued functions with left limits, endowed with the usual Skorohod (J_1) topology; see Billingsley (1999) or Whitt (2002). We assume that A satisfies a functional weak law of large numbers (FWLLN) and a FCLT refinement:

$$\bar{A}_s(t) \Rightarrow \lambda t \quad \text{in} \quad D \quad \text{and} \quad \sqrt{s}(\bar{A}_s(t) - \lambda t) \Rightarrow \sqrt{\lambda c_a^2} B(t) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \tag{13}$$

where B is a standard Brownian motion. That condition will be satisfied if A is a renewal process with an interarrival-time distribution having finite first and second moments. As usual, the arrival process affects the limits for the other random quantities (the estimators) only via the two normalization constants λ and c_a^2. When A is a renewal counting process, c_a^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of an interarrival time. For a Poisson arrival process, $c_a^2 = 1$; for a deterministic arrival process, $c_a^2 = 0$.

We start by considering the Markovian estimator QL_m, which is the best possible estimator for the $G/M/s+M$ model. 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 (8).

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,*

$$E[W_{Q,s}(\lfloor st \rfloor)] \to c(t), \quad sVar(W_{Q,s}(\lfloor st \rfloor) \to d(t) \tag{14}$$

and

$$\hat{W}_{Q,s}(t) \equiv \sqrt{s}\,(W_{Q,s}(\lfloor st \rfloor) - c(t)) \Rightarrow B(d(t)) \quad \textit{in} \quad D \quad \textit{as} \quad s \to \infty, \tag{15}$$

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

$$c(t) \equiv \frac{1}{\alpha} \ln\left(1 + \frac{\alpha t}{\mu}\right) \quad \textit{and} \quad d(t) \equiv \frac{t}{\mu(\mu + \alpha t)}. \tag{16}$$

As a consequence of the stochastic-process limit in (15), we obtain the one-dimensional limit

$$\sqrt{s}\,(W_{Q,s}(\lfloor st \rfloor) - c(t)) \Rightarrow N(0, d(t)) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad s \to \infty \quad \text{for each} \quad t. \tag{17}$$

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

$$\bar{Q}_s(\infty) \equiv \frac{Q_s(\infty)}{s} \Rightarrow q \equiv \frac{\lambda - \mu}{\alpha} \quad \text{as} \quad s \to \infty, \tag{18}$$

from Theorem 2.3 of Whitt (2004) or Theorem 6.1 of Talreja and Whitt (2008).

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

$$sMSE(\theta_{QL_{m,s}}(\lfloor st \rfloor)) \equiv sVar(W_{Q,s}(\lfloor st \rfloor)) \to d(t) \quad \textit{as} \quad s \to \infty \tag{19}$$

for each $t > 0$, where $d(t)$ is given in (16) *and*

$$sMSE(\theta_{QL_{m,s}}(Q_s(\infty))) \equiv sVar(W_{Q,s}(Q_s(\infty)) \Rightarrow d(q) \equiv \frac{q}{\lambda\mu} \equiv \frac{\lambda - \mu}{\lambda\mu\alpha} \quad \textit{as} \quad s \to \infty. \tag{20}$$

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 (which is equivalent to convergence in probability, since the limit is deterministic) in (20), i.e.,

$$sE[MSE(\theta_{QL_{m,s}}(Q_s(\infty)))] \to d(q) \quad \text{as} \quad s \to \infty. \tag{21}$$

From either (20) or (21), we get the approximation

$$E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))] \approx \frac{\lambda - \mu}{s\lambda\mu\alpha}. \tag{22}$$

Note that the FCLT normalization constant c_a^2 does not appear in (20)–(22). 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,*

$$\hat{Q}_s(\infty) \equiv \sqrt{s}(\bar{Q}_s(\infty) - q) \Rightarrow N\left(0, \frac{\lambda(c_a^2+1)}{2\alpha}\right) \quad as \quad s \to \infty \tag{23}$$

and

$$\hat{W}_s(\infty) \equiv \sqrt{s}(W_s(\infty) - w) \Rightarrow N\left(0, \sigma_w^2\right) \quad as \quad s \to \infty, \tag{24}$$

where $\sigma_w^2 \equiv 1/\alpha\mu + (c_a^2-1)/2\lambda\alpha$, with w in (3) and q in (18).

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. (We henceforth assume that uniform integrability holds whenever it is needed.) 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,*

$$sMSE(\theta_{NI,s}) \equiv sVar(W_s(\infty)) \to \frac{1}{\alpha\mu} + \frac{c_a^2-1}{2\lambda\alpha} \quad as \quad s \to \infty. \tag{25}$$

Combining the limits in (20) and (25), we obtain the following

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

$$\frac{MSE(\theta_{NI,s})}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))]} \to \frac{2\lambda+\mu(c_a^2-1)}{2(\lambda-\mu)} > 1 \quad as \quad s\to\infty. \tag{26}$$

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 $W_{QL_{m,s}}(n)$ in (8) and a net-input process $N_s \equiv \{N_s(t) : t \geq 0\}$ over the interval $[0, w]$, i.e.,

$$W_{LES,s}(w) \approx W_{Q,s}(N_s(w)) \equiv \sum_{i=0}^{N_s(w)} X_{s,i}, \tag{27}$$

where $N_s(w)$ counts the number of arrivals in the interval $[0, w]$ who do not abandon, in system s. Formula (27) is not an exact relation because it does not account for the state change since the last customer entered service, but that change is clearly asymptotically negligible in the ED many-server limiting regime.

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

$$E[N_s(t)] = \frac{s\lambda}{\alpha}\left(1-e^{-\alpha t}\right), \quad t \geq 0. \tag{28}$$

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); see Krichagina and Puhalskii (1997) for an extension.

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

$$\bar{N}_s(t) \equiv \frac{N_s(t)}{s} \Rightarrow a(t) \equiv \frac{\lambda}{\alpha}\left(1-e^{-\alpha t}\right) \quad in \quad D \quad as \quad s\to\infty \tag{29}$$

and

$$\hat{N}_s(t) \equiv \sqrt{s}(\bar{N}_s(t) - a(t)) \Rightarrow \hat{G}(t) \quad in \quad D \quad as \quad s\to\infty, \tag{30}$$

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

$$\hat{G}(t) \stackrel{\mathrm{d}}{=} N(0, \sigma_n^2(t)) \quad \textit{where} \quad \sigma_n^2(t) \equiv a(t) + \frac{\lambda(c_a^2 - 1)}{2\alpha}\left(1 - e^{-2\alpha t}\right), \tag{31}$$

for $a(t)$ *defined in* (29) *and* c_a^2 *in* (13).

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 (30) and for $\hat{W}_s(\infty)$ in (24) 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 (13). 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* $s \to \infty$,

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

$$\hat{W}_{LES,s}(W_s(\infty)) \equiv \sqrt{s}\,(W_{LES,s}(W_s(\infty)) - W_s(\infty)) \Rightarrow N(0, \sigma_{LES}^2), \tag{32}$$

$$sE[MSE(\theta_{LES,s}(W_s(\infty))] \to \sigma_{LES}^2, \tag{33}$$

where

$$\sigma_{LES}^2 \equiv d(a(w)) + \frac{\sigma_n^2(w)}{\lambda^2} + \left(\frac{\lambda - \mu}{\lambda}\right)^2 \sigma_w^2 = 2d(q) + \frac{(c_a^2 - 1)(\lambda - \mu)}{\alpha\lambda^2}, \tag{34}$$

for σ_w^2 *in Theorem 2,* $\sigma_n^2(t)$ *in* (31), $a(w) = q$ *and* $d(q) = q/\lambda\mu$.

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

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty))]}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))]} \to 2 \quad \textit{as} \quad s \to \infty. \tag{35}$$

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

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty))]}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))]} \to (2-\rho^{-1}) > 1 \quad as \quad s \to \infty. \tag{36}$$

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

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty))]}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))]} \to r(LES, QL_m) \quad as \quad s \to \infty, \tag{37}$$

where

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

From (36), 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,*

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty))]}{MSE(\theta_{NI,s})} \to \frac{2(\rho-1)}{\rho}, \tag{38}$$

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) *For the $M/M/s+GI$ model in the ED regime, $W_s(\infty) \Rightarrow w$ for w in (9) and*

$$\sqrt{s}(W_s(\infty) - w) \Rightarrow N(0, 1/\lambda f(w)) \quad as \quad s \to \infty, \tag{39}$$

where f is the abandonment-time probability density function.

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), H_2 (hyperexponential with SCV equal to 4 and balanced means) and E_{10} (Erlang, sum of 10 exponentials) distributions. We use a Poisson arrival process because it is usually a good model. We briefly discuss other models in §6.5.

6.1. Description of the Experiments

We vary the number of servers, s, but consider only relatively large values ($s \geq 100$), because we are interested in large service systems. We let the service rate, μ, be equal to 1. We do this without loss of generality, 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, α, be equal to 1 because that seems to be a representative value. We also consider $\alpha = 0.2$ and $\alpha = 5.0$ in the e-companion. We vary λ to get a fixed value of ρ, for alternative values of s. We let $\rho = 1.4$ in all models. This value is chosen to let our systems be significantly overloaded. Because of abandonment, the congestion is not extraordinarily high. For example, with $s = 100$ servers and exponential abandonments, the mean queue length is about $q \approx (\rho - 1)s/\alpha \approx 40$, 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 simulations are steady-state simulations. Therefore, we could potentially encounter estimation error caused by the classical problem of the initial transient, i.e., when the system is not started in steady state. A possible solution is to delete an initial segment of the data, i.e., to have a warmup period which we later discard. We determine the length of this warmup period (roughly) by computing the relative errors that we get for different period lengths: We consider an error of less than 5% to be negligible.

Simulation results for the $M/M/s+M$ and $M/M/s+H_2$ models are based on 10 independent replications of 5 million events each, where an event is either a service completion, an arrival event

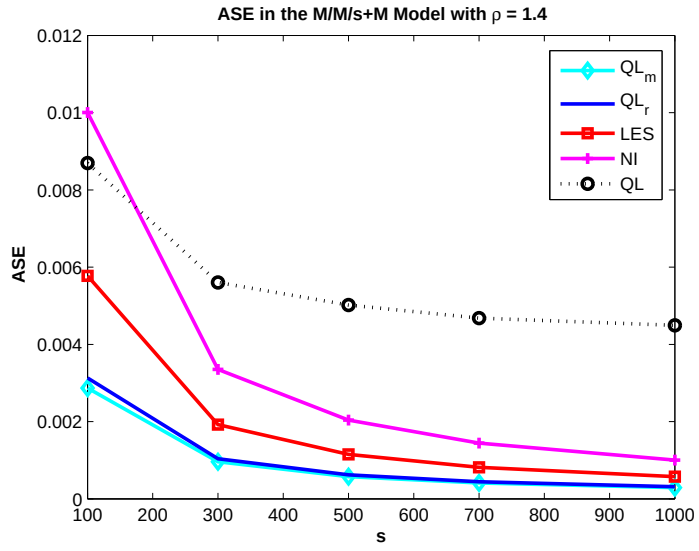

Figure 1

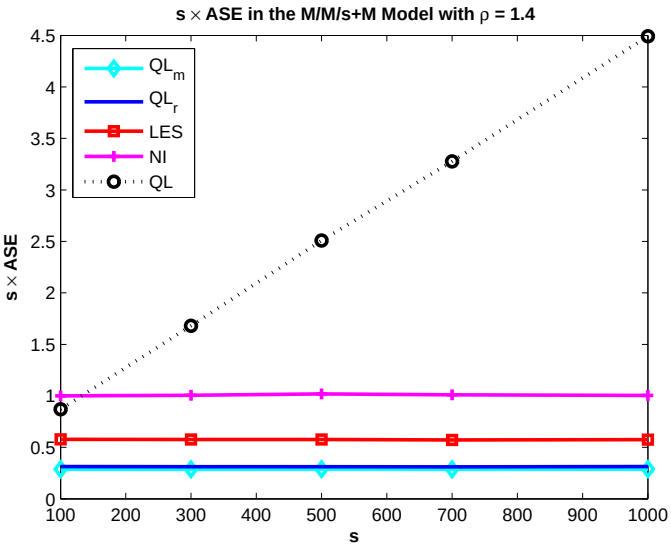

Figure 2

or an abandonment from the system. The effect of the initial transient period is negligible in these models so we do not include any warmup period. Simulation results for the $M/M/s+E_{10}$ model are based on 10 independent replications of 6 million events each, with an initial transient period of 1 million events. In this section we show plots of the simulation results. Corresponding tables with estimates of the 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 possible 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 $s\times\mathrm{ASE}(\mathrm{QL}_m)$, the ASE of QL_m multiplied by the number of servers s, is nearly constant for all values of s considered. In particular, Figure 2 shows that $s\times\mathrm{ASE}(\mathrm{QL}_m)\approx(\lambda-\mu)/(\lambda\mu\alpha)$, as in equation (22) of §5. The relative error between the simulation estimates for $\mathrm{ASE}(\mathrm{QL}_m)$ and the numerical value given by equation (22) is less than 1% throughout.

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

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

On the other hand, we have that

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

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

The LES estimator performs worse than QL_m and QL_r. The ratio ASE(LES)/ASE(QL_m) is close to 2 for all values of s, which provides support to equation (35). This is consistent with the results in Ibrahim and Whitt (2007a) for the $GI/M/s$ model, without customer abandonment. Figure 2 shows that $s\times$ASE(LES) $\approx \sigma^2_{LES}$, consistent with (33). Indeed, the relative error between the simulation estimates and the numerical value given by (34) 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 equation (38) is less than 2% throughout. Figure 2 shows that $s\times$ASE(NI) $\approx 1/\alpha\mu$, as in equation (25), with $c_a^2 = 1$. The relative error between the simulation estimates for ASE(NI) and the numerical value given by equation (25) 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 $s\times$ASE(QL) is monotone increasing in s. This shows the need to go beyond QL when customer abandonment is included.

6.3. Results for the $M/M/s + H_2$ model

Figure 3 shows that the best delay estimator for this model is QL_{ap}. The corresponding RRASE ranges from about 20% for $s = 100$ to about 6% when $s = 1000$. Once more, we see that the

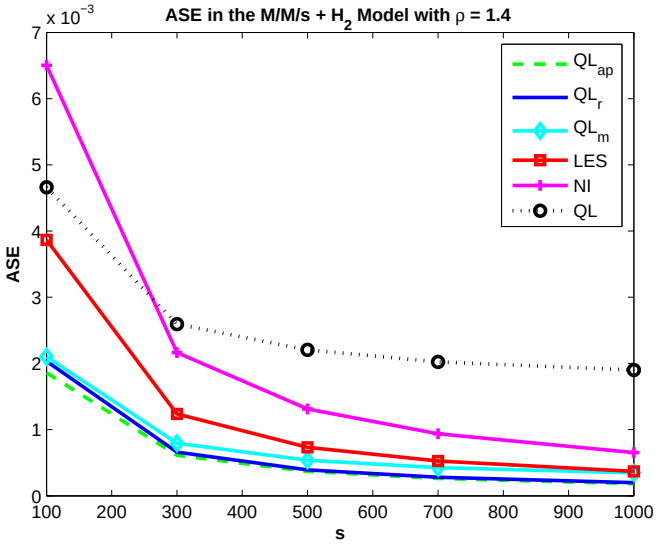

Figure 3

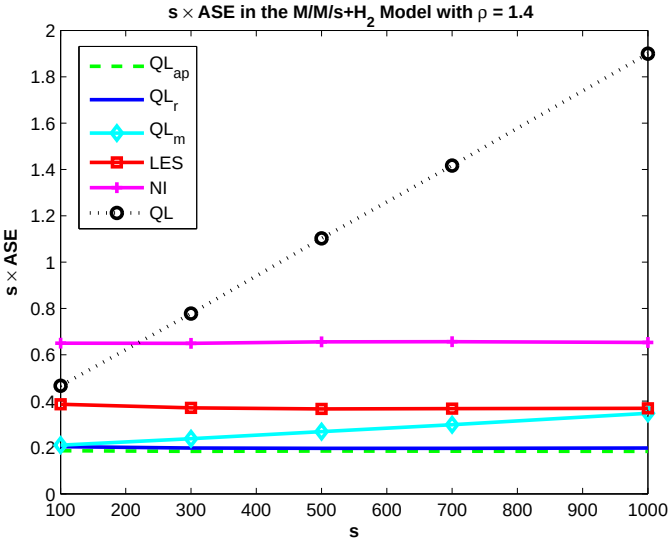

Figure 4

accuracy of this estimator improves as the number of servers increases. The remaining estimators, too, are relatively more accurate for a larger number of servers. For example, the RRASE of the LES estimator ranges from about 30% when $s = 100$ to about 10% when $s = 1000$. The RRASE for QL_m ranges from about 20% when $s = 100$ to about 8% when $s = 1000$.

The QL_r estimator performs nearly the same as QL_{ap}, and is only slightly outperformed. The ratio $\mathrm{ASE}(\mathrm{QL}_r)/\mathrm{ASE}(\mathrm{QL}_{ap})$ is close to 1 for all values of s. The QL_m estimator performs well but it is now slightly outperformed by QL_r. The two are nearly the same when $s = 100$; the ratio $\mathrm{ASE}(\mathrm{QL}_m)/\mathrm{ASE}(\mathrm{QL}_r)$ is close to 1 when $s = 100$ but closer to 2 when $s = 1000$.

The LES estimator performs worse than QL_{ap}, QL_m, and QL_r when $s = 100$ but nearly the same as QL_m when $s = 1000$. The ratio $\mathrm{ASE}(\mathrm{LES})/\mathrm{ASE}(\mathrm{QL}_{ap})$ is close to 2 for all values of s, suggesting that our analytical results of §5 should extend to general abandonment-time distributions.

The NI estimator performs worse than LES but not as bad as QL. The ratio $\mathrm{ASE}(\mathrm{NI})/\mathrm{ASE}(\mathrm{QL}_{ap})$ is close to 3.5 for all values of s considered. As above, the efficiency of QL is degrading as the number of servers increases. The ratio $\mathrm{ASE}(\mathrm{QL})/\mathrm{ASE}(\mathrm{QL}_{ap})$ ranges from about 2 when $s = 100$ to about 10 when $s = 1000$. Once more, the need to go beyond QL is evident.

Consistent with §5, Figure 4 shows that all estimators, except QL and QL_m, have an ASE which is inversely proportional to the number of servers, but mathematical support for the estimators

(besides NI) has yet to be provided, with non-exponential abandonment distributions. Beyond Theorems 2 and 5, the NI behavior is consistent with conjectured stochastic refinements to the fluid limits in Whitt (2006).

6.4. Results for the $M/M/s+E_{10}$ model

Figure 5 shows that QL_{ap} is the best possible delay estimator, for this model, except when s is very large (e.g., $s=700$ or $s=1000$). The corresponding RRASE ranges from about 10% when $s=100$ to about 3% when $s=1000$. The QL_r estimator performs worse than QL_{ap} for smaller values of s, but slightly outperforms QL_{ap} for larger values of s. The ratio $\mathrm{ASE}(\mathrm{QL}_r)/\mathrm{ASE}(\mathrm{QL}_{ap})$ 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_{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 $s\times$ ASE(LES) equals a constant, for all values of s. It is significant that LES is the only estimator with this property here, unlike the previous two models.

The QL_m estimator, which was nearly identical to QL_{ap} before, now performs worse: The corresponding RRASE ranges from about 14% when $s=100$ to about 10% when $s=1000$.

QL_m is relatively effective when $s=100$ but becomes significantly worse than QL_{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 $\mathrm{ASE}(\mathrm{QL})/\mathrm{ASE}(\mathrm{QL}_{ap})$ 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 5 and 6.

6.5. 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 con-

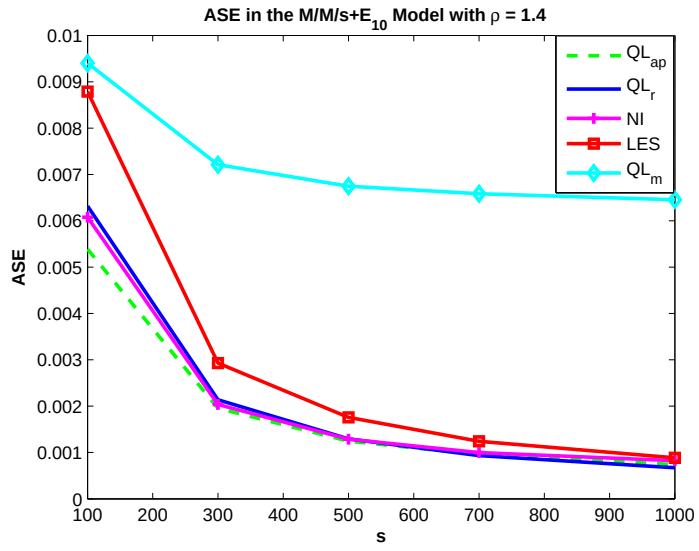

Figure 5

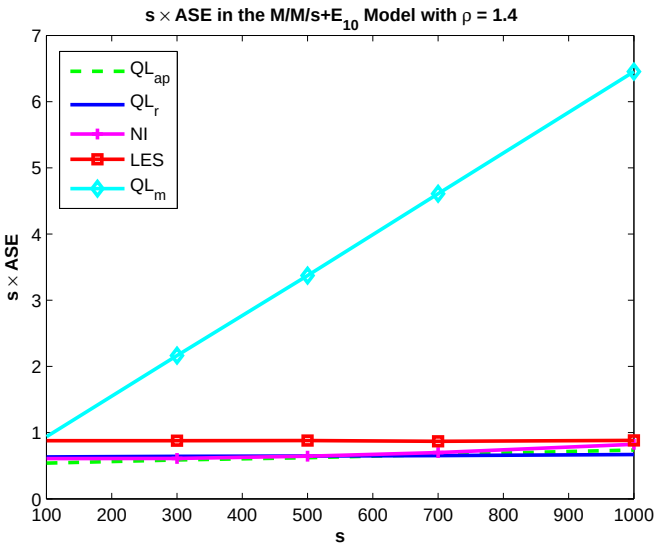

Figure 6

sider 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 (11). Constant service times cause a problem in all cases, but otherwise the estimators perform well; e.g., there is no difficulty when both the service times and abandonments are E_{10}.

7. Conclusions

In this paper, we studied the performance of alternative real-time delay estimators in the overloaded $GI/GI/s+GI$ queueing model, allowing customer abandonment. Customer abandonment makes the system stable. We considered queue-length-based delay estimators exploiting system-state information, including the queue length seen upon arrival. We proposed two new queue-length-based delay estimators - QL_r and QL_{ap} - that effectively cope with non-exponential abandonment-time distributions, and used computer simulation to quantify their effectiveness. We also considered the no-information delay estimator (NI), exploiting no information beyond the model, and the delay of the last customer to have entered service (LES), exploiting customer delay history in the system, but not relying on any model or system-state information. We established heavy-traffic limits for the expected MSE's of the QL_m, LES, and NI delay estimators in the $G/M/s+M$ model, in the ED many-server heavy-traffic limiting regime. For non-exponential abandonment-time distributions, we exploited simulation to study the performance of the candidate delay estimators.

7.1. Performance of the Estimators

We first showed, analytically and by simulation, the need to go beyond the simple queue-length-based estimator, QL, which multiplies the queue length plus one times the mean interval between successive service completions, ignoring customer abandonment. The QL estimator makes consistent estimation error because it ignores customer abandonment; the RRASE (square root of the ASE divided by the mean potential waiting time) can be as high as 40%, and it increases as the number of servers, s, increases.

We considered the Markovian (QL_m) estimator, which assumes exponential service-time and abandonment-time distributions. The QL_m estimator is optimal in the $G/M/s+M$ model, under the MSE criterion, but Figures 5 and 6 show that it can be inferior to all other estimators with non-exponential abandonment-time distributions. Consistent with heavy-traffic theory in Whitt (2006), which shows that the steady-state performance in the ED regime depends strongly upon the time-to-abandon distribution, there can be significant estimation error if we assume that the abandonment-time distribution is exponential when it is not nearly so.

We proposed the simple-refined QL_r estimator, which multiplies the QL estimate by a model-dependent constant, based on heavy-traffic approximations. Simulation shows that QL_r can perform remarkably well, except when the abandonment-time distribution has low variability.

The new approximation-based QL_{ap} estimator was consistently the most effective estimator (with the exception of D service). It approximates the $G/GI/s+GI$ model by the $G/M/s+M(n)$ model, with state-dependent Markovian abandonment rates. The QL_{ap} estimator coincides with QL_m in the setting of the $G/M/s+M$ model, and is thus optimal for that model, under the MSE criterion. Simulation shows that QL_{ap} is significantly better than all other estimators, in all models considered, including models with non-Poisson arrivals and non-exponential service times (with the exception of D service); see the e-companion. The closest competitor to QL_{ap} depends on the model: QL_r is the closest for high to moderately variable abandonment-time distributions, while NI and LES are the closest for low-variability abandonment-time distributions.

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. We obtained an asymptotic approximation for the MSE of NI in the $G/M/s+M$ model, in the ED regime, and found it to be inversely proportional to s.

The LES estimator is particularly appealing because it does not exploit any model or system-state information, and is thereby robust (it responds automatically to changes in system parameters). Simulation suggests that it is an effective estimator in all models considered. That is substantiated by limit theorems for the $G/M/s+M$ model, reviewed next.

7.2. Heavy-traffic Limits for the $G/M/s+M$ Model

We obtained analytical results for the $G/M/s+M$ model, where QL_m is the optimal estimator under the MSE criterion. As can be seen from formulas (19), (25), and (33), the MSE's of QL_m, NI, and LES are inversely proportional to the number of servers, s, in the ED regime. With exponential

abandonments, that is supported by simulation, as shown in Figure 2. Simulation suggests that this remains true with non-exponential abandonment-time distributions for the estimators NI, LES and QL_{ap}, but not for QL_m which assumes exponential abandonments. Figures 4 and 6 show that with hyperexponential and Erlang abandonments, the MSE of QL_m is not nearly inversely proportional to s.

Consistent with analytical results for the $GI/M/s$ model in Ibrahim and Whitt (2007a), we found that the increase in MSE in going from QL_m to LES and NI is primarily due to variability in the arrival process. (The variability of the arrival process is quantified by the squared coefficient of variation, c_a^2.) We quantified the difference in performance between the estimators in formulas (20), (26), (34), and (35): We found that NI is less effective than QL_m, and that the difference in performance is greater when the arrival process is highly variable. We also found that LES is less effective than QL_m, especially when the arrival process has high variability. In the $M/M/s+M$ model, these formulas reduce to simple expressions: Formula (35) shows that the MSE for LES is roughly twice as large as that of QL_m, for all values of $\rho > 1$. On the other hand, formula (38) shows that LES is more effective than NI if $\rho < 2$ and less effective if $\rho > 2$.

7.3. Future Research Directions

It remains to mathematically analyze the performance of the QL_{ap} delay estimator. Simulation suggests that it is uniformly the most effective among all estimators considered, under the MSE criterion, except for the case of D service. It thus also remains to develop special methods for D service.

Our simulations of the $M/M/s+GI$ model in §6, with a general abandonment-time distribution, and Theorem 5, suggest that the analytical results of §5 should extend to that model too. For example, the ASE of the LES estimator is roughly equal to twice the ASE of the QL_{ap} estimator in the $M/M/s+H_2$ model, with hyperexponential abandonments, for all values of s considered. We would like to understand the dependence of this multiplying factor on the abandonment-time and service-time distributions.

We would also like to study the performance of our delay estimators, and possibly new ones, in more general queueing models, such as multiclass call-center models with skill-based routing. We would like to consider time-dependent arrival processes, because arrival processes to a service system are rarely ever stationary. With a non-stationary arrival process, we can use the forecasted or estimated time-varying arrival rate $\lambda(t)$ in the estimators QL_r and QL_{ap}, i.e., in (9) and (10) for QL_r and in (11) for QL_{ap}, but it remains to see when that is appropriate.

Acknowledgments

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

References

Aksin, O.Z., Armony, M. and Mehrotra, V. 2007. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research, *Production and Operations Management*, 16:6, 665 – 688.

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information, *Operations Research*, 52: 527 – 545.

Armony, M., N. Shimkin and W. Whitt. 2008. The impact of delay announcements in many-server queues with abandonments. *Operations Research*, forthcoming.
Available at http://columbia.edu/∼ww2040.

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

Gans, N., G. Koole and A. Mandelbaum. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing and Service Oper. Mgmt.* 5: 79–141.

Garnett, O., A. Mandelbaum, M.I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 5: 79-141

Guo, P. 2007. Analysis and comparison of queues with different levels of delay information, PhD thesis, Duke University.

Hui, M. and D. Tse. 1996. What to tell customers in waits of different lengths: an integrative model of service evaluation. *Journal of Marketing.* 60: 81–90.

Ibrahim, R. and W. Whitt. 2007a. Real-time delay estimation based on delay history. *Manufacturing and Service Oper. Mgmt.* Forthcoming.

Ibrahim, R. and W. Whitt. 2007b. Supplement to "Real time delay estimation based on delay history." IEOR Department, Columbia University, New York, NY. Available at http://columbia.edu/∼ww2040.

Ibrahim, R. and W. Whitt. 2008. Supplement to "Real-time delay estimation in overloaded multiserver queues with abandonments" IEOR Department, Columbia University, New York, NY. Available at http://columbia.edu/∼ww2040.

Jouini, O. 2006. *Stochastic Modeling in Call Centers Operations Management*, PhD thesis, Ecole Centrale Paris.

Jouini, O. Y. Dallery and Z. Aksin. 2007. Modeling call centers with delay information. *Working Paper.*

Maister 1984. D. Psychology of waiting lines. *Harvard Business School Cases.* 71–78.

Mandelbaum A., A. Sakov and S. Zeltyn. 2000. Empirical analysis of a call center. Technical Report, Faculty of Industiral Engineering and Management, The Technion, Israel.

Nakibly, E. 2002. *Predicting Waiting Times in Telephone Service Systems*, MS thesis, the Technion, Haifa, Israel.

Taylor, S. 1994. Waiting for service: the relationship between delays and evaluations of service. *Journal of Marketing,* 58:56-69.

Whitt, W. 1989. Planning queueing simulations. *Management Sc.* 35: 1341–1366.

Whitt, W. 1999a. Predicting queueing delays. *Management Sci.* 45: 870–888.

Whitt, W. 1999b. Improving service by informing customers about anticipated delays. *Management Sci.* 45: 192–207.

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

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

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* 54: 37–54.

Xu, S.H., L. Gao and J. Ou. 2007. Service performance analysis and improvement for a ticket queue with balking customers. *Management Sci.* 53: 971–990.

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

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, corresponding to Figures 1-6 of §6. We present additional experimental results for non-exponential service-time distributions in §EC.4. 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 α. 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 (32). 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

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

for the processes defined in (15), (30) and (24), 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 (32). (In the case of the $G/M/s+M$ model, we assume that the joint limit of $(\hat{N}_s(t), \hat{W}_s(\infty))$ is the same as if these were independent.)

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

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

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

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

jointly with the limit in (EC.2), where B is the given standard Brownian motion and $a(t)$ is defined in (29). We can now apply a random-time-change argument one more time with $W_s(\infty)$ to obtain the limit

$$\hat{Z}_s \equiv \sqrt{s}\left(W_{Q,s}(N_s(W_s(\infty))) - c(\bar{N}_s(W_s(\infty)))\right) \Rightarrow B(d(a(w))) \stackrel{\mathrm{d}}{=} N(0, d(a(w))) \quad \text{in} \quad \mathbb{R} \tag{EC.4}$$

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 (32) by writing

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

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

$$\hat{Y}_s \equiv \sqrt{s}\left(c(\bar{N}_s(W_s(\infty))) - W_s(\infty)\right) \tag{EC.6}$$

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 $(\hat{N}_s(t), \hat{W}_s(\infty))$ in (EC.2), we exploit a Taylor series expansion for the functions $c(t)$ and $a(t)$ in (16) and (29). Note that

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

Hence, $d(a(w)) = d(q) = q/(\lambda\mu)$.

We write

$$\hat{Y}_s \equiv \sqrt{s}\left(c(\bar{N}_s(W_s(\infty))) - W_s(\infty)\right) \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3}, \tag{EC.8}$$

where

$$\begin{aligned}
\hat{Y}_{s,1} &\equiv \sqrt{s}\left(c(\bar{N}_s(W_s(\infty)) - c(a(W_s(\infty)))\right), \\
\hat{Y}_{s,2} &\equiv \sqrt{s}\left(c(a(W_s(\infty))) - c(a(w))\right), \\
\hat{Y}_{s,3} &\equiv \sqrt{s}\left(c(a(w)) - W_s(\infty)\right),
\end{aligned} \tag{EC.9}$$

Using a Taylor series expansion of c, we see that

$$\hat{Y}_{s,1} - c'(w)\sqrt{s}\left(\bar{N}_s(W_s(\infty)) - a(W_s(\infty))\right) \Rightarrow 0, \tag{EC.10}$$

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

$$\hat{Y}_{s,1} \Rightarrow \frac{1}{\lambda}\hat{G}(w) \stackrel{\mathrm{d}}{=} N(0, \sigma_n^2(w)/\lambda^2) \quad \text{as} \quad s \to \infty. \tag{EC.11}$$

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

$$\hat{Y}_{s,2} - c'(a(w))a'(w)\sqrt{s}\left(W_s(\infty) - w\right)) \Rightarrow 0, \tag{EC.12}$$

so that, by Theorem 2,

$$\hat{Y}_{s,2} \Rightarrow \frac{\mu}{\lambda}N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty. \tag{EC.13}$$

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

$$\hat{Y}_{s,3} - c'(a(w))a'(w)\sqrt{s}\left(w - W_s(\infty)\right) \Rightarrow 0, \tag{EC.14}$$

so that, by Theorem 2 again,

$$\hat{Y}_{s,3} \Rightarrow N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty, \tag{EC.15}$$

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,

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

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, λ, to keep the traffic intensity, ρ, fixed for alternative values of s ($\rho = 1.4$). We let the abandonment rate, α, 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 equation (35), 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 (35) is less than 1% throughout. Consistent with equation (26), 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 (26) 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 $\alpha=1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r^m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.867×10^{-3}	2.869×10^{-3}	3.130×10^{-3}	8.693×10^{-3}	5.772×10^{-3}	1.00×10^{-2}
	$\pm1.76\times10^{-5}$	$\pm1.78\times10^{-5}$	$\pm1.89\times10^{-5}$	$\pm3.20\times10^{-5}$	$\pm2.79\times10^{-5}$	$\pm5.97\times10^{-5}$
300	9.587×10^{-4}	9.601×10^{-4}	1.039×10^{-3}	5.602×10^{-3}	1.922×10^{-3}	3.351×10^{-3}
	$\pm6.86\times10^{-6}$	6.92×10^{-6}	$\pm6.41\times10^{-6}$	$\pm2.64\times10^{-5}$	$\pm1.50\times10^{-5}$	$\pm6.03\times10^{-5}$
500	5.761×10^{-4}	5.661×10^{-4}	6.224×10^{-4}	5.017×10^{-3}	1.153×10^{-3}	2.038×10^{-3}
	$\pm1.94\times10^{-6}$	$\pm3.86\times10^{-6}$	$\pm2.94\times10^{-6}$	$\pm2.41\times10^{-5}$	$\pm9.99\times10^{-6}$	$\pm2.26\times10^{-5}$
700	4.104×10^{-4}	4.201×10^{-4}	4.440×10^{-4}	4.682×10^{-3}	8.166×10^{-4}	1.441×10^{-3}
	$\pm1.82\times10^{-6}$	2.839×10^{-4}	$\pm2.71\times10^{-6}$	$\pm2.40\times10^{-5}$	$\pm5.78\times10^{-6}$	$\pm1.57\times10^{-5}$
1000	2.892×10^{-4}	2.839×10^{-4}	3.136×10^{-4}	4.492×10^{-3}	5.752×10^{-4}	1.019×10^{-3}
	$\pm3.48\times10^{-6}$	$\pm3.86\times10^{-6}$	$\pm3.09\times10^{-6}$	$\pm1.54\times10^{-5}$	$\pm6.91\times10^{-6}$	$\pm3.00\times10^{-5}$

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

when s is large (e.g., ASE(QL$_m$)/ASE(QL$_{ap}$) 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 ASE(QL)/ASE(QL$_{ap}$) ranges from about 2 when $s=100$ to about 10 when $s=1000$.

With Erlang abandonments, Table EC.3 shows that QL$_{ap}$ 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$_{ap}$ for relatively small values of s, but slightly outperforms QL$_{ap}$ 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$_{ap}$, particularly when s is large. The LES estimator also fares well, but is slightly outperformed by QL$_{ap}$, QL$_r$ and NI. The QL$_m$ estimator performs significantly worse than QL$_{ap}$ when s is large (e.g., when $s=1000$, ASE(QL$_m$)/ASE(QL$_{ap}$) ≈ 9). Finally, QL is yet again the least effective estimator for this model. The ratio ASE(QL)/ASE(QL$_{ap}$) 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 $\alpha=1.0$

s	$ASE[\theta_{QL_{ap}}]$	$ASE[\theta_{QL_m}]$	$ASE[\theta_{QL_r}]$	$ASE[\theta_{QL}]$	$ASE[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	1.859×10^{-3}	2.100×10^{-3}	2.032×10^{-3}	4.662×10^{-3}	3.866×10^{-3}	6.503×10^{-3}
	$\pm6.52\times10^{-6}$	$\pm5.54\times10^{-6}$	$\pm6.31\times10^{-6}$	$\pm1.83\times10^{-5}$	$\pm8.10\times10^{-6}$	$\pm3.85\times10^{-5}$
300	6.116×10^{-4}	7.933×10^{-4}	6.599×10^{-4}	2.593×10^{-3}	1.236×10^{-3}	2.165×10^{-3}
	$\pm4.64\times10^{-6}$	$\pm7.62\times10^{-6}$	$\pm8.82\times10^{-6}$	$\pm2.25\times10^{-5}$	$\pm1.76\times10^{-5}$	$\pm2.09\times10^{-5}$
500	3.695×10^{-4}	5.367×10^{-4}	3.921×10^{-4}	2.205×10^{-3}	7.331×10^{-4}	1.311×10^{-3}
	$\pm2.19\times10^{-6}$	$\pm2.12\times10^{-6}$	$\pm2.47\times10^{-6}$	$\pm9.97\times10^{-6}$	$\pm5.41\times10^{-6}$	$\pm1.03\times10^{-5}$
700	2.630×10^{-4}	4.257×10^{-4}	2.802×10^{-4}	2.024×10^{-3}	5.250×10^{-4}	9.378×10^{-4}
	$\pm1.43\times10^{-6}$	$\pm1.89\times10^{-6}$	$\pm1.00\times10^{-5}$	$\pm2.35\times10^{-6}$	$\pm2.52\times10^{-6}$	$\pm1.07\times10^{-5}$
1000	1.833×10^{-4}	3.474×10^{-4}	1.978×10^{-4}	1.900×10^{-3}	3.691×10^{-4}	6.533×10^{-4}
	$\pm1.55\times10^{-6}$	$\pm1.43\times10^{-6}$	$\pm6.90\times10^{-7}$	$\pm5.93\times10^{-6}$	$\pm3.00\times10^{-6}$	1.14×10^{-5}

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 $\alpha=1.0$

s	$ASE[\theta_{QL_{ap}}]$	$ASE[\theta_{QL_m}]$	$ASE[\theta_{QL_r}]$	$ASE[\theta_{QL}]$	$ASE[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	5.388×10^{-3}	9.400×10^{-3}	6.317×10^{-3}	8.097×10^{-2}	8.810×10^{-3}	6.077×10^{-3}
	$\pm1.54\times10^{-5}$	$\pm3.48\times10^{-5}$	4.51×10^{-5}	$\pm2.47\times10^{-4}$	$\pm3.91\times10^{-5}$	$\pm2.63\times10^{-5}$
300	1.955×10^{-3}	7.211×10^{-3}	2.139×10^{-3}	7.211×10^{-2}	2.933×10^{-3}	2.040×10^{-3}
	$\pm5.13\times10^{-6}$	$\pm3.86\times10^{-5}$	$\pm1.83\times10^{-5}$	$\pm3.301\times10^{-4}$	$\pm3.22\times10^{-5}$	$\pm2.23\times10^{-5}$
500	1.244×10^{-3}	6.746×10^{-3}	1.293×10^{-3}	7.049×10^{-2}	1.760×10^{-3}	1.288×10^{-3}
	$\pm1.54\times10^{-5}$	$\pm2.68\times10^{-5}$	$\pm1.35\times10^{-5}$	$\pm2.48\times10^{-4}$	$\pm2.44\times10^{-5}$	$\pm2.61\times10^{-5}$
700	9.572×10^{-4}	6.584×10^{-3}	9.319×10^{-4}	6.975×10^{-2}	1.241×10^{-3}	9.966×10^{-4}
	$\pm8.31\times10^{-6}$	$\pm1.43\times10^{-6}$	$\pm1.00\times10^{-5}$	$\pm1.00\times10^{-5}$	$\pm2.35\times10^{-6}$	$\pm1.30\times10^{-5}$
1000	7.369×10^{-4}	6.454×10^{-3}	6.694×10^{-4}	6.902×10^{-2}	8.830×10^{-4}	8.242×10^{-4}
	$\pm1.96\times10^{-5}$	$\pm1.70\times10^{-5}$	$\pm1.13\times10^{-5}$	$\pm1.68\times10^{-4}$	$\pm1.28\times10^{-5}$	$\pm1.17\times10^{-5}$

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

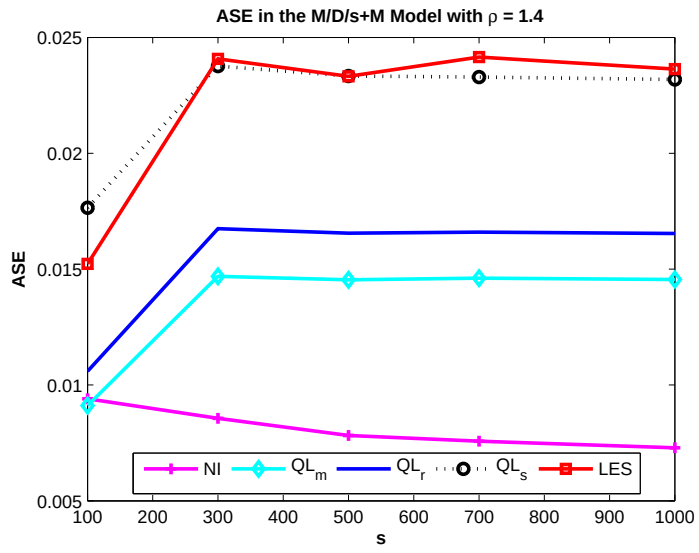

Figure EC.1

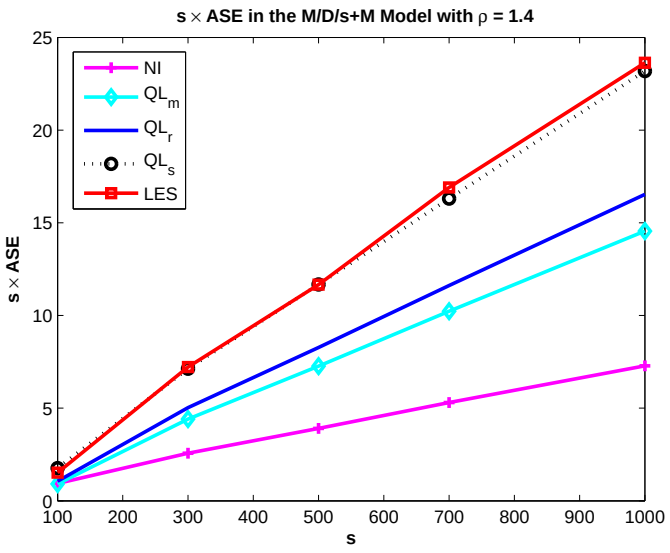

Figure EC.2

equal to 1) distributions. We let $\mu = \alpha = 1.0$, and vary λ, 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., $\text{ASE}(\text{QL}_m)$ 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.

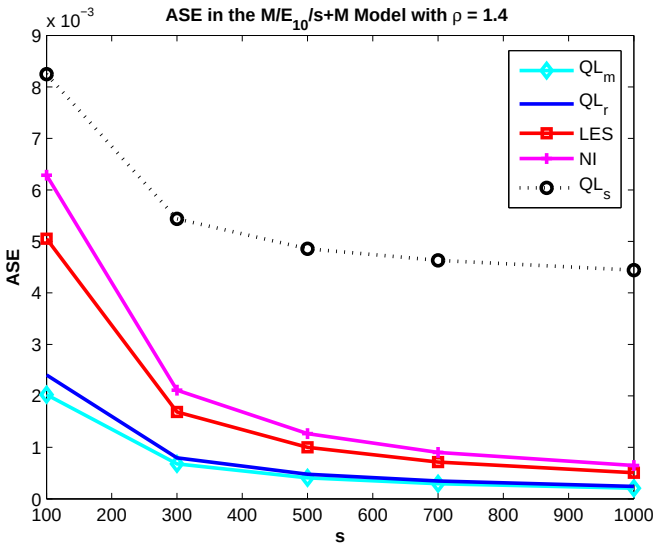

Figure EC.3

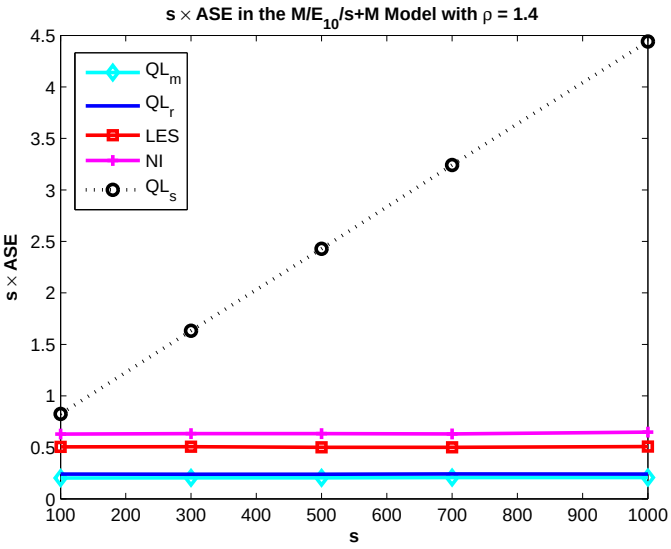

Figure EC.4

Figures EC.3 and EC.4 show that QL_m is the most effective delay estimator for this model. The QL_r estimator is nearly identical to QL_m, particularly when s is large enough ($s \geq 300$). Once more, the relative accuracy of the delay estimators improves as s increases. The RRASE for QL_m 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 $\mathrm{ASE(QL)}/\mathrm{ASE}(\mathrm{QL}_m)$ 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_m is the most effective delay estimator for this model. The RRASE for QL_m ranges from approximately 14% when $s = 100$ to approximately 5% when

Efficiency of the estimators in the $M/LN(1,1)/s+M$ model with $\rho=1.4$ and $\alpha=1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.359×10^{-3}	2.596×10^{-3}	8.207×10^{-3}	5.248×10^{-3}	9.089×10^{-3}
	$\pm7.00\times10^{-6}$	$\pm9.02\times10^{-6}$	$\pm4.45\times10^{-5}$	$\pm2.37\times10^{-5}$	$\pm4.80\times10^{-5}$
300	7.810×10^{-4}	8.506×10^{-4}	5.394×10^{-3}	1.716×10^{-3}	3.032×10^{-3}
	$\pm5.14\times10^{-6}$	$\pm5.68\times10^{-6}$	3.36×10^{-5}	$\pm1.25\times10^{-5}$	$\pm5.30\times10^{-5}$
500	4.663×10^{-4}	5.0685×10^{-4}	4.836×10^{-3}	1.029×10^{-3}	1.826×10^{-3}
	$\pm2.04\times10^{-6}$	$\pm2.12\times10^{-6}$	$\pm2.085\times10^{-5}$	$\pm7.29\times10^{-6}$	$\pm8.10\times10^{-6}$
700	3.346×10^{-4}	3.635×10^{-4}	4.615×10^{-3}	7.438×10^{-4}	1.290×10^{-3}
	$\pm2.71\times10^{-6}$	$\pm3.37\times10^{-6}$	$\pm1.77\times10^{-5}$	$\pm6.47\times10^{-6}$	$\pm1.12\times10^{-5}$
1000	2.340×10^{-4}	2.548×10^{-4}	4.443×10^{-3}	5.290×10^{-4}	8.942×10^{-4}
	$\pm1.84\times10^{-6}$	$\pm2.81\times10^{-6}$	$\pm2.54\times10^{-5}$	$\pm5.90\times10^{-6}$	$\pm2.46\times10^{-5}$

Table EC.4 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 $\mathrm{ASE}(\mathrm{QL}_r)/\mathrm{ASE}(\mathrm{QL}_m)$ 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 $\mathrm{ASE}(\mathrm{QL})/\mathrm{ASE}(\mathrm{QL}_m)$ ranges from approximately 4 when $s=100$ to approximately 19 when $s=1000$.

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 α. Thus, changing α 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 λ to get a fixed value of ρ

($\rho = 1.4$), 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 $\alpha = 0.2$

Table EC.5 compares the efficiencies of the alternative delay estimators in the $D/M/s+M$ model with $\alpha = 0.2$. Consistent with theory, QL_m is the optimal delay estimator for this model, under the MSE criterion. The RRASE of QL_m ranges from approximately 35% when $s = 100$ to approximately 11% when $s = 1000$. The QL_r estimator is slightly less efficient than QL_m: $\mathrm{ASE}(\mathrm{QL}_r)/\mathrm{ASE}(\mathrm{QL}_m)$ 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 $\mathrm{ASE}(\mathrm{QL})/\mathrm{ASE}(\mathrm{QL}_m)$ ranges from approximately 8 when $s = 100$ to approximately 71 when $s = 1000$.

Tables EC.6 and EC.7 substantiate equations (36) and (26) of §5, that compare the performances of QL_m, LES and NI in the $D/M/s+M$ model. Consistent with equation (36), Table EC.6 shows that the performance of LES is close to that of QL_m, when the arrival process is deterministic. The simulation estimates of $\mathrm{ASE}(\mathrm{LES})/\mathrm{ASE}(\mathrm{QL}_m)$, for alternative values of s, are remarkably close to the numerical value, approximately 1.286, predicted by equation (36); the relative error (RE) observed is less than 1% for all values of s considered. Consistent with equation (26), Table EC.7 shows that the performance of NI is worse than that of LES and QL_m. The simulation estimates of $\mathrm{ASE}(\mathrm{NI})/\mathrm{ASE}(\mathrm{QL}_m)$ are also remarkably close to the numerical value, 2.25, predicted by equation (26); the RE observed is less than 4% for all values of s considered.

EC.5.2. Results for the $H_2/M/s+M$ model

Table EC.8 compares the efficiencies of the alternative delay estimators in the $H_2/M/s+M$ model with $\alpha = 5.0$, which makes the model more lightly loaded. Consistent with theory, QL_m is the optimal delay estimator for this model, under the MSE criterion. The RRASE of QL_m 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 $\alpha=0.2$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	1.436×10^{-2}	1.492×10^{-2}	1.192×10^{-1}	1.863×10^{-2}	3.266×10^{-2}
	$\pm9.78\times10^{-5}$	$\pm9.40\times10^{-5}$	$\pm1.57\times10^{-4}$	$\pm1.64\times10^{-4}$	$\pm5.33\times10^{-4}$
300	4.798×10^{-3}	5.005×10^{-3}	1.071×10^{-1}	6.172×10^{-3}	1.056×10^{-2}
	$\pm5.99\times10^{-5}$	$\pm6.08\times10^{-5}$	$\pm1.41\times10^{-4}$	$\pm7.45\times10^{-5}$	$\pm1.92\times10^{-4}$
500	2.865×10^{-3}	2.966×10^{-3}	1.044×10^{-1}	3.672×10^{-3}	6.641×10^{-3}
	$\pm5.43\times10^{-5}$	$\pm5.24\times10^{-5}$	$\pm1.071\times10^{-4}$	$\pm6.67\times10^{-5}$	$\pm2.933\times10^{-4}$
700	2.091×10^{-3}	2.170×10^{-3}	1.033×10^{-1}	2.691×10^{-3}	4.802×10^{-3}
	$\pm2.39\times10^{-5}$	$\pm1.90\times10^{-5}$	$\pm1.53803\times10^{-4}$	$\pm3.23\times10^{-5}$	$\pm2.26\times10^{-4}$
1000	1.435×10^{-3}	1.507×10^{-3}	1.026×10^{-1}	1.859×10^{-3}	3.030×10^{-3}
	$\pm1.15\times10^{-5}$	$\pm1.52\times10^{-5}$	$\pm1.20\times10^{-4}$	$\pm2.06\times10^{-5}$	$\pm1.05\times10^{-4}$

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 $\alpha=0.2$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{LES}]$/ASE$[\theta_{QL_m}]$	Predicted ratio by (36)	RE (%)
100	1.436×10^{-2}	1.863×10^{-2}	1.297	1.286	0.885
	$\pm9.78\times10^{-5}$	$\pm1.642\times10^{-4}$			
300	4.798×10^{-3}	6.172×10^{-3}	1.286	1.286	0.0421
	$\pm5.99\times10^{-5}$	$\pm7.45\times10^{-5}$			
500	2.865×10^{-3}	3.672×10^{-3}	1.281	1.286	-0.329
	$\pm5.43\times10^{-5}$	$\pm6.67\times10^{-5}$			
700	2.091×10^{-3}	2.691×10^{-3}	1.287	1.286	0.107
	$\pm2.39\times10^{-5}$	$\pm3.23\times10^{-5}$			
1000	1.435×10^{-3}	1.859×10^{-3}	1.296	1.286	0.765
	$\pm1.15\times10^{-5}$	$\pm2.05\times10^{-5}$			

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$_m$: ASE(QL$_r$)/ASE(QL$_m$) ranges from approximately

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

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{NI}]$	ASE$[\theta_{NI}]$/ASE$[\theta_{QL_m}]$	Predicted ratio by (26)	RE (%)
100	1.436×10^{-2} $\pm9.78\times10^{-5}$	3.266×10^{-2} $\pm5.33\times10^{-4}$	2.275	2.25	1.09
300	4.798×10^{-3} $\pm5.99\times10^{-5}$	1.056×10^{-2} $\pm1.92\times10^{-4}$	2.201	2.25	-2.18
500	2.865×10^{-3} $\pm5.43\times10^{-5}$	6.641×10^{-3} $\pm2.933\times10^{-4}$	2.318	2.25	3.01
700	2.091×10^{-3} $\pm2.39\times10^{-5}$	4.802×10^{-3} $\pm2.26\times10^{-4}$	2.297	2.25	2.08
1000	1.435×10^{-3} $\pm1.15\times10^{-5}$	3.130×10^{-3} $\pm1.05\times10^{-4}$	2.111	2.25	-3.08

Table EC.7

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

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	7.193×10^{-4} $\pm2.63\times10^{-6}$	1.059×10^{-3} $\pm4.47\times10^{-6}$	2.217×10^{-3} $\pm1.01\times10^{-5}$	2.393×10^{-3} $\pm6.72\times10^{-6}$	3.101×10^{-3} $\pm1.42\times10^{-5}$
300	2.008×10^{-4} $\pm7.85\times10^{-7}$	2.675×10^{-4} $\pm1.28\times10^{-6}$	7.240×10^{-4} $\pm2.63\times10^{-6}$	7.569×10^{-4} $\pm2.70\times10^{-6}$	1.169×10^{-3} $\pm5.82\times10^{-6}$
500	1.167×10^{-4} $\pm7.05\times10^{-7}$	1.495×10^{-4} $\pm8.78\times10^{-7}$	4.792×10^{-4} $\pm2.68\times10^{-6}$	4.540×10^{-4} $\pm1.71\times10^{-6}$	7.624×10^{-4} $\pm6.07\times10^{-6}$
700	8.277×10^{-5} $\pm4.12\times10^{-7}$	1.042×10^{-4} $\pm6.52\times10^{-7}$	3.856×10^{-4} $\pm2.50\times10^{-6}$	3.280×10^{-4} $\pm1.27\times10^{-6}$	5.714×10^{-4} $\pm4.72\times10^{-6}$
1000	5.733×10^{-5} $\pm2.48\times10^{-7}$	7.141×10^{-5} $\pm2.44\times10^{-7}$	3.184×10^{-4} $\pm1.34\times10^{-6}$	2.302×10^{-4} $\pm1.19\times10^{-6}$	4.0951×10^{-4} $\pm4.15\times10^{-6}$

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 ASE(QL)/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 QL$_m$ in the $H_2/M/s+M$ model with $\rho=1.4$ and $\alpha=5.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{LES}]$/ASE$[\theta_{QL_m}]$	Predicted by (37)	RE (%)
100	7.193×10^{-4} $\pm2.63\times10^{-6}$	2.393×10^{-3} $\pm6.72\times10^{-6}$	3.326	4.143	-19.7
300	2.008×10^{-4} $\pm7.85\times10^{-7}$	7.569×10^{-4} $\pm2.70\times10^{-6}$	3.769	4.143	-9.03
500	1.167×10^{-4} $\pm7.05\times10^{-7}$	4.540×10^{-4} $\pm1.71\times10^{-6}$	3.891	4.143	-6.09
700	8.277×10^{-5} $\pm4.12\times10^{-7}$	3.280×10^{-4} $\pm1.27\times10^{-6}$	3.962	4.143	-4.36
1000	5.733×10^{-5} $\pm2.48\times10^{-7}$	2.302×10^{-4} $\pm1.19\times10^{-6}$	4.014	4.143	-3.10

Table EC.9

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

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{NI}]$	ASE$[\theta_{NI}]$/ASE$[\theta_{QL_m}]$	Predicted ratio by (26)	RE (%)
100	7.193×10^{-4} $\pm2.63\times10^{-6}$	3.101×10^{-3} $\pm1.42\times10^{-5}$	4.310	7.25	-40.5
300	2.008×10^{-4} $\pm7.85\times10^{-7}$	1.169×10^{-3} $\pm5.82\times10^{-6}$	5.821	7.25	-19.7
500	1.167×10^{-4} $\pm7.05\times10^{-7}$	7.624×10^{-4} $\pm6.07\times10^{-6}$	6.533	7.25	-9.89
700	8.277×10^{-5} $\pm4.12\times10^{-7}$	5.714×10^{-4} $\pm4.72\times10^{-6}$	6.904	7.25	-4.78
1000	5.733×10^{-5} $\pm2.48\times10^{-7}$	4.0951×10^{-4} $\pm4.15\times10^{-6}$	7.143	7.25	-1.48

Table EC.10

Tables EC.9 and EC.10 substantiate equations (37) and (26) of §5, that compare the performances of QL$_m$, LES and NI in the $H_2/M/s+M$ model. Consistent with equation (37), Table EC.9 shows that the performance of LES is significantly worse than that of QL$_m$, when the arrival process is highly variable. The simulation estimates of ASE(LES)/ASE(QL$_m$), for alternative values of s, are close to the numerical value, approximately 4.143, predicted by equation (37), 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 (37) when the number of servers is small. That is understandable because the system is not very heavily loaded when $\alpha = 5.0$. Consistent with equation (26), Table EC.10 shows that the performance of NI is much worse than that of QL_m, when the arrival process is highly variable. The approximation in (26) performs poorly when $s = 100$ (RE $\approx -40\%$) but becomes remarkably accurate when $s = 1000$ (RE $\approx -1.5\%$).

References

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. J. Amer. Statist. Assoc. 100: 3650.

Ibrahim, R. and W. Whitt. 2008. Supplement to "Real-time delay estimation in overloaded multiserver queues with abandonments" IEOR Department, Columbia University, New York, NY. Available at http://columbia.edu/∼ww2040.