Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Moench, O. Rose, eds.

REAL-TIME DELAY ESTIMATION IN CALL CENTERS

Rouba Ibrahim

Department of Industrial Engineering
Columbia University
500 West 120th Street 313 Mudd
New York, NY 10027, U.S.A.

Ward Whitt

Department of Industrial Engineering
Columbia University
500 West 120th Street 313 Mudd
New York, NY 10027, U.S.A.

ABSTRACT

We use computer simulation to study the performance of alternative real-time delay estimators in heavily loaded multiserver queueing models. These delay estimates may be used to make delay announcements in call centers and related service systems. We consider the classical delay estimator based on the queue length, QL_s, which multiplies the queue length plus one times the mean interval between successive service completions, ignoring customer abandonment. We show that QL_s has a superior performance in the $GI/M/s$ model, but that there is a need to go beyond it in the $GI/GI/s+GI$ model, allowing abandonment. To this end, we propose new, simple and effective, delay estimators based on the queue length. We also consider a delay estimator based on recent customer delay history in the system: the delay of the last customer to enter service, LES.

1 INTRODUCTION

We investigate alternative ways to estimate, in real time, the delay (before entering service) of an arriving customer in a service system. There is empirical evidence suggesting that long waits lead to poor service evaluation, especially when coupled with feelings of uncertainty about the length of the wait; see Taylor (1994). Announcing delay estimates to arriving customers is a relatively inexpensive way to reduce this uncertainty, thereby increasing customer satisfaction with the service provided. We are thinking of making these delay announcements in call centers, where queues are invisible, so that customers are unable to estimate their own delays upon arrival to the system; see Gans et al. (2003) for background on call centers.

We quantify the effectiveness of a delay estimator by the *mean squared error* (MSE), which we estimate via simulation by computing the *average squared error* (ASE). A lower MSE (or ASE) corresponds to a more effective delay estimator. Alternative delay estimators also 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. Comparing the alternative delay estimators is therefore complicated: We would like to have an 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. 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 a new effective and simple way to do better delay estimations and (ii) to describe results of simulation experiments evaluating a wide range of alternative delay estimators in heavily loaded many-server queues.

The rest of this paper is organized as follows: In §2, we describe our delay estimation framework and in §3 we explain how we quantify the performance of a delay estimator. In §4, we introduce candidate delay estimators and in §5 use simulation to compare their performance in the $GI/M/s$ model. The first five sections summarize some results from Ibrahim and Whitt (2007). There, we studied the performance of alternative delay-history-based estimators, both analytically and with simulation in the $GI/M/s$ model. Here, we only show a sample of the simulation results.

In the second part of this paper, we give initial results of ongoing research on delay estimation in the overloaded $GI/GI/s+GI$ model, with customer abandonment. We introduce more queue-length-based delay estimators for this case with customer abandonment in §6 and show simulation results quantifying their performance in §7. We make concluding remarks in §8. Extensions of these results will appear in Ibrahim and Whitt (2008).

2 THE DELAY ESTIMATION FRAMEWORK

We are interested in large, heavily loaded service systems, in which the arrival rate approaches or exceeds the total

service rate. These systems mimic existing call centers, particularly service-oriented ones in which emphasis is placed on efficiency rather than on quality of service. In heavily loaded systems, many customers will be delayed before receiving service and these delays will often be relatively long. This setting is appropriate for our delay estimation problem because we are only concerned here with delayed customers. Moreover, when the delays are negligible, there is little incentive to give delay announcements.

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.

3 QUANTIFYING PERFORMANCE

3.1 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:

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

where a_i is the *potential* delay of customer i, e_i is the delay estimate given to customer i and k is the number of customers in our sample. As in Garnett et al. (2002), a customer's potential delay is the delay he would experience, if he had infinite patience (his patience is quantified by his abandon time). In our simulation experiments, we measure a_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 *mean squared error* (MSE) in steady state.

3.2 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. 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.

4 CANDIDATE DELAY ESTIMATORS

4.1 The Simple Queue-Length-Based Estimator (QL_s)

As in Whitt (1999), a system having s agents each of whom, on average, completes one service request in m minutes, may predict that a customer finding n customers in queue upon arrival, will be able to begin service in $(n+1)m/s$ minutes. Let QL_s refer to this simple queue-length-based delay estimator, commonly used in practice. Let the estimator, as a function of n, be:

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

The QL_s estimator is appealing due to its simplicity and its ease of implementation: It uses information about the system that usually is readily available.

4.2 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 make any assumptions about the model and uses very little information about the system. It is robust, i.e., it responds automatically to changes in system parameters (e.g., number of servers, mean service time, arrival rate), because it does not require knowledge of these parameters. It also extends directly to unconventional queueing models, including multi-class skill-based routing scenarios.

4.3 Alternative Delay-History-Based Estimators

In Ibrahim and Whitt (2007), we 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. We emphasize that 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 concluded 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 DELAY ESTIMATION IN THE $GI/M/s$ MODEL

The standard $GI/M/s$ 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 on the interarrival times when simulating this model: It is not required for the implementation of our delay estimators. There are s homogeneous servers working in parallel and the service times are i.i.d. exponential with mean μ^{-1}. Let the traffic intensity be $\rho \equiv \lambda/s\mu$. It is well known that the $GI/M/s$ model is stable if and only if $\rho < 1$. This model has unlimited waiting space and no customer abandonment. Arriving customers are served in the order of their arrival times; i.e., we use the first-come-first-served (FCFS) service discipline.

5.1 QL_s in the $GI/M/s$ Model

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 rate μ, which makes them i.i.d. exponential with mean $1/s\mu$. The optimal delay estimator, under the MSE criterion, is the one announcing the conditional mean, $E[W_Q(n)]$. But, following the analysis,

$$E[W_Q(n)] = (n+1)/s\mu \ . \tag{2}$$

Since (2) coincides exactly with the QL_s delay estimation, $\theta_{QL}(n)$ in (1), the optimality of QL_s in the $GI/M/s$ model, under the MSE criterion, is mathematically demonstrated. We also have an expression for the MSE of the QL_s estimator: $MSE(\theta_{QL_s}) = Var[W_Q(n)] = (n+1)/(s\mu)^2$, where "*Var*" denotes the variance of a random variable.

As discussed in Whitt (1999), QL_s has the desirable property that it is relatively more accurate for larger values of n. The squared coefficient of variation is given by

$$c^2_{W_Q(n)} \equiv \frac{Var[W_Q(n)]}{(E[W_Q(n)])^2} = \frac{1}{n+1} \ , \tag{3}$$

so that $c^2_{W_Q(n)} \to 0$ as $n \to \infty$.

To help judge the performance of QL_s, we can compare it to the no-information (NI) steady-state delay estimator $\theta_{NI} \equiv E[(W_\infty | W_\infty > 0)]$, with W_∞ denoting the steady-state waiting time before beginning service. Then, $MSE(\theta_{NI}) = Var[(W_\infty | W_\infty > 0)]$. The NI estimator exploits no state information at all, so any other estimator exploiting additional real-time information should do at least as well to be worth serious consideration.

For the $M/M/s$ model, there are simple expressions for the average MSE's of those two estimators, in steady state. Let Q^w_∞ be a random variable with the conditional distribution of the steady-state queue length upon arrival, given that the customer must wait before beginning service. In the $M/M/s$ model, it is well known that $Q^w_\infty + 1$ has a geometric distribution with mean $1/(1-\rho)$ and that $(W_\infty | W_\infty > 0)$ has an exponential distribution with mean $1/s\mu(1-\rho)$ (e.g., this follows from section 5.14 of Cooper (1981)). Hence,

$$\begin{aligned} E[MSE(\theta_{QL_s})] &= \sum_{n=0}^{\infty} Var[W_Q(n)] P(Q^w_\infty = n) \\ &= E[Var(W_Q(Q^w_\infty))] \\ &= \frac{1}{(s\mu)^2(1-\rho)} \ , \end{aligned}$$

so that the corresponding ratio is:

$$\begin{aligned} \frac{MSE(\theta_{NI})}{E[MSE(\theta_{QL_s})]} &= \frac{Var[(W_\infty | W_\infty > 0)]}{E[Var(W_Q(Q^w_\infty))]} \\ &= \frac{1/(s\mu)^2(1-\rho)^2}{1/(s\mu)^2(1-\rho)} \\ &= \frac{1}{1-\rho} \ . \end{aligned}$$

Clearly, as ρ approaches 1, the MSE ratio increases. For example, when $\rho = 0.95$ the MSE for NI is 20 times greater than the average MSE for QL_s, and when $\rho = 0.98$ it is 50 times greater. In contrast, we will see in section 5.3.2 that

the LES estimator produces a corresponding ratio of only approximately $c_a^2+1=2$ for values of ρ approaching 1.

5.2 LES in the $GI/M/s$ Model

Let $W_{LES}(w)$ be a random variable with the conditional distribution of the waiting time of a new arrival given that the new arrival must join the queue and given that the delay of the last customer to have entered service is w. $W_{LES}(w)$ has a relatively complicated exact distribution because we do not know precisely what happens in the interval between the time that the LES customer arrived and the new arrival epoch.

Since the current queue length is approximately equal to the number of arrivals during this LES waiting time, w, the increase in MSE in going from QL_s to LES is primarily due to the variability in the arrival process. That is confirmed by our simulation experiments.

5.3 Simulation Experiments For the $GI/M/s$ Model

5.3.1 Description of the Experiments

In Tables 1 and 2, we report point and confidence interval estimates for the ASE's of QL_s and LES in the $M/M/100$ and $D/M/100$ (deterministic interarrival times) models, as a function of the traffic intensity ρ. We only consider values of ρ larger than or equal to 0.9 since we are interested in heavily loaded systems (but we have $\rho<1$ to guarantee the stability of the system). We fix μ and vary λ to get different values of ρ.

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 $GI/M/s$ model are based on 10 independent replications of 5 million events each. An event is either a service completion or an arrival event. That is, we end each simulation run when the sum of the number of service completions and arrival events equals 5 million. We collect statistics without deleting an initial segment of the data because we found that the impact of having a warmup period is negligible in this setting.

5.3.2 The Simulation Results for $GI/M/s$

Table 1 shows that the QL_s estimator significantly outperforms the LES estimator in the $M/M/100$ model, where the interarrival times have squared coefficient of variation (SCV, denoted by c_a^2, equal to the variance divided by the square of the mean) equal to 1. In this case, we see that the ratio ASE(LES)/ASE(QL_s) is approximately equal to $2/\rho$, especially when ρ is large (e.g., $\rho=0.98$). To better assess the accuracy of the estimators, we compute the corresponding RASE (relative average squared error, equal to the ASE divided by the mean squared). For the QL_s estimator, the RASE ranges from about 10% when $\rho=0.9$ to about 2% when $\rho=0.98$. As implied by (3), we see that the QL_s estimation is relatively more accurate as ρ approaches 1. The RASE for LES ranges from about 22% when $\rho=0.9$ to about 4% when $\rho=0.98$. These results suggest that the LES estimator, too, is relatively more accurate as ρ approaches 1. In Ibrahim and Whitt (2007), we give supporting analytical results: We show that the relative error of the LES estimator is asymptotically negligible in heavy-traffic.

For the $D/M/100$ model, where the interarrival times have SCV equal to 0, Table 2 shows that the reported ASE's are closer. Here, ASE(LES)/ASE(QL_s) is approximately equal to $1/\rho$. The RASE of the QL_s estimator ranges from about 20% when $\rho=0.9$ to about 4% when $\rho=0.98$. The RASE of the LES estimator ranges from about 25% when $\rho=0.9$ to about 4% when $\rho=0.98$. Both estimators are relatively more accurate as ρ approaches 1.

These results confirm our initial observation: When the variability of the arrival process is low, these two delay estimators perform nearly the same. In Ibrahim and Whitt (2007), we analyze the performance of LES and QL_s in the $GI/M/s$ model and prove that the ratio of the respective ASE's, ASE(LES)/ASE(QL_s), is asymptotically equal to $(c_a^2+1)/\rho$ as s grows or ρ approaches 1.

5.4 Simulation Experiments for the $GI/GI/s$ Model

Simulation experiments for the $GI/GI/s$ model, with generally distributed i.i.d. service times, yield similar results. For the service-time distribution, we consider D and H_2 (hyperexponential with SCV equal to 4 and balanced means). We try different combinations of interarrival and service-time distributions and study the performance of QL_s and LES in each case. As before, QL_s is more effective than LES, and the difference in performance is greater when the arrival process is highly variable; e.g., the ratio ASE(LES)/ASE(QL_s) is slightly larger than 1 (roughly equal to 1.06) in the $D/H_2/s$ model, but is close to 2.6 in the $H_2/D/s$ model, for all values of ρ considered.

6 CUSTOMER ABANDONMENT

We now consider delay estimation when there is customer abandonment. We also allow for non-exponential distributions. The $GI/GI/s+GI$ model has i.i.d. service times again with mean μ^{-1}, but now with a general distribution.

Table 1: ASE in the $M/M/100$ model in units of 10^{-3}

ρ	QL_s	LES
0.98	5.115 $\pm$0.0473	10.67 $\pm$0.0987
0.95	2.053 $\pm$0.00383	4.311 $\pm$0.00801
0.93	1.432 $\pm$0.00319	3.014 $\pm$0.00719
0.90	0.9948 $\pm$0.00192	2.215 $\pm$0.00442

Table 2: ASE in the $D/M/100$ model in units of 10^{-3}

ρ	QL_s	LES
0.98	2.448 $\pm$0.00424	2.603 $\pm$0.00472
0.95	1.012 $\pm$0.00347	1.162 $\pm$0.00358
0.93	0.7363 $\pm$0.00318	0.8830 $\pm$0.00314
0.9	0.5356 $\pm$0.00689	0.6840 $\pm$0.00747

Associated with each arriving customer is an abandonment time quantifying his patience. Abandonment times are i.i.d. with mean α^{-1} and a general distribution.

6.1 The Need to go Beyond QL_s

Intuitively, we expect that when there is significant customer abandonment, the QL_s estimator will overestimate the potential delay, because many customers in queue may abandon before entering service, and QL_s fails to take that into account. That is confirmed by our simulation results for the $GI/GI/s+GI$ model in §7. Consequently, we are motivated to consider new delay estimators. Here, we go beyond QL_s, and propose two other delay estimators based on the queue-length in the system upon arrival: (i) the Markovian queue-length-based delay estimator, QL_m, and (ii) the simple-refined queue-length-based delay estimator, QL_{sr}. In Table 3, we summarize the information needed for the implementation of each queue-length-based delay estimator considered in this paper.

6.2 Markovian Queue-Length-Based Delay Estimator

This delay estimator approximates the service-time and abandonment-time distributions by the exponential distribution. That is, it 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:

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

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 is defined by the corresponding delay estimate, $\theta_{QL_m}(n)$, given to a customer finding n customers in queue upon arrival:

$$\theta_{QL_m}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + i\alpha} \ .$$

Note that QL_m is the best possible, under the MSE criterion, in the $GI/M/s+M$ model since the corresponding delay estimate is equal to the conditional mean, which minimizes the MSE, but we find that it is not always so good for the more general $GI/GI/s+GI$ model.

6.3 The Simple-Refined Estimator (QL_{sr})

We design a simple refinement of QL_s by making use of the steady-state fluid approximations to the general $G/GI/s+GI$ model, in the efficiency driven (ED) limiting regime, as developed by Whitt (2006). The ED approximations are appropriate when the number of servers s and the arrival rate λ are large, with the traffic intensity ρ held fixed at a value greater than 1. We first introduce some notation. Let F be the cumulative distribution function (cdf) of the abandon-time distribution. In the steady-state fluid limit, all customers wait the same deterministic amount of time, w, and they all see the same number of customers, q, in queue upon arrival. These deterministic quantities 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:

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

and

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

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

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

For QL_{sr}, 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_s estimator. That is, the new delay estimate is:

$$\theta_{QL_{sr}}(n) \equiv \beta \times \theta_{QL_s}(n) = \beta(n+1)/s\mu \ .$$

The QL_{sr} estimator is appealing because it makes use of the simple form of the QL_s estimator, but performs much better in models with customer abandonment, as we show next. Note that in addition to s, n and μ, we need to know ρ or, equivalently, λ in order to implement QL_{sr}. It is significant that both QL_s and QL_m, on the other hand, are independent of the arrival process: For these two estimators, the arrival process can be arbitrary, even non-stationary.

7 MORE SIMULATION RESULTS

In this section, we present simulation results quantifying the performance of our alternative delay estimators in the $GI/GI/s+GI$ model. Here, we let the interarrival and service times be exponential and vary the abandonment-time distribution. We use a Poisson arrival process because it is usually a good model. We use exponential service times because Whitt (2006) showed that the steady-state performance in the ED regime depends strongly upon the time-to-abandon distribution beyond its mean, but hardly at all upon the service-time distribution beyond its mean. We also conducted simulations with non-exponential interarrival and service times, which we describe in Ibrahim and Whitt (2008).

7.1 Description of the Simulation Experiments

Figures 1, 2 and 3 show point estimates of the ASE's of our alternative delay estimators as a function of the number of servers, s. For the abandonment-time distribution, we consider M (exponential), H_2 (hyperexponential), and E_{10} (Erlang, sum of 10 exponentials) distributions. We consider large values of s since we are interested in large service systems ($s = 100, 300, 500, 700$ and 1000). We let $\rho = 1.4$ in all models. This value is chosen to let our systems be significantly overloaded. We let $\mu = \alpha = 1$ and vary λ to get a fixed value of ρ for alternative values of s. Because of the 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).

Table 3: Information required for queue-length estimators

QL_s	Q(t), s, μ
QL_{sr}	Q(t), s, μ, $F(x)$, λ
QL_m	Q(t), s, μ, α

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: The effect of the initial transient period is negligible in these models (relative error less than 5%) so we don't include any warmup period. Note that an event in this case is either a service completion, an arrival event or an abandonment from the system. 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.

7.2 Simulation Results

7.2.1 $M/M/s+M$ Model

For this model, Figure 1 shows that the Markovian queue-length-based delay estimator, QL_m, is the best possible, under the MSE criterion. The RASE for QL_m ranges from about 2% for $s = 100$ to about 0.2% 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 QL_s being the least accurate). For example, the RASE of LES ranges from about 5% for $s = 100$ to about 0.5% for $s = 1000$.

It is interesting to note that QL_{sr} performs nearly as well as QL_m, particularly when the number of servers s is large. This is so because the fluid approximation is more appropriate with a large number of servers (under heavy loading). The ratio ASE(QL_{sr})/ASE(QL_m) is close to 1 for all values of s.

The LES estimator performs worse than QL_m and QL_{sr}. The ratio ASE(LES)/ASE(QL_m) is close to 2 for all values of s, but mathematical support for this has yet to be provided.

The QL_s estimator performs significantly worse than the other three estimators and its performance gets worse as s increases. The ratio ASE(QL_s)/ASE(QL_m) ranges from about 3 when $s = 100$ to nearly 16 when $s = 1000$. This shows the need to go beyond QL_s when customer abandonment is included.

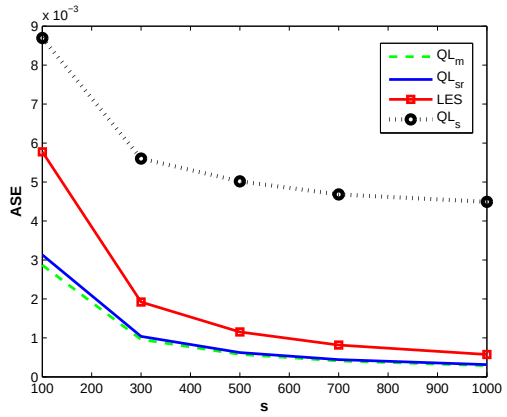

Figure 1: M/M/s+M model.

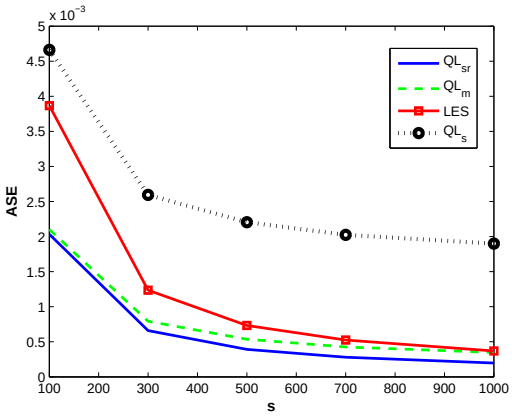

Figure 2: M/M/s+H_2 model.

7.2.2 $M/M/s+H_2$ Model

Figure 2 shows that the best delay estimator for this model is the QL_{sr} delay estimator. The corresponding RASE ranges from about 5% for $s = 100$ to about 0.6% when $s = 1000$. Once more, we see that the 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 RASE of the LES estimator ranges from about 9% when $s = 100$ to about 1% when $s = 1000$.

The QL_m estimator performs well but it is now slightly outperformed by QL_{sr}. The two are nearly the same when $s = 100$; the ratio $\mathrm{ASE}(\mathrm{QL}_m)/\mathrm{ASE}(\mathrm{QL}_{sr})$ is close to 1 when $s = 100$ but closer to 2 when $s = 1000$.

The LES estimator performs worse than both QL_m and QL_{sr} when $s = 100$ but nearly the same as QL_m when $s = 1000$. The ratio $\mathrm{ASE}(\mathrm{LES})/\mathrm{ASE}(\mathrm{QL}_{sr})$ is close to 2 for all values of s.

As above, the efficiency of QL_s is degrading as the number of servers increases. The ratio $\mathrm{ASE}(\mathrm{QL}_s)/\mathrm{ASE}(\mathrm{QL}_{sr})$ ranges from about 2 when $s = 100$ to about 10 when $s = 1000$. Once more, the need to go beyond QL_s is evident.

7.2.3 $M/M/s+E_{10}$ Model

Figure 3 shows that, in contrast to previous cases, LES is the most effective delay estimator here. The corresponding RASE ranges from about 2% when $s = 100$ to about 0.1% when $s = 1000$. Compared to LES, QL_{sr} performs worse as the number of servers increase: $\mathrm{ASE}(\mathrm{QL}_{sr})/\mathrm{ASE}(\mathrm{LES})$ ranges from nearly 1 when $s = 100$ to nearly 3 when $s = 1000$.

The QL_m estimator is effective when $s = 100$ but becomes significantly worse than LES when $s = 1000$ (in that case, the ratio of respective ASE's is close to 6).

The QL_s estimator is consistently the least effective delay estimator in this model too; the ratio $\mathrm{ASE}(\mathrm{QL}_s)/\mathrm{ASE}(\mathrm{LES})$ ranges from about 9 when $s = 100$ to nearly 60 when $s = 1000$. That is why the corresponding ASE curve is not even included in Figure 3.

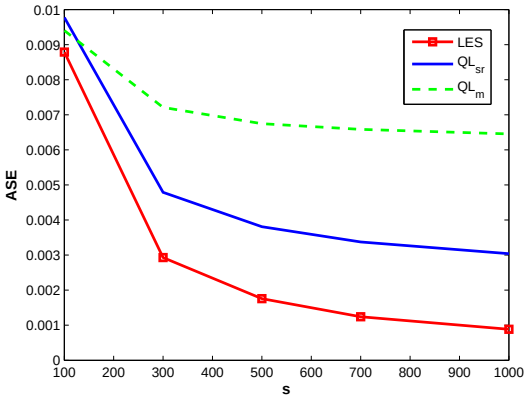

Figure 3: M/M/s+E_{10} model.

8 CONCLUSIONS

In this work, we used computer simulation to compare the performance of alternative real-time delay estimators in the $GI/M/s$ and $GI/GI/s+GI$ queueing models. Simulation plays a critical role in studying the performance of all the candidate estimators, because analytical results are hard to come by.

The delay estimation problem is easier in the $GI/M/s$ model without abandonment, where it is easy to justify that QL_s is the optimal single-number delay estimator, under the MSE criterion. However, LES has the important advantage of not relying on any model information. We quantified the performance of QL_s and LES and we showed that the

difference in performance between the two need not be great, especially when the variability of the arrival process is low.

The $GI/GI/s+GI$ model, with abandonment, is more complicated. We showed the need to go beyond QL_s. We proposed a new simple and effective queue-length-based delay estimator, QL_{sr}. Our simulation results in §7 suggest that the performance of the alternative delay estimators proposed differ according to the particular model at hand. Nevertheless, these delay estimators can perform remarkably well in the models considered: This good performance is quantified by the RASE reported in §7.

In Ibrahim and Whitt (2008), we hope to propose other queue-length-based delay estimators for the $GI/GI/s+GI$ model, that are even more effective than the estimators that we consider here.

ACKNOWLEDGMENTS

This research was supported by NSF Grant DMI-0457095.

REFERENCES

Armony, M., N. Shimkin and W. Whitt. 2006. The impact of delay announcements in many-server queues with abandonments. *Operations Research*, forthcoming.

Cooper, R. B. 1981. *Introduction to Queueing Theory*, second edition, North-Holland, New York.

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

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

Ibrahim, R. and W. Whitt. 2007. Real time delay estimation based on delay history. *Manufacturing and Service Operations Management*, forthcoming.

Ibrahim, R. and W. Whitt. 2008. Real time delay estimation in overloaded multiserver queues with abandonments. IEOR Department, Columbia University, New York, NY.

Taylor, S. 1994. Waiting for Service: The Relationship Between Delays and Evaluations of Service. *Journal of Marketing* 58:56–69.

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

Whitt, W. 2006. Fluid Models for Multiserver Queues with Abandonments. *Operations Research* 54: 37–54.

AUTHOR BIOGRAPHIES

ROUBA IBRAHIM is a doctoral student in the Department of Industrial Engineering and Operations Research at Columbia University. She joined this doctoral program in 2004. Her research interests lie in queueing theory and simulation modeling.

WARD WHITT is a professor in the Department of Industrial Engineering and Operations Research at Columbia University. He joined the faculty there in 2002 after spending 25 years in research at AT&T. He received his Ph.D. from Cornell University in 1969. His recent research has focused on stochastic models of customer contact centers, using both queueing theory and simulation.