MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 3, Summer 2009, pp. 397–415
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1103 | 0397
DOI 10.1287/msom.1080.0223


# Real-Time Delay Estimation Based on Delay History

Rouba Ibrahim, Ward Whitt
IEOR Department, Columbia University, New York, New York 10027
{rei2101@columbia.edu, ww2040@columbia.edu}

Motivated by interest in making delay announcements to arriving customers who must wait in call centers and related service systems, we study the performance of alternative real-time delay estimators based on recent customer delay experience. The main estimators considered are: (i) the delay of the last customer to enter service (LES), (ii) the delay experienced so far by the customer at the head of the line (HOL), and (iii) the delay experienced by the customer to have arrived most recently among those who have already completed service (RCS). We compare these delay-history estimators to the standard estimator based on the queue length (QL), commonly used in practice, which requires knowledge of the mean interval between successive service completions in addition to the QL. We characterize performance by the mean squared error (MSE). We do an analysis and conduct simulations for the standard $GI/M/s$ multiserver queueing model, emphasizing the case of large $s$. We obtain analytical results for the conditional distribution of the delay given the observed HOL delay. An approximation to its mean value serves as a refined estimator. For all three candidate delay estimators, the MSE relative to the square of the mean is asymptotically negligible in the many-server and classical heavy-traffic (HT) limiting regimes.

*Key words*: delay estimation; real-time delay estimation; delay prediction; delay announcements; many-server queues; call centers; heavy traffic
*History*: Received: May 27, 2007; accepted: April 8, 2008. Published online in *Articles in Advance* September 9, 2008.

## 1. Introduction

In this paper, we study alternative ways to estimate the delay (before entering service) of an arriving customer in a service system. These delay estimates may be used to make delay announcements to arriving customers, especially when the delay will be relatively long. Such real-time delay announcements can be very helpful with invisible queues, as in call centers, where service requests are made by telephone; see Gans et al. (2003) for a background on call centers.

Since the steady-state waiting-time distribution tends to be quite highly variable (e.g., often exponential or approximately so), good real-time delay estimation necessarily relies on state information; see Whitt (1999). From the perspective of statistical precision, for a single-number estimate, we would ideally want to use the conditional expected delay given all information available at the arrival epoch, but complexity leads to considering more elementary alternatives.

### 1.1. Standard Queue Length (QL) Delay Estimator

The standard state-dependent delay estimator, commonly used in practice (assuming service from a queue in first-come first-served order, but without any other specific model assumptions), is the *queue-length* (*QL*) *delay estimator*, defined as

$$\theta_{\mathrm{QL}}(t) \equiv \frac{Q(t)+1}{r(t)}, \tag{1}$$

where the notation $\equiv$ means "defined as," $t$ is the current time (time of the arrival for which the announcement is made), $Q(t)$ is the QL (number of customers waiting), and $r(t)$ is the rate at which customers enter service (typically, not known precisely). If the number of servers is $s(t)$, and can be assumed to remain at that level in the near future, with each server serving a single customer without interruption, and the current average service time is $m(t)$, then the rate customers enter service may be approximated by $r(t) = s(t)/m(t)$. Furthermore, when the mean service time is stable, we can replace $m(t)$ by a long-run average service time $m$. The QL delay estimator then becomes $\theta_{\mathrm{QL}}(t) \equiv m(Q(t)+1)/s(t)$, which requires knowledge of only $s(t)$, the number of servers, and $Q(t)$, the QL, at each time $t$, which is information that usually is readily available.

### 1.2. Estimators Based on Delay History

In this paper, we examine alternative estimators based on the delays actually experienced by recent customers, in particular: (i) the delay of the last customer to enter service (LES), (ii) the delay experienced so far by the customer at the head of the line (HOL), (iii) the delay experienced by the customer to have arrived most recently among those who have completed service (RCS).

These delay estimators based on recent delay history are appealing because they are easy to interpret, and because they are simple and robust, applying to a broad range of models, without requiring knowledge of the model or its parameters. If somehow the QL, $Q(t)$, or the rate at which customers enter service, $r(t)$, is unknown or incorrect, then we would have difficulties with the standard QL estimator. With any prediction system, it is good to monitor its performance, but that is often not possible for the customer. A delay-history delay estimator has the advantage that the basis for the prediction is evident.

The HOL delay estimator was used as an announcement in an Israeli bank studied by Mandelbaum et al. (2000) and is mentioned as a candidate delay announcement by Nakibly (2002) in her study of delay predictions. Something similar to LES or RCS is used by the U.S. Citizenship and Immigration Service; they publish the arrival time of recently completed applications to give an idea about upcoming delays. In this study, we are motivated, in part, by recent work by Armony et al. (2009), who studied delay announcements in many-server queues with customer abandonment, focusing on customer response to the announcements, leading to balking and new abandonment behavior. They developed ways to approximately describe the equilibrium system performance using LES delay announcements. Armony et al. (2009) discuss the motivation for the LES delay estimator and other delays estimators based on recent delay history.

### 1.3. Quantifying the Effectiveness

We quantify the effectiveness of the delay estimators through the mean squared error (MSE), which we approximate analytically and estimate via simulation. To illustrate, let $W_{\mathrm{LES}}(w)$ denote the random delay of a new arrival, conditional on that customer having to wait and an observed LES delay of $w$ (under specified conditions, e.g., in steady state). Let $\theta_{\mathrm{LES}}(w)$ be a candidate estimator based on this information. We will primarily be concerned with the direct estimator $\theta^{d}_{\mathrm{LES}}(w) \equiv w$, the refined estimator $\theta^{r}_{\mathrm{LES}}(w) \equiv E[W_{\mathrm{LES}}(w)]$, and approximations of the refined estimator, since the refined estimator is difficult to determine. The MSE of such an estimator is

$$\mathrm{MSE} \equiv \mathrm{MSE}(\theta_{\mathrm{LES}}(w)) \equiv E[(W_{\mathrm{LES}}(w) - \theta_{\mathrm{LES}}(w))^2]. \qquad (2)$$

For the refined estimator $\theta^{r}_{\mathrm{LES}}(w)$, the MSE coincides with the variance $\mathrm{Var}(W_{\mathrm{LES}}(w))$. It is well known that the mean minimizes the MSE (using that information).

To estimate these MSEs via simulation, we use the average squared error (ASE), defined by

$$\mathrm{ASE} \equiv \frac{1}{n}\sum_{j=1}^{n}(a_j - e_j)^2, \qquad (3)$$

where $a_j$ is the actual delay and $e_j$ is the estimated delay for appropriate customers. For example, if we want to estimate the performance of LES when the observed delay is $w = 0.40$, then we consider all arrivals who must wait ($a_j > 0$) for which the LES delay $e_j$ falls in an interval such as $[0.39, 0.41]$. On the other hand, if we wish to consider the overall average performance of LES, then we consider all $j$ such that $a_j > 0$.

### 1.4. Study in an Idealized Setting

In this paper, we study the performance of the delay-history delay estimators and compare them to the standard QL delay estimator in the relatively simple idealized setting of the $GI/M/s$ queueing model, which has a renewal arrival process, $s$ homogeneous servers working in parallel, unlimited waiting space, a first-come first-served service discipline and i.i.d. exponential service times with mean $m$, which are independent of the arrival process. For this $GI/M/s$ model, the QL estimator $\theta_{\mathrm{QL}}(t) \equiv m(Q(t)+1)/s$ is an ideal delay estimator. Indeed, there are no serious competitors, as far as statistical precision is concerned (provided that we have no information about remaining service times). Given the queue length, the future evolution of the system is independent of the past. (This even remains true for more general arrival processes.) Consequently, $\theta_{\mathrm{QL}}(t)$ is the conditional mean delay given all information available at time $t$, so that it minimizes the MSE.

We study the alternative delay-history delay estimators in this simple context to gain insight about the relative performance of alternative estimators in more complex scenarios (which are much more difficult to analyze directly). We know that the QL estimator will have superior performance for the $GI/M/s$ model, but we want to understand by how much. That knowledge will help us understand the advantage of the QL estimator over these alternative delay estimators when the QL estimator is appropriate, and will provide useful background when considering these alternative delay estimators for more complicated systems for which these alternative estimators may be preferred.

### 1.5. Motivation for Considering Alternative Delay Estimators

Whenever the actual service system is well modeled by a $GI/M/s$ queueing model and the system state is known accurately at each time, then there is little motivation for considering other delay estimators besides the standard QL estimator. However, real service systems rarely are as simple as the $GI/M/s$ model. First, the service-time distribution might well be nonexponential, as shown for call centers by Brown et al. (2005). Second, the number of servers and mean service times often change over time, in part, because the servers are humans who serve in different shifts and may well have different service-time distributions. Third, the QL may not be 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 QL is not known to ticket-holding customers or even to system managers, because they do not observe customer abandonments.

Finally, the system is often much more complicated: For one example, there may be multiple customer classes and multiple service pools with some form of skill-based routing; see Gans et al. (2003). For a second example, with Web chat, servers may serve several customers simultaneously, different servers may participate in a single service, and there may be interruptions in the service times, as the customers explore material on the Web in between conversations with agents. For a third example, when delays are large—which is when we most want to make delay announcements—customers often abandon from queue. In these more complicated settings, the QL is typically known, but the rate customers enter service is often not known and/or difficult to estimate reliably. That causes problems for the QL estimator.

When the $GI/M/s$ model is not appropriate for one of these reasons, the QL estimator may not perform well.

Example (Nonexponential Service Times). To dramatically illustrate the possible difficulties with the QL delay estimator in the presence of a nonexponential service-time distribution (without trying to be realistic), we consider a limiting hyperexponential ($H_2$) distribution, in which each service time is either an exponential with mean 10, with probability 1/10, or the deterministic value 0, with probability 9/10. Thus the service time has mean 1, but busy servers will only be serving customers with the exponential distribution. Let $s = 100$ and suppose that an arrival finds the queue empty but all the servers busy. Then, the QL delay estimate for this new arrival is $1/s = 1/100$, but the actual delay is exponentially distributed with mean 1/10 (the minimum of 100 exponential random variables, each with mean 10). Hence the actual mean delay is 10 times greater than predicted by the QL delay estimator. Consistent with this extreme example, we have found that our alternative delay estimators actually outperform the QL delay estimator in the $D/H_2/100$ model with moderately variable $H_2$ distributions.

Similarly, when there is a large amount of customer abandonment, the QL estimator will tend to overestimate the potential delay (the delay assuming that the customer has infinite patience), because many customers in queue may abandon before entering service, and the standard QL estimator fails to take that into account. As discussed in Whitt (1999), the QL estimator can be revised to provide an accurate estimate of delays with abandonments when the time-to-abandon distribution is exponential. However, as discussed in Whitt (2006), the performance measures in the overloaded $M/M/s+GI$ model, with nonexponential time-to-abandon distribution, depend strongly on the time-to-abandon distribution beyond its mean. Since the time-to-abandon distribution has been found to be nonexponential in practice, see Brown et al. (2005), there also are potential difficulties with the generalized QL estimator based on the $M/M/s+M$ model. We investigate alternative delay estimators in the presence of abandonments in a sequel to this paper, Ibrahim and Whitt (2008). There we give examples with nonexponential distributions in which both the standard QL estimator and the refinement for the $M/M/s+M$ model are outperformed by delay estimators based on recent delay history.

From the above discussion, we conclude that other estimators besides the standard QL estimator are worth considering; we do not conclude that the standard QL estimator or other estimators based on the QL are necessarily bad. Indeed, we will show advantages of the QL estimator when it can be used.

### 1.6. This Study

Here, we study the performance of the delay estimators based on delay history in the relatively simple idealized setting of the $GI/M/s$ model. Motivated by call centers, we are especially interested in the case of large $s$, but we consider all possible $s$.

For this more elementary $GI/M/s$ model, we obtain strong analytical results and make comparisons through computer simulations. Unlike Armony et al. (2009), here we do not consider customer response and we do not consider balking or customer abandonment, although we recognize that those phenomena are important. Moreover, here we are not concerned with what to announce, for which we should consider interpretation and response, but only with the effectiveness of the candidate delay estimators in predicting the actual delay encountered (assuming no customer response).

We find that the conditional distribution of the delay to be estimated, given the observed past delay, is often approximately normally distributed, implying that the conditional distribution is approximately characterized by its mean and variance. The observed delay is the natural *direct estimator* of the delay to be encountered by the new arrival, while the mean of the conditional distribution of the delay of the new arrival, given that observed delay, is a natural *refined estimator* based on the same information. (In general, these are different!) The refined estimator depends on the model and its parameters. Because the conditional mean is complicated, we develop approximations for it.

For the $GI/M/s$ model, we will show that the QL estimator does indeed perform better than the alternative estimators based on recent delays, and we will quantify the difference. Roughly, the MSE differs by the constant factor $c_a^2+1$, where $c_a^2$ is the squared coefficient of variation (SCV; variance divided by the square of the mean) of an interarrival time. Thus the MSEs of the delay-history estimators are about the same as the MSE of the QL estimator when the arrival-process variability is low, but considerably greater when the arrival-process variability is high.

### 1.7. Related Literature

There is a large body of related literature with somewhat different goals. We are doing statistical inference for queues, but as in Avramidis et al. (2004), Brown et al. (2005), and Glynn and Whitt (1989), most statistical inference for queues aims to estimate the *model* or the *steady-state performance*. There is an interesting stream of literature related to estimating *past* performance in a partially observed system from transactional data, stemming from Larson (1990). There has been much interesting recent inference work, including delay estimation, related to the *Internet*, as surveyed by Coates et al. (2002), but our setting and time scales tend to be very different. In addition to Whitt (1999), delay estimation for real-time delay prediction is investigated by Ward and Whitt (2000) and Nakibly (2002); these focus on processor sharing and priority disciplines, respectively. Our real-time focus is in the spirit of real-time queueing, as in Doytchinov et al. (2001) and references therein.

### 1.8. Organization of the Paper

We start in §2 by defining alternative delay estimators based on recent delay history and giving some expressions for them for the $GI/M/s$ model. We present results of initial simulation experiments in §3. We establish properties of two basic delay estimators—LES and the HOL estimator—in §4. We present confirming simulations related to those analytical results in §5. We discuss insights from heavy-traffic (HT) limits in §6. Finally, we draw conclusions in §7. We present additional material in an e-companion, including more experimental results, more HT limits, and a cautionary example showing the possible pitfalls of the LES and HOL delay estimators for nonexponential service-time distributions. We present even more experimental results in an online supplement (Ibrahim and Whitt 2007).

## 2. Alternative Estimators

### 2.1. $GI/M/s$ Model

We now specify the $GI/M/s$ model: The service times are independent and identically distributed (i.i.d.) exponential random variables $V_n$ with mean 1. The interarrival times are i.i.d. positive random variables $U_n$ with a nonlattice cumulative distribution function (cdf) $F$. (We will also consider the deterministic arrival process, which violates this condition; consequently, it will require slightly different analysis.) We omit the subscripts from $U$ and $V$ when the specific index is not important. Let $F$ have finite third moment, characterized by $\nu_3^a \equiv E[U^3]/(E[U])^3$. Then, $F$ necessarily has finite first and second moments. Assume that $E[U] = 1/(s\rho)$, where $s$ is the number of servers and $\rho \equiv E[V]/(sE[U])$ is the traffic intensity. Let $F$ have SCV $c_a^2 \equiv \mathrm{Var}(U)/(E[U]^2)$. Let $A \equiv \{A(t){:}\ t \geq 0\}$ be the renewal counting process (arrival process) associated with $U_n$, defined by

$$A(t) \equiv \max\{n \geq 0{:}\ U_1 + \cdots + U_n \leq t\}, \quad t \geq 0. \tag{4}$$

The $GI/M/s$ system is well known to be stable, and have a proper limiting steady-state behavior, if and only if $\rho < 1$. All our simulation results are for the $GI/M/s$ model in steady state, even though the estimation procedures apply more generally.

### 2.2. No-Information (NI) Steady-State Estimator

The candidate delay estimators differ depending on the information used. If no information at all is used beyond the model, then it is natural to use the steady-state distribution. In particular, with $W_\infty$ denoting the steady-state waiting time before beginning service, the no-information (NI) steady-state delay estimator for a customer who must wait before beginning service is $\theta_{\mathrm{NI}} \equiv E[(W_\infty \mid W_\infty > 0)]$. It serves as a useful reference point. Any other estimator exploiting additional real-time information should do at least as well to be worth serious consideration.

For the $GI/M/s$ model, it is well known that $(W_\infty \mid W_\infty > 0)$ has an exponential distribution—see §XII.3 of Asmussen (2003)—so that the SCV is 1. Since the SCV is 1, the NI estimator is quite highly variable, and so necessarily has low predictive power. For the $M/M/s$ special case, the mean is $1/s(1-\rho)$, so that $\mathrm{MSE} = \mathrm{Var}((W_\infty \mid W_\infty > 0)) = 1/s^2(1-\rho)^2$.

### 2.3. Full-Information QL Delay Estimator

The other extreme would be full information at the arrival epoch, which we take to mean that we know: (i) the queueing model, (ii) the number of customers in the system at that arrival epoch, and (iii) the elapsed service times of all customers in service. If we knew the remaining service times as well, then we could compute the exact delay, but we assume that the remaining service times are unknown. Of course, for exponential service times, the elapsed service times give no useful information about the remaining service times because of the lack-of-memory property of the exponential distribution. Thus the (full-information) QL estimator for the $GI/M/s$ model only exploits the QL $Q(t)$ and knowledge of the model.

Let $W_Q(n)$ represent a random variable with the conditional distribution of the delay of a new arriving customer at some time $t$, given that the arriving customer must wait before starting service and given that the queue length at that time (not counting the new arrival) is $Q(t) = n$. (For $n \geq 1$, the customer must necessarily wait; for $n = 0$, our conditioning implies that all servers are busy but the QL is 0.) For the $GI/M/s$ model, the random variable $W_Q(n)$ can be represented as

$$W_Q(n) \equiv \sum_{i=1}^{n+1} (V_i/s), \tag{5}$$

when $Q(t) = n$. The natural QL delay estimator, based on the observed QL $Q(t) = n$, is the mean $\theta_{\mathrm{QL}}(n) \equiv E[W_Q(n)] = (n+1)/s$. The QL estimator requires knowledge of $s$ and the mean service time $E[V]$ (here taken to be 1) as well as $Q(t)$.

We have the division by $s$ in (5) because 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 1, which makes the minimum exponential with mean $1/s$. It is significant that this estimator is independent of the arrival process, and thus also of the traffic intensity. It applies equally well to steady-state and transient settings.

As discussed in Whitt (1999), $W_Q(n)$ has the desirable property that the estimation gets relatively more accurate as the observed QL $n$ increases:

$$E[W_Q(n)] = \frac{n+1}{s}, \quad \mathrm{Var}[W_Q(n)] = \frac{n+1}{s^2} \quad \text{and}$$
$$c^2_{W_Q(n)} \equiv \frac{\mathrm{Var}[W_Q(n)]}{(E[W_Q(n)])^2} = \frac{1}{n+1}, \tag{6}$$

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

Thus, whenever the QL is large, the QL estimator $E[W_Q(n)]$ will be relatively accurate. If we consider HT regimes, where the QL approaches infinity, as we will do later, then this QL delay estimator will perform well. For example, the half width of a 95% confidence interval is about $2/\sqrt{n}$, which is about 20% of a mean conditional waiting time when $n = 100$. Such a large value of $n$ is not uncommon when $s$ too is large.

For the $M/M/s$ model, there is a simple expression for the average MSE in steady state, which helps judge the performance of other estimators; the MSEs for the other delay estimators should all fall between the QL estimator (best possible) and the NI estimator (worst possible, knowing the model). Let $Q^w_\infty$ be a random variable with the conditional distribution of the steady-state QL upon arrival, given that the customer must wait before beginning service. In the $M/M/s$ model, $Q^w_\infty + 1$ has a geometric distribution with mean $1/(1-\rho)$. That is easily deduced from the time reversibility of the $M/M/s$ model, which implies that $Q^w_\infty$ has the steady-state distribution of the number in the system in an $M/M/1$ queue with traffic intensity $\rho$; e.g., see Proposition 5.6.3 of Ross (1996). Hence

$$E[\mathrm{MSE}(W_Q(Q^w_\infty))] \equiv \sum_{n=0}^{\infty} \mathrm{MSE}(W_Q(n)) P(Q^w_\infty = n)$$
$$= E[\mathrm{Var}(W_Q(Q^w_\infty))] = \frac{1}{s^2(1-\rho)}, \tag{7}$$

so that the ratio between the worst possible NI MSE and the best possible QL MSE is

$$\frac{\mathrm{MSE}(\theta_{\mathrm{NI}})}{\mathrm{MSE}(\theta_{\mathrm{QL}}(Q^w_\infty))} = \frac{\mathrm{Var}(W_\infty \mid W_\infty > 0)}{E[\mathrm{Var}(W_Q(Q^w_\infty))]}$$
$$= \frac{1/s^2(1-\rho)^2}{1/s^2(1-\rho)} = \frac{1}{1-\rho}. \tag{8}$$

For example, a case of principle interest for call centers has $s = 100$ and $\rho = 0.95$. Then, the average MSE for NI is 20 times greater than the average MSE for QL. We will show that the delay-history estimators produce a corresponding ratio of approximately $c^2_a + 1 = 2$.

### 2.4. Last Customer to Enter Service (LES)

The first candidate *direct* delay estimator is the delay (before starting service) of the LES. The direct LES estimator is appealing because it is relatively easy to obtain and interpret, but there also are a variety of *refined* LES estimators we can consider; all are based on the LES observation.

To a large extent, the alternative refined LES delay estimators (and others as well) are obtained by replacing the known QL $n$ in (5) by random variables that estimate the QL, based on the available delay history. Let $W_{\mathrm{LES}}(w, d)$ be the delay of a new arrival, given that the new arrival must wait before starting service and given that the LES experienced delay $w$ before entering service and there was elapsed time $d$ since that customer entered service. Let $t_a$ be the arrival epoch of the new customer and $t_e$ be the time the last customer entered service prior to $t_a$. (Throughout this paper, we use the fact that, almost surely, no two events—arrivals or service completions—will occur simultaneously.) Necessarily, $d = t_a - t_e$ and $t_e - w$ is the arrival epoch of the customer entering service at $t_e$. A key observation is that the QL at time $t_e$ must be distributed as $A(w)$, because customers enter service from the queue in order of arrival. However, $W_{\mathrm{LES}}(w, d)$ has a relatively complicated exact distribution, because we do not know precisely what happens in the interval $[t_e, t_a]$.

If we impose an extra condition, then this random variable $W_{\mathrm{LES}}(w, d)$ has a relatively simple distribution. The *extra condition* is that the epoch $t_e$ is also simultaneously the last service completion prior to $t_a$. That extra condition will necessarily hold if at least one customer remains in the queue at time $t_e$. In turn, that sufficient condition is very likely to be satisfied if $w$ is relatively large (the case of primary interest). Under the extra condition that $t_e$ is also the last service completion before $t_a$, we have the simple representation

$$W_{\mathrm{LES}}(w, d) \equiv \sum_{i=1}^{A(w+d)+1} (V_i/s), \tag{9}$$

where the summands are i.i.d. and independent of $A(w+d)$, because the QL seen by the new arrival at time $t_a$ will be $A(w+d)$, the number of arrivals in the interval of length $w+d$ preceding the arrival epoch $t_a$. Formula (9) allows us to characterize the distribution of $W_{\mathrm{LES}}(w, d)$, under the assumed extra condition. Just like (5), (9) requires knowledge of $s$ and the mean service time as well as $w$. Here, we also require knowledge of the renewal arrival process or, equivalently, the interarrival-time distribution.

An important reference point for the refined LES estimator in (9) is the $D/M/s$ model, with a deterministic arrival process, having constant interarrival times, because under the extra condition leading to (9), we then have $W_{\mathrm{LES}}(w, d) = W_Q(Q(t_a))$, since $A(w+d) = Q(t_a)$, making (5) coincide with (9). Thus we see that the loss of efficiency in going from QL to LES (direct or refined) is primarily because of the variability in the arrival process.

We assume that the experienced LES waiting time $w$ is always available, but we might not know $d$, so that we might want to consider as an alternative refined estimator the mean of the random variable $W_{\mathrm{LES}}(w)$, which assumes $d$ is unavailable, but dropping $d$ makes the distribution even more complicated. If we can assume that $w >> d$, then there should be negligible difference. In general, we have the natural approximations based on (9):

$$W_{\mathrm{LES}}(w) \approx \sum_{i=1}^{A(w+(V_0/s))+1} (V_i/s) \approx \sum_{i=1}^{A(w+(1/s))+1} (V_i/s), \tag{10}$$

where $V_0$ is an exponential random variable with mean 1 independent of $V_i$ for $i \geq 1$, because the time between successive service completions when all servers are busy is distributed as $V_0/s$. (Assuming that the queue is nonempty at time $t_e$, that time is a service completion epoch. Then, $d$ is the age of the Poisson all-servers-busy departure process with rate $s$ under Poisson inspection by the arrival process.) The second approximation is obtained by inserting the expected value. It is also based on the extra condition, which will hold approximately for large $w$.

### 2.5. Head of the Line (HOL) Estimator

A second candidate direct delay estimator, which is closely related to the direct LES estimator, is the elapsed waiting time of the customer at the HOL (queue), assuming that there is at least one customer waiting at the new arrival epoch. The direct HOL delay estimator was used as an announcement in an Israeli bank studied by Mandelbaum et al. (2000) and mentioned as a candidate delay announcement by Nakibly (2002). It is appealing compared to LES because the conditional distribution of the delay to be estimated is more tractable given the HOL information.

The customer at the HOL will enter service after the next service completion. That remaining time is exponential with mean $1/s$. Let $W_{\mathrm{HOL}}(w)$ be a random variable with the conditional distribution of the waiting time (before starting service) of a new arrival given that the new arrival must join the queue, given that there already is at least one customer in queue, and given that the customer at the HOL has already spent time $w$ in queue. The random variable $W_{\mathrm{HOL}}(w)$ is closely related to the random variable $W_{\mathrm{LES}}(w, d)$, but has the advantage that we do not need to use $d$. Moreover, we do not need to impose the extra condition that we made for $W_{\mathrm{LES}}(w, d)$, but, instead, we need to impose a new one: The *extra condition* now is the assumption that there is at least one customer in queue at the arrival epoch $t_a$; otherwise there would be no customer at the head of the line. We propose the random variable $W_{\mathrm{HOL}}(w)$ as an approximation for the random variable $W_{\mathrm{LES}}(w)$, where we omit the lag $d$, as well as for its own sake. Closely paralleling the previous formulas, we have

$$W_{\mathrm{HOL}}(w) \equiv \sum_{i=1}^{A(w)+2} (V_i/s). \tag{11}$$

### 2.6. Delay of the Last Customer to Complete Service (LCS)

A third candidate direct delay estimator is the delay of the 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.

Let $W_{\mathrm{LCS}}(w, v, d)$ be the delay of a new arrival, given that the new arrival must wait before starting service and given that the last customer to complete service experienced delay $w$ before entering service, had individual service time $v$, and there was elapsed time $d$ since that customer completed service. As before, let $t_a$ be the arrival epoch of the new customer; let $t_c$ be the time the last customer completed service prior to $t_a$. The mean of the random variable $W_{\mathrm{LCS}}(w, v, d)$ is a natural refined estimator, but this random variable has a relatively complicated distribution. Some data may be unavailable, so that we may want to consider as alternative refined estimators the means of the random variables $W_{\mathrm{LCS}}(w, d)$, which assumes $v$ is unavailable, and $W_{\mathrm{LCS}}(w)$, which assumes that neither $v$ nor $d$ is available. Dropping $v$ or the pair $(v, d)$ makes the representation even more complicated.

### 2.7. Delay of the Most Recent Arrival to Complete Service (RCS)

Under some circumstances, the LCS and LES direct estimators will be similar, but they actually can be very different when $s$ is large, because the LCS may have experienced his waiting time much before the LES. We emphasize that customers need not depart in order of arrival. Indeed, with exponential service times, when all $s$ servers are busy, each of the $s$ servers is equally likely to generate the next service completion. Thus, for large $s$, the LCS estimator is not really a viable alternative, as we will show. Consequently, we propose other candidate delay estimators based on the delay experience of customers who have already completed service. The first is the delay experienced by the customer who arrived most recently (and thus entered service most recently) among those customers who have already completed service (RCS). We find that RCS is far superior to LCS when $s$ is large.

### 2.8. Among the Last $c\sqrt{s}$ Customers to Complete Service (RCS-$c\sqrt{s}$)

A disadvantage of the RCS estimator is that we must analyze a lot of data, going arbitrarily far back in the past. From HT analysis in §6 and the e-companion, we deduce that the most recent arrival time of a customer who has completed service is very likely to occur among the last $c\sqrt{s}$ customers when $s$ is large (and the system is normally loaded). So we introduce a new estimator, which requires less information processing: Let RCS-$c\sqrt{s}$ be the delay of the customer to have arrived most recently among the last $c\sqrt{s}$ customers who have already completed service. Clearly, these last three estimators LCS, RCS, and RCS-$c\sqrt{s}$ are complicated, so that we primarily rely on simulation to evaluate their relative performance. Through extensive simulation experiments, we found that the ASE of RCS-$c\sqrt{s}$ is essentially identical to that of RCS when $c = 4$, differs by at most 1% when $c = 2$ and differs by at most 10% when $c = 1$.

### 2.9. Averages

Our main estimators are individual delays experienced by a recent customer, rather than an average over many past delays. Only the no-information steady-state estimator $(W_\infty \mid W_\infty > 0)$ can be said to use averages. We can extend the LES, LCS, RCS, and RCS-$c\sqrt{s}$ estimators to get LES-$k$, LCS-$k$, RCS-$k$, and RCS-$c\sqrt{s}-k$ by averaging over the last $k$ experienced delays. With the exception of LCS with large $s$ (which does not have desirable properties), we have found that averages do not help, when the delays are relatively large (the case of primary interest to us). There is a simple explanation: When delays are large, the delays change relatively slowly compared to the size of the delays. Theoretically, this can be explained by the HT snapshot principle; see §6. In this setting, it is better to use recent information than to eliminate noise by averaging.

## 3. Initial Simulation Experiments: Comparing the Estimators

In this section, we present initial simulation experiments, aiming to compare the alternative estimators defined in §2. We focus on the ASE of the estimator, defined in (3). For large samples, the ASE should agree with the MSE in steady state.

Table 1 shows the ASEs for seven different delay estimators in the $GI/M/s$ model with $s = 100$. We consider three categories of estimators: (i) the two reference estimators QL and NI, (ii) the direct delay

**Table 1** Comparison of Efficiency of Different Real-Time Delay Estimators for $GI/M/100$ Queue as a Function of Traffic Intensity $\rho$ and Interarrival-Time Distribution ($M$, $D$, and $H_2$)

| | Estimated ASE in units of $10^{-3}$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $\rho$ | QL | LES | HOL | RCS | RCS-$\sqrt{s}$ | LCS | NI |
| | $M/M/s$ model with $s=100$ | | | | | | |
| 0.98 | 5.03 ±0.02 | 10.2 ±0.05 | 10.2 ±0.05 | 12.5 ±0.05 | 12.9 ±0.05 | 26.7 ±0.06 | 255 ±36 |
| 0.95 | 2.04 ±0.02 | 4.3 ±0.05 | 4.3 ±0.05 | 6.4 ±0.05 | 6.7 ±0.05 | 16.5 ±0.06 | 41.8 ±2.7 |
| 0.93 | 1.44 ±0.002 | 3.07 ±0.003 | 3.08 ±0.003 | 5.06 ±0.003 | 5.32 ±0.003 | 13.1 ±0.13 | 20.8 ±1.2 |
| 0.90 | 0.99 ±0.003 | 2.2 ±0.006 | 2.2 ±0.006 | 3.9 ±0.008 | 4.2 ±0.009 | 9.4 ±0.27 | 9.7 ±0.7 |
| | $D/M/s$ model with $s=100$ | | | | | | |
| 0.98 | 2.48 ±0.05 | 2.62 ±0.05 | 2.62 ±0.05 | 3.77 ±0.05 | 3.94 ±0.05 | 10.3 ±0.11 | 61.5 ±3.9 |
| 0.95 | 1.01 ±0.02 | 1.15 ±0.02 | 1.15 ±0.02 | 2.20 ±0.03 | 2.34 ±0.03 | 6.38 ±0.12 | 10.1 ±0.40 |
| 0.93 | 0.73 ±0.02 | 0.87 ±0.02 | 0.87 ±0.02 | 1.85 ±0.03 | 1.96 ±0.03 | 4.90 ±0.13 | 5.20 ±0.32 |
| 0.90 | 0.52 ±0.015 | 0.67 ±0.016 | 0.66 ±0.017 | 1.54 ±0.035 | 1.63 ±0.037 | 3.44 ±0.15 | 2.68 ±0.23 |
| | $H_2/M/s$ model with $s=100$ | | | | | | |
| 0.98 | 12.4 ±0.70 | 60.4 ±3.2 | 60.4 ±3.2 | 66.1 ±3.2 | 67.0 ±3.2 | 103.4 ±34.0 | 1,505 ±226 |
| 0.95 | 4.82 ±0.095 | 22.5 ±0.46 | 22.5 ±0.47 | 27.7 ±0.45 | 28.4 ±0.45 | 56.3 ±0.58 | 243.3 ±22.7 |
| 0.93 | 3.44 ±0.094 | 15.5 ±0.44 | 15.5 ±0.44 | 20.4 ±0.49 | 21.1 ±0.50 | 44.5 ±1.02 | 121.4 ±10.2 |
| 0.90 | 2.35 ±0.040 | 10.2 ±0.21 | 10.2 ±0.21 | 14.6 ±0.24 | 15.2 ±0.24 | 33.1 ±0.53 | 55.4 ±2.9 |

*Notes.* Only the direct estimators are considered. Estimates of the ASE are shown together with the half width of the 95% confidence interval. The units are $10^{-3}$ throughout.

estimators LES and HOL, and (iii) the three estimators based on delays of customers who have already completed service—LCS, RCS, and RCS-$\sqrt{s}$. We consider three interarrival-time distributions—$M$, $D$, and $H_2$—and four values of the traffic intensity $\rho$—0.98, 0.95, 0.93, and 0.90. The $H_2$ distribution has SCV $c_a^2 = 4$ and balanced means (the two component exponential distributions contribute equally to the mean). We performed 10 independent replications of long runs in each case. The half width of the 95% confidence interval is shown below each estimate. Corresponding results for other values of $s$—1, 10, 400, and 900—are contained in the online supplement, Ibrahim and Whitt (2007). The cases $s = 10$ and $s = 1$ are shown in the e-companion.

These estimators appear in Table 1 with the better performance toward the left; i.e., in terms of efficiency (low ASE), the estimators are ordered by

$$\mathrm{QL} > \mathrm{LES} \approx \mathrm{HOL} > \mathrm{RCS} \approx \mathrm{RCS} - \sqrt{s} > \mathrm{LCS} > \mathrm{NI}. \quad (12)$$

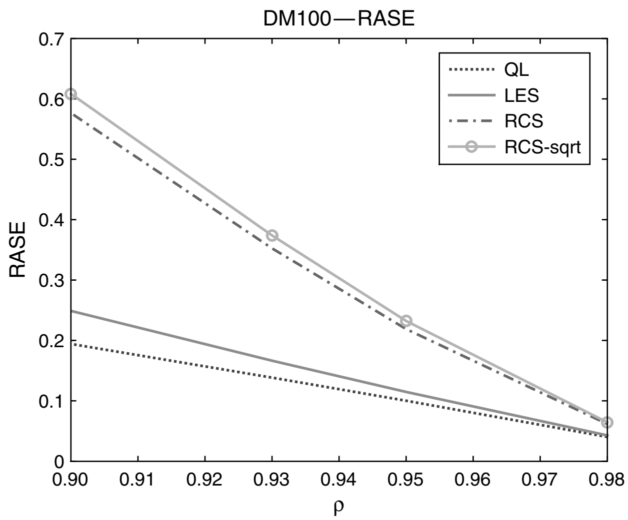
As expected, the full-information QL estimator performs best, while the no-information NI estimator performs worst. The performance of LES and HOL are very close, while the performance of RCS and RCS-$\sqrt{s}$ are very close. The QL estimator is significantly better than LES; LES is slightly better than RCS; RCS is significantly better than LCS; and LCS is significantly better than NI. Very roughly, $\mathrm{ASE(LES)/ASE(QL)} \approx (c_a^2 + 1)/\rho$, so LES performs nearly as well as QL for low-variability arrival processes such as the $D$ arrival process, but much worse for high-variability arrival processes such as the $H_2$ arrival process.

It is instructive to look at the relative average squared error (RASE), which is obtained by dividing the ASE by $E[W_\infty \mid W_\infty > 0]^2$, because the associated steady-state relative mean squared error (RMSE), defined as $\mathrm{MSE}/E[W_\infty \mid W_\infty > 0]^2$, is *linear* as a function of $\rho$ for the QL estimator: $\mathrm{RMSE(QL)} = (1-\rho)$. (The RMSE is identically 1 for the NI estimator.) We show the RASE plots for the $D/M/100$ model in Figure 1. With the $D$ arrival process, LES and HOL are virtually identical (with the plots lying on top of each other), so we only show LES. Both LES and HOL are nearly as good as QL and much better than RCS; LCS is so bad that it is not even shown. Corresponding plots for other interarrival-time distributions and other $s$ appear in the online supplement. The plots for the $M/M/100$ and $H_2/M/100$ models are in the e-companion.

Experience shows that the NI estimator performs especially poorly in very HT, while LCS performs especially poorly with large $s$ in light traffic. For large $s$ and small $\rho$, LCS even performs worse than the NI estimator. There is only one case in Table 2; more cases can be seen when $s = 400$ and $s = 900$ in the supplement.

Because delay estimates are more relevant when the observed delays in the system are longer, it

**Figure 1** ***RASE* for $D/M/100$ Model**




is natural to consider the behavior of the estimators for larger delays. We have complemented the experiments described above by considering how the delay estimators perform when we only consider actual delays that fall in one of the intervals: $(E[W \mid W>0], 2E[W \mid W>0])$, $(2E[W \mid W>0], 4E[W \mid W>0])$, $(4E[W \mid W>0], 6E[W \mid W>0])$, and $(6E[W \mid W>0], \infty)$. Table 2 illustrates the results for the $M/M/100$ model when the observed delays fall in the interval $(4E[W \mid W>0], 6E[W \mid W>0])$. Other cases appear in the online supplement. The performance of the estimators for these larger delays is

**Table 2** **Comparison of Efficiency of Different Real-Time Delay Estimators Conditional on Level of Delay Experienced for $M/M/100$ Model as a Function of Traffic Intensity $\rho$**

| $\rho$ | QL | LES | HOL | RCS | RCS-$\sqrt{s}$ | LCS | NI |
|---|---|---|---|---|---|---|---|
| 0.99 | 49.4<br>±7.0 | 86.6<br>±6.9 | 86.3<br>±6.9 | 89.4<br>±7.2 | 90.1<br>±7.2 | 108.8<br>±10.2 | 11,586<br>±1,250 |
| 0.98 | 24.8<br>±1.8 | 47.5<br>±3.1 | 47.3<br>±3.0 | 50.1<br>±3.1 | 50.6<br>±3.1 | 69.6<br>±3.7 | 3,542<br>±431 |
| 0.95 | 10.5<br>±0.23 | 20.4<br>±0.63 | 20.1<br>±0.62 | 23.5<br>±0.82 | 24.0<br>±0.80 | 50.4<br>±3.3 | 564<br>±27 |
| 0.93 | 7.54<br>±0.20 | 15.2<br>±0.31 | 14.9<br>±0.29 | 18.7<br>±0.43 | 19.3<br>±0.45 | 52.0<br>±3.2 | 286<br>±8.0 |
| 0.90 | 5.62<br>±0.21 | 11.1<br>±0.38 | 10.7<br>±0.38 | 15.3<br>±0.61 | 16.1<br>±0.66 | 50.9<br>±25.2 | 137.4<br>±6.7 |

*Notes.* Actual delays are considered that fall in the interval $(4E[W \mid W>0], 6E[W \mid W>0])$. Estimates of the conditional ASE are shown together with the half width of the 95% confidence interval. The units are $10^{-3}$ throughout.

approximately as in Table 1. As should be expected, the NI estimator fares even worse in this comparison.

## 4. Analysis of HOL and LES Estimators

The representation (11) allows us to characterize the probability distribution of the random variable $W_{\mathrm{HOL}}(w)$, which we do both for its own sake and as an approximation for the random variables $W_{\mathrm{LES}}(w)$ and $W_{\mathrm{RCS}}(w)$. When we use the HOL estimator, we assume that there is at least one customer in queue at the new arrival epoch $t_a$. Very similar formulas hold for the LES estimator based on formula (9), under the extra assumption given there. Because the formulas are virtually identical, we do not display separate results for LES.

We emphasize that the random variable $W_{\mathrm{HOL}}(w)$ applies to both transient and steady-state scenarios. We can have arbitrary traffic intensity $\rho$, including $\rho > 1$, under which there is no proper steady state. We assume that the renewal arrival process $\{A(t): t \geq 0\}$ and the traffic intensity $\rho$ are specified and unchanging in the interval $[t_a - w, t_a]$, which is the relevant system history for our estimation at time $t_a$.

We start by showing that the distribution of $W_{\mathrm{HOL}}(w)$ depends on $s$ in a relatively simple way. For that purpose, we introduce an extra subscript $s$ to indicate the dependence on $s$, getting $W_{\mathrm{HOL},s}(w)$. Let $\stackrel{\mathrm{d}}{=}$ denote equality in distribution.

Theorem 4.1 (Dependence on $s$). *For the $GI/M/s$ model,*

$$W_{\mathrm{HOL},s}(w) \stackrel{\mathrm{d}}{=} \frac{W_{\mathrm{HOL},1}(sw)}{s} \tag{13}$$

*for all $\rho$, $w$, and $s$.*

Proof. We show the equality in distribution by establishing equality w.p.1 for a special construction. We construct a convenient family of systems indexed by $s$. For each $s$, let the service times be exponential random variables $V_n$ with mean 1 as before. Start by defining interarrival times $U_n$ with mean $1/\rho$ to use for the case of $s=1$. Then, in the system with $s>1$, let the $n$th interarrival time be $U_n/s$. Let $\{A_s(t): t \geq 0\}$ be the renewal counting process in system $s$, having interarrival times $U_n/s$. Then, $A_s(w/s) = A_1(w)$ for all $s$ and $w$; since we have rescaled the interarrival times, we just rescale time in the associated renewal counting process. This construction yields equality for the random variables in (13) and all $w \geq 0$. Because the distribution is independent of the construction, that implies the claimed relation (13). □

We now show that we get relatively simple asymptotic expressions characterizing the distribution of $W_{\mathrm{HOL},s}(w)$ when $sw \to \infty$. That applies when $w \to \infty$ for fixed $s$, but it also can apply when $s \uparrow \infty$ and $w \downarrow 0$, as occurs in the quality-and-efficiency-driven (QED) many-server HT limiting regime, to be discussed in §6; then $w = O(1/\sqrt{s})$, so that $sw \to \infty$ while $w \to 0$.

Let $N(m, \sigma^2)$ denote a normally distributed random variable with mean $m$ and variance $\sigma^2$. Let $\Rightarrow$ denote convergence in distribution.

Theorem 4.2 (Distribution of $W_{\mathrm{HOL},s}(w)$). *Consider the $GI/M/s$ queue with traffic intensity $\rho$ operating in the time interval $[t_a - w, t_a]$.*

(a) *For any $\rho > 0$, $s \geq 1$, and $w > 0$,*

$$E[W_{\mathrm{HOL},s}(w)] = \frac{E[A(w)] + 2}{s} \tag{14}$$

*and*

$$\mathrm{Var}[W_{\mathrm{HOL},s}(w)] = E[A(w)+2]\mathrm{Var}(V/s) + \mathrm{Var}(A(w)+2)(E[V/s])^2. \tag{15}$$

(b) *If the arrival process is Poisson, then*

$$E[W_{\mathrm{HOL},s}(w)] = \rho w + \frac{2}{s} \tag{16}$$

*and*

$$\mathrm{Var}[W_{\mathrm{HOL},s}(w)] = \frac{2\rho w}{s} + \frac{2}{s^2}, \tag{17}$$

*so that*

$$c^2_{W_{\mathrm{HOL},s}(w)} = \frac{2}{\rho s w} - \frac{6}{(\rho s w)^2} + O\left(\frac{1}{(\rho s w)^3}\right) \quad \text{as } sw \to \infty. \tag{18}$$

(c) *For a general renewal arrival process with a nonlattice interrenewal-time distribution if $sw \to \infty$, then*

$$sE[W_{\mathrm{HOL},s}(w)] - \rho s w \to \frac{(c_a^2 + 3)}{2}, \tag{19}$$

$$\frac{W_{\mathrm{HOL},s}(w)}{w} \to \rho \quad \text{w. p. 1}, \quad \text{and} \qquad \frac{E[W_{\mathrm{HOL},s}(w)]}{w} \to \rho, \tag{20}$$

$$s^2\mathrm{Var}(W_{\mathrm{HOL},s}(w)) - \rho s w(c_a^2 + 1) \to \left(\frac{5(c_a^2+1)^2}{4} - \frac{2\nu_a^3}{3} + 1\right), \tag{21}$$

$$s^2 E[(W_{\mathrm{HOL},s}(w) - \rho w)^2] - \rho s w(c_a^2 + 1) \to K, \tag{22}$$

$$s^2 E[(W_{\mathrm{HOL},s}(w) - w)^2] - (sw)^2(1-\rho)^2 - sw[(2\rho - 1)c_a^2 + 4\rho - 3] \to K, \tag{23}$$

*where*

$$K \equiv K(c_a^2, \nu_a^3) \equiv \left(\frac{3c_a^4}{2} + 4c_a^2 + \frac{9}{2} - \frac{2\nu_a^3}{3}\right), \tag{24}$$

$$swc^2_{W_{\mathrm{HOL},s}(w)} \to \frac{c_a^2+1}{\rho} \quad \text{and} \qquad \frac{W_{\mathrm{HOL},s}(w) - \rho w}{\sqrt{\rho w(c_a^2+1)/s}} \Rightarrow N(0,1). \tag{25}$$

Proof. Since $W_{\mathrm{HOL}}(w)$ in (11) is a random sum of i.i.d. random variables, where $A(w)$ is independent of the summands $V_i/s$, we have (14). Formula (15) follows from the conditional variance formula, e.g., Ross (1996, p. 51). For (18), we use elementary operations on series, as in 3.6.22 in Abramowitz and Stegun (1972). When we let $sw$ increase, we first apply Theorem 4.1 to reduce the analysis to the case $s = 1$. Henceforth assume that $s = 1$. When we restrict attention to $s = 1$, it suffices to simply let $w \to \infty$. When we let $w$ increase,

$$E[A(w)+2] - \rho w \to \frac{(c_a^2+1)}{2} + 1 \quad \text{as } w \to \infty, \tag{26}$$

see Corollary 3.4.7 of Ross (1996) or (10) and (11) of Whitt (1982), which review a classic result. Combining (26) and (14) gives (19), which immediately implies the second limit in (20). For the w.p.1 limit in (20), we apply the strong law of large numbers for the partial sums of $V_n$ and the renewal arrival process $A(w)$: With probability one,

$$\frac{\sum_{i=1}^{n} V_i}{n} \to E[V] = 1 \quad \text{and} \qquad \frac{A(w)+2}{w} \to \frac{1}{E[U]} = \rho, \tag{27}$$

so that

$$\frac{\sum_{i=1}^{A(w)+2} V_i}{w} = \frac{A(w)+2}{w} \times \frac{\sum_{i=1}^{A(w)+2} V_i}{A(w)+2} \to \rho \quad \text{w. p. 1.} \tag{28}$$

The asymptotic variance formula (21) follows from (15) and the asymptotic form of the variance for a renewal process, e.g., as in (10) and (11) of Whitt (1982):

$$\begin{aligned}\mathrm{Var}(A(w)+2) &= \mathrm{Var}(A(w))\\ &= \rho w c_a^2 + \frac{5(c_a^2+1)^2}{4} - \frac{2\nu_a^3}{3}\\ &\quad - \frac{(c_a^2+1)}{2} + o(1) \quad \text{as } w \to \infty. \end{aligned} \tag{29}$$

The associated limits (22) and (23) follow from (21). For (22), we use

$$\begin{aligned} &E[(W_{\mathrm{HOL},s}(w) - \rho w)^2]\\ &\quad = \mathrm{var}(W_{\mathrm{HOL},s}(w) - \rho w) + (E[W_{\mathrm{HOL},s}(w) - \rho w])^2\\ &\quad = \mathrm{var}(W_{\mathrm{HOL},s}(w)) + (E[W_{\mathrm{HOL},s}(w) - \rho w])^2. \end{aligned} \tag{30}$$

The calculation for (23) is similar. The first limit in (25) follows immediately from (19) and (21). The central limit theorem in (25) follows from the central limit theorem for renewal-reward processes, e.g., Whitt (2002, Theorem 7.4.1). We use the convergence-together theorem, Theorem 11.4.7 of Whitt (2002), to justify neglecting the asymptotically negligible terms. □

Remark 4.1 (Exact Values by Numerical Inversion). It is possible to exploit (14) and (15) to compute the exact means and variances. To do so, we can exploit numerical transform inversion of Laplace transforms, as discussed in §13 of Abate and Whitt (1992). The Laplace transform of $E[A(t)]$ is $\widehat{m}_1(s) \equiv \hat{f}(s)/[s(1-\hat{f}(s))]$, where $\hat{f}(s)$ is the Laplace transform of the density function of the interarrival-time cdf $F$ (here assumed to exist). The associated Laplace transform of $E[A(t)^2]$ is $2\widehat{m}_1(s)^2 - \widehat{m}_1(s)$, as can be seen from exercise XI.13 on p. 386 of Feller (1971). Because we are interested in estimation for relatively large delays, we will rely on the asymptotic approximations.

Remark 4.2 (Nonhomogeneous Poisson Arrival Process). We can also analyze the random variable $W_{\mathrm{HOL},s}(w)$ in the case of a nonhomogeneous Poisson arrival process with intensity function $\{\lambda(t)\colon t \ge 0\}$. The exact relations (16) and (17) have natural extensions to that case. We again have representation (11), but now with $A(w)$ being a Poisson random variable having mean

$$m_a(w) \equiv \int_{t_a - w}^{t_a} \lambda(t)\, dt, \tag{31}$$

which depends on the arrival time $t_a$ and the intensity function as well as the experienced waiting time $w$. Unless we specify how the intensity function behaves, we have no simple asymptotic story as $w$ increases though.

Theorem 4.2 shows that the first-order asymptotic behavior of the random variable $W_{\mathrm{HOL},s}(w)$ as $sw$ increases depends on the general interarrival-time distribution $F$ only through its first two moments or, equivalently, through the mean $E[U] = 1/\rho s$ and the SCV $c_a^2$. Equations (21) and (25) show that both the variance $\mathrm{Var}(W_{\mathrm{HOL},s}(w))$ and the SCV $c^2_{W_{\mathrm{HOL},s}(w)}$ are approximately proportional to $c_a^2 + 1$ for large $sw$.

Theorem 4.2 shows that it may be useful to consider various refined estimators instead of the direct estimator $\theta^d_{\mathrm{HOL}} \equiv w$. We would want to use the refined estimator $\theta^r_{\mathrm{HOL}} \equiv E[W_{\mathrm{HOL},s}(w)]$, because the mean necessarily minimizes the MSE, but we do not have a convenient formula for the mean. Theorem 4.2 leads us to consider two other refined estimators: the *simple refined estimator* $\theta^{sr}_{\mathrm{HOL}} \equiv \rho w$ and the *asymptotic refined estimator* $\theta^{ar}_{\mathrm{HOL}} \equiv \rho w + (c_a^2 + 3)/(2s)$, based on the the limit (19) as $sw \to \infty$. Note that the formulas for the mean and variance for Poisson arrivals in (16) and (17) are exact, whereas the formulas for non-Poisson formulas are only approximations.

For fixed $\rho < 1$, the three refined estimators $\theta^r_{\mathrm{HOL}}(w)$, $\theta^{sr}_{\mathrm{HOL}}(w)$, and $\theta^{ar}_{\mathrm{HOL}}(w)$ are all relatively consistent and asymptotically relatively efficient as $sw \to \infty$, whereas the direct HOL estimator $w$ has neither of these properties. By *relatively consistent*, we mean that the ratio of the estimator to the quantity being estimated (here $W_{\mathrm{HOL},s}(w)$) converges to 1; by *asymptotically relatively efficient*, we mean that the RMSE ($\mathrm{RMSE} \equiv \mathrm{MSE}/\mathrm{Mean}^2$) converges to 0.

At first glance, the simple refined estimator looks very appealing, because it combines simplicity with good asymptotic properties. However, we found that the direct estimator consistently outperforms the simple refined estimator in experiments evaluating the steady-state performance for typical parameter values. Evidently, the extra constant term in $\theta^{ar}_{\mathrm{HOL}}$ helps. The following (somewhat loosely stated) theorem supports that empirical observation. Let $\mathrm{MSE}(\theta_{\mathrm{HOL}}(W_\infty))$ denote the steady-state MSE of the estimator $\theta_{\mathrm{HOL}}(w)$ when $w$ is averaged with respect to the conditional delay $(W_\infty \mid W_\infty > 0)$, where $W_\infty$ is the steady-state delay.

Theorem 4.3 (Comparison of Alternative HOL Estimators). *Consider the $GI/M/s$ queue with traffic intensity $\rho < 1$ in steady state. If the arrival process is Poisson or if we take the limit in (19) as the exact mean, then the steady-state MSEs are ordered by*

$$\mathrm{MSE}(\theta^{ar}_{\mathrm{HOL}}(W_\infty)) < \mathrm{MSE}(\theta^{d}_{\mathrm{HOL}}(W_\infty)) < \mathrm{MSE}(\theta^{sr}_{\mathrm{HOL}}(W_\infty)). \tag{32}$$

*Moreover,*

$$\begin{aligned}&\mathrm{MSE}(\theta^{d}_{\mathrm{HOL}}(W_\infty)) - \mathrm{MSE}(\theta^{ar}_{\mathrm{HOL}}(W_\infty))\\ &\quad = E\left[\left((1-\rho)(W_\infty \mid W_\infty > 0) - \frac{(c_a^2+3)}{2s}\right)^2\right] < \frac{(c_a^2+3)^2}{4s^2}\\ &\quad = \mathrm{MSE}(\theta^{sr}_{\mathrm{HOL}}(W_\infty)) - \mathrm{MSE}(\theta^{ar}_{\mathrm{HOL}}(W_\infty)).\end{aligned} \tag{33}$$

Proof. The MSE formulas in (33) are obtained by directly adding and subtracting the mean inside the MSE formula, with the mean here regarded as being given by (19). The key inequality in (33) follows from a bound on the mean steady-state waiting time in the $GI/M/1$ queue. The conditional delay $(W_\infty \mid W_\infty > 0)$ in the $GI/M/s$ model has the same exponential distribution as in the $GI/M/1$ model; e.g., see Wolff (1989, p. 398). Its mean is $(1-\omega)^{-1}$, where $\omega$ is the root of the transform equation $\hat{f}(1-\omega) = \omega$, where $\hat{f}(s)$ is the Laplace-Stieltjes transform of the interarrival-time cdf. However, it is known that $1-\omega > 2(1-\rho)/(c_a^2+1)$; e.g., apply Theorem 2 of Whitt (1984), noting that in the $D/M/1$ queue $1-\omega > 2(1-\rho)$, which follows from elementary inequalities for the exponential function: $e^{-2(1-\rho)} \geq 1 - 2(1-\rho)$. From (33), we see that $\mathrm{MSE}(\theta^{d}_{\mathrm{HOL}}(W_\infty)) < \mathrm{MSE}(\theta^{sr}_{\mathrm{HOL}}(W_\infty))$ if and only if

$$E\left[\left((1-\rho)(W_\infty \mid W_\infty > 0) - \frac{(c_a^2+3)}{2s}\right)^2\right] < \frac{(c_a^2+3)^2}{4s^2}, \tag{34}$$

which, upon expanding the quadratic and using the fact that the second moment is twice the square of the first moment, holds if and only if

$$E[W_\infty \mid W_\infty > 0] < \frac{c_a^2+3}{s(1-\rho)}, \tag{35}$$

which is implied by the delay bound. □

To illustrate, we show numerical results in Table 3 for the candidate delay estimators $\theta^{d}_{\mathrm{HOL}}$, $\theta^{sr}_{\mathrm{HOL}}$, and $\theta^{ar}_{\mathrm{HOL}}$ in the $H_2/M/s$ model with $s = 100$ and $s = 1$. We display the values of their approximate MSEs in steady state predicted by formulas (23), (22), and (21), and we show the contributing terms, displayed in the order given in Theorem 4.2. In each case, one term grows without bound as $\rho$ increases, while the other terms remains constant or nearly constant. We take the expected value of each MSE formula, where $w$ is distributed randomly as the steady-state conditional delay $(W_\infty \mid W_\infty > 0)$. We use the simulation estimates of the first two moments of the conditional delay. Table 3 is consistent with Theorem 4.3. As a consequence of Theorem 4.3, we suggest using the asymptotic refined estimator $\theta^{ar}_{\mathrm{HOL}}$.

We remark that the limit in (25) implies that $W_{\mathrm{HOL},\,s}(w)$ should be approximately normally distributed when $sw$ is not too small. Our simulation experiments show that all the random variables $W_{\mathrm{HOL},\,s}(w)$, $W_{\mathrm{LES},\,s}(w)$, and $W_{\mathrm{RCS},\,s}(w)$ tend to be normally distributed when $sw$ is not too small.

We can combine (25) and (6) to compare the efficiency of the QL and refined HOL estimators under high congestion. Let $W(t)$ be the virtual waiting time at time $t$, the time an arrival at time $t$ would have to wait before beginning service. Since

$$W(t) = \sum_{i=1}^{Q(t)+1} (V_i/s), \tag{36}$$

the law of large numbers implies that $W(t)/Q(t) \to 1/s$ as $Q(t) \to \infty$. Thus, when $Q(t)$ is large, we have $W(t) \approx Q(t)/s$ (even if $W(t)$ itself is not large). Assuming that $n$ is large with $w \approx n/s$ in (25) and (6), we have both $sw$ and $n$ large and

$$\frac{c^2_{W_{\mathrm{HOL},\,s}(w)}}{c^2_{W_{Q,\,s}(n)}} \approx \frac{(c_a^2+1)/\rho s w}{1/(n+1)} \approx \frac{c_a^2+1}{\rho}. \tag{37}$$

Since we have introduced HOL partly as an approximation for LES, it is interesting to consider the difference between the HOL and LES observed delays and the difference between the random variables $W_{\mathrm{HOL},\,s}(w)$ and $W_{\mathrm{LES},\,s}(w, d/s)$. (We let $t_a - t_e = d/s$ because it should be proportional to $1/s$ with $s$ servers.) First, note that if at least one customer remains in queue after the LES at time $t_e$, then the HOL customer at time $t_e$ (after the customer entered service) will remain the HOL customer at time $t_a$. As a consequence, the HOL customer arrived immediately

**Table 3** Evaluation of MSE Approximations for Estimators $\theta^{d}_{\mathrm{HOL}}$, $\theta^{sr}_{\mathrm{HOL}}$, and $\theta^{ar}_{\mathrm{HOL}}$ in Steady State Using (23), (21), and (22) Together with Simulation Estimates of First Two Moments of Conditional Delay $E[W_\infty \mid W_\infty > 0]$

| Evaluating alternative HOL estimators | | | | |
|---|---|---|---|---|
| $\rho$ | 0.88 | 0.92 | 0.96 | 0.98 |
| Approximations in $H_2/M/100$ model | | | | |
| $E[W \mid W > 0]$ | 0.1902 | 0.2964 | 0.6114 | 1.307 |
| conf. int. | ±0.0030 | ±0.0067 | ±0.029 | ±0.17 |
| $E[W^2 \mid W > 0]$ | 0.07205 | 0.1761 | 0.7446 | 3.436 |
| conf. int. | ±0.0022 | ±0.0095 | ±0.060 | ±0.67 |
| MSE($\theta^{d}$) | 0.00826 | 0.0135 | 0.0293 | 0.0640 |
| Term 1 | 0.00103 | 0.00113 | 0.00119 | 0.00137 |
| Term 2 | 0.00677 | 0.0120 | 0.0276 | 0.0622 |
| Term 3 | 0.00045 | 0.00045 | 0.00045 | 0.00045 |
| MSE($\theta^{sr}$) | 0.00882 | 0.00141 | 0.00298 | 0.0645 |
| Term 1 | 0.00837 | 0.0136 | 0.0293 | 0.0640 |
| Term 2 | 0.00045 | 0.00045 | 0.00045 | 0.00045 |
| MSE($\theta^{ar}$) | 0.00759 | 0.0129 | 0.0286 | 0.0632 |
| Term 1 | 0.00837 | 0.0136 | 0.0293 | 0.0640 |
| Term 2 | −0.000775 | −0.000775 | −0.000775 | −0.000775 |
| $\rho$ | 0.85 | 0.90 | 0.95 | 0.98 |
| Approximations in $H_2/M/1$ model | | | | |
| $E[W \mid W > 0]$ | 15.01 | 23.50 | 48.64 | 115.7 |
| conf. int. | ±0.18 | ±0.42 | ±1.6 | ±8.80 |
| $E[W^2 \mid W > 0]$ | 446.2 | 1,105.7 | 4,707.1 | 25,650.5 |
| conf. int. | ±8.03 | ±39.2 | ±263.2 | ±3,280 |
| MSE($\theta^{d}$) | 62.59 | 104.9 | 230.3 | 565.7 |
| Term 1 | 10.04 | 11.06 | 11.76 | 10.26 |
| Term 2 | 48.04 | 89.3 | 214.0 | 550.9 |
| Term 3 | 4.5 | 4.5 | 4.5 | 4.5 |
| MSE($\theta^{sr}$) | 68.31 | 110.3 | 235.5 | 571.6 |
| Term 1 | 63.81 | 105.8 | 231.0 | 567.1 |
| Term 2 | 4.5 | 4.5 | 4.5 | 4.5 |
| MSE($\theta^{ar}$) | 56.06 | 98.02 | 223.3 | 559.3 |
| Term 1 | 63.81 | 105.8 | 231.0 | 567.1 |
| Term 2 | −7.75 | −7.75 | −7.75 | −7.75 |

*Note.* The $H_2/M/s$ model is considered as a function of the traffic intensity $\rho$ for $s = 100$ and $s = 1$.

after the LES customer. Thus the HOL customer waits more than the LES customer by the time $t_a - t_e$ but less by the single interarrival time between them. Clearly, these differences should become asymptotically negligible in the appropriate scaling.

We now compare the random variables $W_{\mathrm{HOL},s}(w)$ and $W_{\mathrm{LES},s}(w, d)$. We establish a stochastic bound between these random variables. Let $\le_{st}$ denote ordinary stochastic order; see §9.1 of Ross (1996). The following bound shows that the difference between $W_{\mathrm{HOL},s}(w)$ and $W_{\mathrm{LES},s}(w, d)$ is stochastically bounded, and thus asymptotically negligible compared to $w$ and these individual random variables as $sw \to \infty$. We say that a family of random variables $\{X(w)\colon w > 0\}$ is *stochastically bounded* if for any $\epsilon > 0$ there exists a positive constant $K(\epsilon)$ such that $P(|X(w)| > K(\epsilon)) < \epsilon$. By Markov's inequality, for nonnegative random variables, it suffices to have the means $E[X(w)]$ uniformly bounded: $P(|X(w)| > K(\epsilon)) \le E[X(w)]/K(\epsilon)$.

Theorem 4.4 (Bound on Difference Between $W_{\mathrm{HOL},s}(w)$ and $W_{\mathrm{LES},s}(w, d/s)$). *Consider the $GI/M/s$ model. Assume that there is at least one customer in queue at the new arrival epoch, so that* (11) *is valid for HOL and* (9) *is valid for LES. Then,*

$$W_{\mathrm{LES},s}(w, d/s) - X(s, w, d) \le_{st} W_{\mathrm{HOL},s}(w) \le_{st} W_{\mathrm{LES},s}(w, d/s) + X(s, w, d), \tag{38}$$

*where $X(s, w, d)$ is distributed as*

$$X(s, w, d) \equiv \sum_{i=1}^{A(w+(d/s))-A(w)+1} (V_i/s). \tag{39}$$

*As $w \to \infty$ for fixed $s$, $E[X(s, w, d)] \to (\rho d + 1)/s$; as $sw \to \infty$, $E[X(s, w, d)]/w \to 0$, so that*

$$\frac{|W_{\mathrm{HOL}}(w) - W_{\mathrm{LES}}(w, d)|}{w} \to 0 \quad \text{as } sw \to \infty. \tag{40}$$

*For the $M/M/s$ model,*

$$X(s, w, d) = \sum_{i=1}^{A(d/s)+1} (V_i/s), \tag{41}$$

*so that*

$$E[X(s, w, d)] = (\rho d + 1)/s \quad \text{and}$$
$$\mathrm{Var}(X(s, w, d)) = (2\rho d + 1)/s^2. \tag{42}$$

Proof. Without altering the individual distributions of $W_{\mathrm{HOL},s}(w)$ and $W_{\mathrm{LES},s}(w, d/s)$, we can make a special construction in which we use exactly the same exponential random variables $V_i/s$ for the two estimators. The random numbers of summands differ by $A(w + (d/s)) - A(w) - 1$, which is bounded above by $A(w + (d/s)) - A(w) + 1$, which we use in (39). Since the renewal process $A$ has rate $\rho s$, we can then apply Blackwell's renewal theorem, Asmussen (2003, p. 155), to get $E[A(w + d/s) - A(w)] \to \rho d$ as $sw \to \infty$. Recall that we have assumed that the interarrival time cdf $F$ is nonlattice. Hence we get $E[X(s, w, d)]/w \to 0$ as $sw \to \infty$, which implies (40). □

## 5. Simulations Related to Theorem 4.2

Based on (23) in Theorem 4.2, we approximate the MSE of the direct HOL, LES, and RCS estimators by

$$\mathrm{MSE}(\theta^d_{\mathrm{HOL}}(w)) \approx (1-\rho)^2 w^2 + \frac{((2\rho-1)c_a^2+4\rho-3)w}{s} + \frac{K}{s^2} \quad (43)$$

for $K$ in (24). As above, let $\mathrm{MSE}(\theta^d_{\mathrm{HOL}}(W_\infty))$ denote the MSE in steady state, i.e., when we replace $w$ in (43) by $(W_\infty \mid W_\infty > 0)$. We obtain

$$\mathrm{MSE}(\theta^d_{\mathrm{HOL}}(W_\infty)) \approx (1-\rho)^2 E[W_\infty^2 \mid W_\infty > 0] + \frac{((2\rho-1)c_a^2+4\rho-3)E[W_\infty \mid W_\infty > 0]}{s} + \frac{K}{s^2}, \quad (44)$$

where $W_\infty$ is the steady-state delay.

We have compared the ASE for HOL, LES, and RCS to $\mathrm{MSE}(\theta^d_{\mathrm{HOL}}(W_\infty))$ and found close agreement, with the agreement being slightly better for HOL and LES than for RCS. In making this comparison, we substitute the simulation estimates of the two moments $E[W_\infty \mid W_\infty > 0]$ and $E[W_\infty^2 \mid W_\infty > 0]$ into (44). We must calculate or approximate these conditional moments to have a full approximation, but we do not consider that step here. We obtain good results comparing approximation (44) to the ASE for the cases of exponential ($M$), hyperexponential ($H_2$ with $c_a^2 = 4$), and Erlang ($E_2$) interrenewal-time distributions. We did experiments for $s = 1$, 10, 100, 400, 900, each for four values of $\rho$, increasing with $s$ to represent typical cases. The errors were consistently less than 5% for HOL and LES in these experiments, as illustrated by the results for LES with $M$ and $H_2$ interarrival-time distributions in Table 4.

We found that the approximation in (44) does not perform nearly as well for the case of a deterministic ($D$) arrival process, which should not be surprising, because the deterministic interrenewal-time distribution is a lattice distribution not covered by Theorem 4.2. Instead of (43), we propose the following approximation for the direct estimator with a $D$ arrival process:

$$\mathrm{MSE}(\theta^d_{\mathrm{HOL},D}(w)) \approx (1-\rho)^2 w^2 + \frac{\rho w + (2/s)}{s}, \quad (45)$$

**Table 4** **Evaluation of Approximations for Steady-State MSE of HOL in (44) and (46) by Comparing to Simulation Estimates of ASE for LES in $GI/M/100$ Model as a Function of Interarrival-Time Distribution and Traffic Intensity $\rho$**

| Testing the MSE(HOL$_\infty$) approximations in the $GI/M/100$ model | | | | | | |
|---|---|---|---|---|---|---|
| $\rho$ | $M$ | % diff. (%) | % diff. $D$ | (%) | $H_2$ | % diff. (%) |
| 0.98 | 10.20 | −0.3 | 2.67 | −1.9 | 62.8 | −3.9 |
| 0.95 | 4.20 | 1.4 | 1.20 | −4.1 | 22.9 | −1.9 |
| 0.93 | 3.06 | 0.4 | 0.92 | −5.8 | 15.9 | −2.1 |
| 0.90 | 2.20 | −1.5 | 0.72 | −7.5 | 10.5 | −3.2 |

*Notes.* The simulation estimates appear in Table 1. The approximations in units of $10^{-3}$ and the relative percent differences are shown here.

which is obtained by making the simple approximation $A(w) \approx \rho s w$. We then obtain the following analog of the steady-state approximation (44):

$$\mathrm{MSE}(\theta^d_{\mathrm{HOL},D}(W_\infty)) \approx (1-\rho)^2 E[W_\infty^2 \mid W_\infty > 0] + \frac{\rho E[W_\infty \mid W_\infty > 0] + (2/s)}{s}. \quad (46)$$

Approximation (46) performs much better than approximation (44) with $c_a^2 = 0$, yielding errors of about 5% (ranging up to 11%), instead of about 5%–25%, as shown in Table 4. For the refined estimator, we would also change the mean estimator to (16) instead of (19).

To evaluate the approximations for a specified observed delay $w$, we consider data from the simulation where the observed HOL delay falls in a small interval about $w \equiv 2E[W_\infty \mid W_\infty > 0]$. (We choose interval widths to make roughly reasonable, comparable sample sizes.) Table 5 shows the results of such an experiment for the $GI/M/100$ model with $\rho = 0.95$. (The width of the sampling interval in each case was chosen to have roughly comparable sample sizes.) Table 5 shows that the approximations for the HOL conditional mean and variance are remarkably accurate approximations for all three estimators: HOL, LES, and RCS, with the variance being slightly higher for RCS. We found that the estimated distribution of the actual delay is approximately normally distributed in each case, as predicted by the limit in (25).

## 6. Insights from Heavy-Traffic (HT) Limits

We can gain additional insight about the performance of the different estimators by considering HT limits

**Table 5** **Comparing Approximations for $E[W_{\mathrm{HOL}}(w)]$ and $\mathrm{Var}(W_{\mathrm{HOL}}(w))$ for Fixed $w$ Following from (19) and (21) with Simulation Estimates of Mean and Variance of HOL, LES, and RCS Estimators in $GI/M/100$ Model with $\rho = 0.95$ as a Function of Interarrival-Time Distribution**

| Testing the approximations (19) and (21) with observed $w$ in a small interval about $2E[W_\infty \mid W_\infty > 0]$ | | | |
|---|---|---|---|
| Interarrival-time dist. | $M$ | $D$ | $H_2$ |
| $2E[W_\infty \mid W_\infty > 0]$ | 0.40 | 0.20 | 0.96 |
| Selected HOL $w$ interval | [0.39, 0.41] | [0.19, 0.21] | [0.94, 0.98] |
| Sample size | 128,287 | 99,747 | 151,556 |
| Sample mean observed | 0.3998 | 0.2000 | 0.9597 |
| $E[W_{\mathrm{HOL}}](w)$ est. | 0.4003 | 0.1996 | 0.9625 |
| $\mathrm{Var}(W_{\mathrm{HOL}})(w)$ est. | 0.0080 | 0.0020 | 0.0448 |
| $E[W_{\mathrm{LES}}(w)]$ est. | 0.3996 | 0.1995 | 0.9617 |
| $\mathrm{Var}(W_{\mathrm{LES}}(w))$ est. | 0.0081 | 0.0021 | 0.0450 |
| $E[W_{\mathrm{RCS}}(w)]$ est. | 0.3938 | 0.1929 | 0.9586 |
| $\mathrm{Var}(W_{\mathrm{RCS}}(w))$ est. | 0.0103 | 0.0029 | 0.0507 |
| Predicted mean by (19) | 0.400 | 0.205 | 0.947 |
| Pred. variance by (21) | 0.0076 | 0.0021 | 0.0455 |

*Notes.* Data are collected for observed waiting times contained in a small interval about $2E[W_\infty \mid W_\infty > 0]$. The resulting sample sizes are shown.

for the $GI/M/s$ model. To do so, we consider a family of models indexed by the parameter $\rho$, so we introduce a second subscript $\rho$ in addition to $s$. We let the service times remain unchanged. We assume that we start with interarrival times $U_n$ having mean $1/s$. In system $(s, \rho)$, we use interarrival times $U_n/\rho$, so that they have mean $1/s\rho$. That makes the traffic intensity in model $\rho$ be $\rho$.

We consider both the classical HT regime in which $\rho \uparrow 1$ for fixed $s$ and the QED many-server HT regime in which both $\rho \uparrow 1$ and $s \to \infty$ with $(1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$; see Whitt (2002, chaps. 5, 9–10) for background. The queue length tends to be of order $1/(1-\rho)$ in both limiting regimes, but the delays behave differently. The delays are of order $1/(1-\rho)$ in the classical HT regime, but are of order $1-\rho$ or $1/\sqrt{s}$ in the QED HT regime.

### 6.1. HT Snapshot Principle

Just as in the application of HT limits to plan queueing simulations reviewed in Whitt (2002, §5.8), the time scaling in the HT stochastic-process limits provides important insight. In particular, we can apply the celebrated *HT snapshot principle*, see Reiman (1982) and Whitt (2002, p. 187), which, in our context, tells us that the waiting times (of other customers) tend to change negligibly during the time a customer spends waiting when the system is in HT. In other words, the snapshot principle immediately implies that the LES and HOL estimators are asymptotically exact in heavy-traffic limits (specifically, the ratio converges to one). It also shows that, asymptotically in the HT limit, there is no advantage in averaging over delays of past customers.

Since we are primarily concerned with waiting times, it is appropriate to focus on the virtual waiting time stochastic process, which describes the waiting time of a potential arrival who would come at time $t$. We first consider the classical HT regime. Let $W_{s,\rho}(t)$ be the virtual waiting time at time $t$ in model $(s, \rho)$. The waiting time of the $k$th arrival at time $A_{k,s,\rho}$ is just $W_{s,\rho}(A_{k,s,\rho}-)$, where $g(t-)$ is the left limit of the function $g$ at time $t$.

The classical HT stochastic-process limit for the virtual waiting time process states that

$$(1-\rho)W_{s,\rho}((1-\rho)^{-2}t) \;\Rightarrow\; \mathrm{RBM}(t) \quad \text{as } \rho \uparrow 1, \qquad (47)$$

where the limit stochastic process $RBM(t)$ is a reflected Brownian motion, which has continuous sample paths, and the convergence in distribution is for the entire stochastic process with sample paths in the function space $D$; see Whitt (2002). The space scaling in (47) implies that the waiting times will be of order $O(1/(1-\rho))$, while the time scaling in (47) implies that the waiting times will only change significantly over time intervals of length of order $O(1/(1-\rho)^2)$. As a consequence, we conclude that the HOL and LES estimators are relatively consistent in the classical HT regime.

A similar story holds in the QED HT regime. The stochastic-process limit for the virtual waiting time process in the QED regime is obtained by Puhalskii and Reiman (2000). Let $W_{s,\rho}(t)$ be the virtual waiting time at time $t$ in model $(s, \rho)$. Paralleling (47), in the QED regime, we have the stochastic-process limit

$$\sqrt{s}W_{s,\rho}(t) \;\Rightarrow\; Y(t) \quad \text{as} \quad \rho \uparrow 1, \qquad (48)$$

where the limit process $Y(t)$ is no longer RBM, but again is a diffusion process with continuous sample paths and again the convergence in distribution is for the entire stochastic process with sample paths in the function space $D$.

The time and space scaling in (48) is drastically different from (47), but we nevertheless obtain the same conclusions about our estimators. Now, the waiting times are getting small instead of large, being of order $O(1/\sqrt{s})$, but there is no time scaling at all, so that the waiting times will only change significantly over time intervals of length of order $O(1)$. As a consequence, we conclude that the HOL and LES estimators are also relatively consistent in the QED HT regime. Again, we conclude that there will be no advantage to averaging the delays experienced over past customers.

### 6.2. Steady-State HT Limits

In the e-companion, we also establish HT limits in both regimes for steady-state random variables. We focus on the HOL estimator; by Theorem 4.4, the LES estimator behaves the same. We see what happens "on average" to the random variable $W_{\mathrm{HOL},s,\rho}(w)$ (where the observed delay $w$ has the steady-state distribution). From the steady-state HT limits, we deduce that both the direct QL and HOL estimators are (weakly) relatively consistent: the ratio of the estimator to the random quantity being estimated converges to 1. We also develop limits establishing the asymptotic efficiency of the different estimators (comparing MSEs). In these HT limits, the direct and refined estimators have asymptotically the same efficiency, while the QL estimator is asymptotically more efficient than these delay-history estimators by the constant factor $c_a^2+1$, consistent with Theorem 4.2. Because associated heavy-traffic stochastic-process limits have been established for other models, the estimators should have similar nice properties for other models.

## 7. Conclusions

### 7.1. Insights That Can Be Generalized

Even though we are primarily interested in service systems that are more complex than the $GI/M/s$ queueing model, in this paper, we studied the performance of alternative delay estimators in this relatively simple idealized $GI/M/s$ setting. Our goal has been to gain insight into how the estimators will perform in more complex settings. Our results for the $GI/M/s$ model indicate what to expect more generally. Although it remains to be verified in each specific context, we anticipate that many of the performance conclusions for the $GI/M/s$ model (reviewed below) will extend to other settings. At a minimum, the results here serve as a basis for comparison in further examination of delay estimation.

### 7.2. Performance of Estimators

An important reference point for the delay estimators based on delay history is the standard QL estimator based on the observed QL, defined in (1). For QL, the only source of uncertainty is the remaining service times of the customers ahead of the arrival. That uncertainty can be reduced if the remaining service times can be reliably estimated, as emphasized by Whitt (1999).

As can be seen from formulas (9)–(11), to a large extent, the LES and HOL estimators can be regarded as the QL estimator modified by replacing the known QL by an estimate of that QL. Because the QL is equal (or approximately equal) to the number of arrivals during the observed waiting time, the QL is estimated by the expected number of arrivals during the observed waiting time. Thus the increase in MSE in going from QL to the LES, HOL, and RCS estimators is primarily because of variability in the arrival process. The MSE tends to be larger for LES and HOL than QL by the constant factor $(c_a^2+1)$, where $c_a^2$ is the SCV of an interarrival time, a common measure of variability for a renewal arrival process; see Whitt (1982).

As a consequence, the delay estimators based on delay history will perform about the same as the QL estimator when the arrival process has very low variability, but the relative performance will degrade as that arrival-process variability increases. From the perspective of statistical precision, the QL estimator should be preferred to the delay-history estimators if it is available, unless there is negligible arrival-process variability. The delay-history estimators offer the advantage of transparency, but that is obtained at the expense of statistical precision. This insight should apply very broadly.

Overall, we conclude that the greatest source of estimation uncertainty is the remaining service times. After that, it is the arrival-process variability, as partially characterized by the SCV $c_a^2$. We conclude that the estimators $\theta_{\mathrm{QL}}(n)$, $\theta^d_{\mathrm{LES}}(w)$, $\theta^d_{\mathrm{HOL}}(w)$, and $\theta^d_{\mathrm{RCS}}(w)$ can be very useful, but they are not extraordinarily accurate. The refined estimators for HOL, LES, and RCS can remove all or nearly all of the bias, but nonnegligible variance remains. The greatest hope for more reliable estimation seems to lie in being able to better predict the remaining service times, which is certainly possible if the service times are actively *controlled*, and is possible to some extent if either the service-time distribution is nonexponential or if it is possible to classify the customers, as discussed in Whitt (1999). An important direction for further research is to develop more sophisticated estimators that exploit much more of the information. Nevertheless, there may always be a role for the transparent delay estimators based on recent delay history considered here.

We considered several different delay estimators based on recent delay history, notably LES, HOL, and RCS. Through analysis and extensive simulation experiments, we conclude that the LES and HOL delay estimators are very similar, with both being more accurate than the others based on delay history, but less accurate than the full-information QL estimator. For large $s$, RCS is far superior to the delay of the LCS, because customers need not complete service in the same order they arrive. For low traffic intensities with large $s$, LCS was even outperformed by the NI estimator. The reason is that the LCS customer may have arrived too long ago. We conclude that RCS should only be preferred to HOL and LES if delay information is not available until after customers complete service, but the MSE is not much greater for RCS than for LES and HOL.

For the $GI/M/s$ model, the random delay $W_{\mathrm{HOL}}(w)$ given the HOL observation $w$ is remarkably tractable, as can be seen from the representation (11). Theorem 4.2 gives the exact mean and variance of $W_{\mathrm{HOL}}(w)$ for Poisson arrivals. It is significant that the mean $E[W_{\mathrm{HOL}}(w)]$ is not simply $w$, but instead is a linear function of it: $\rho w + (2/s)$ with Poisson arrivals. That mean serves as a refined estimate, which has lower MSE than the direct estimator, but it requires extra information. Bias in the direct estimators can be expected more generally.

For general renewal arrivals, Theorem 4.2 establishes asymptotic results that generate simple approximations, which may well describe the behavior of these estimators in other settings. As $sw$ increases, the random variable $W_{\mathrm{HOL}}(w)$ is asymptotically normally distributed with explicit mean and variance (§4), which has been substantiated by simulation, as discussed in §5. From (25), we see that the squared coefficient of variation $c^2_{W_{\mathrm{HOL},s}(w)}$ is asymptotically proportional to $(c_a^2+1)/\rho s w$ as $sw \to \infty$. That implies very accurate prediction when $sw$ is large. These properties of $W_{\mathrm{HOL}}(w)$ (and $W_{\mathrm{LES}}(w)$) can be expected to hold more generally.

In §6 and the e-companion, we showed that heavy-traffic limits provide important insight. The heavy-traffic snapshot principle provides strong support for all these delay-history estimation procedures, and shows that there should be little benefit from averaging over past customer delays, under heavy loads. The relative errors of the LES and HOL estimators are asymptotically negligible in both the classical and many-server HT regimes. The MSE relative to the mean is asymptotically negligible for all the candidate delay estimators based on delay history. The QL estimator is asymptotically more efficient than HOL and LES by the constant factor $c_a^2+1$ in both HT regimes. Since similar HT limits have already been established for much more general models, these HT properties can be expected to hold more generally.

### 7.3. Possible New Applications

For call centers as well as other service systems (e.g., delays in receiving new products or getting an application processed), there may be new applications of these alternative delay estimators based on recent delay history. They can also be used by customers and third parties who do not have access to all the state information available to the service provider. This might work as follows: Large groups of customers might voluntarily route their delay experience electronically to a centralized consumer-group monitor that makes this information available to its customer base in real time. The customers, in turn, could have their communication equipment set up to simultaneously query the monitor whenever the customer contacts the service provider. In this way, the flow of critical information could take place in milliseconds, which is far shorter than a short telephone call. This is not beyond current technology.

In the same spirit, the LES delay estimator could be used by outside parties to verify that the service provider is providing accurate delay estimates. The service provider could agree to publish its delay estimates, providing extra coded information giving the customer identification for each observed LES delay. Customers or authorized third parties could then verify that the delays, appropriately recorded, coincided with that same delay when it was quoted as an LES delay. The information available to each customer would not go beyond its own delay experience, and yet, collectively, customers could verify the accuracy of the delay predictions. Such verification might well be regarded as a legitimate customer concern. And service providers might want to offer the verification as a way to provide better service.

## Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

## Acknowledgment

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

## References

Abate, J., W. Whitt. 1992. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* **10** 5–88.

Abramowitz, M., I. A. Stegun. 1972. *Handbook of Mathematical Functions*. National Bureau of Standards, U. S. Dept. of Commerce, Washington, D.C.

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

Asmussen, S. 2003. *Applied Probability and Queues*, 2nd ed. Springer, New York.

Avramidis, A. N., A. Deslauriers, P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Sci.* **50** 896–908.

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

Coates, M., A. O. Hero, III, R. Nowak, B. Yu. 2002. Internet tomography. *IEEE Signal Processing Magazine* **19** 47–65.

Doytchinov, B., J. Lehoczky, S. Shreve. 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* **11** 332–378.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed. John Wiley and Sons, New York.

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

Glynn, P. W., W. Whitt. 1989. Indirect estimation via $L = \lambda W$. *Oper. Res.* **37** 82–103.

Ibrahim, R. E., W. Whitt. 2007. Supplement to "Real-time delay estimation based on delay history." IEOR Department, Columbia University, New York. http://columbia.edu/~ww2040.

Ibrahim, R. E., W. Whitt. 2008. Real-time delay announcements in service systems with customer abandonment. Working paper, Columbia University, New York.

Larson, R. C. 1990. The queue inference engine: Deducing queue statistics from transactional data. *Management Sci.* **36**(5) 586–601.

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

Nakibly, E. 2002. Predicting waiting times in telephone service systems. Master's thesis, The Technion, Haifa, Israel.

Puhalskii, A. A., M. I. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595.

Reiman, M. I. 1982. The heavy traffic diffusion approximation for Sojourn times in Jackson networks. R. L. Disney, T. J. Ott, eds. *Applied Probability—Computer Science, The Interface, II*. Birhauser, Boston, 409–422.

Ross, S. M. 1996. *Stochastic Processes*, 2nd ed. John Wiley and Sons, New York.

Ward, A. W., W. Whitt. 2000. Predicting response times in processor-sharing queues. D. R. McDonald, S. R. E. Turner, eds. *Analysis of Communication Networks*: *Call Centres, Traffic and Performance*. Fields Institute Communications 28, American Math. Society, Providence, RI, 1–29.

Whitt, W. 1982. Approximating a point process by a renewal process: Two basic methods. *Oper. Res.* **30** 125–147.

Whitt, W. 1984. On approximations for queues, I: Extremal distributions. *AT&T Bell Lab. Tech. J.* **63**(1) 115–138.

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

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

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

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.

Xu, S. H., L. Gao, J. Ou. 2007. Service performance analysis and improvement for a ticket queue with balking customers. *Management Sci.* **53**(6) 971–990.