REAL-TIME DELAY ESTIMATION BASED ON DELAY HISTORY IN MANY-SERVER SERVICE SYSTEMS WITH TIME-VARYING ARRIVALS

Abstract

Motivated by interest in making delay announcements in service systems, we study real-time delay estimators in many-server service systems, both with and without customer abandonment. Our main contribution here is to consider the realistic feature of time-varying arrival rates. We focus especially on delay estimators exploiting recent customer delay history. We show that time-varying arrival rates can introduce significant estimation bias in delay-history-based delay estimators when the system experiences alternating periods of overload and underload. We then introduce refined delay-history estimators that effectively cope with time-varying arrival rates together with non-exponential service-time and abandonment-time distributions, which are often observed in practice. We use computer simulation to verify that our proposed estimators outperform several natural alternatives.

(*Delay Estimation; Delay Announcements; Time-Varying Arrival Rates; Simulation.*)

April 8, 2010

1. Introduction

We investigate alternative ways to estimate, in real time, the delay (before entering service) of an arriving customer in a service system with time-varying arrival rates. We consider time-varying arrival rates because arrival processes to service systems, in real life, typically vary significantly over time.

Our delay estimators may be used to make delay announcements. Delay announcements may be especially helpful when delays are sometimes long, as in a hospital emergency department (ED). In many cases waiting customers are unable to accurately estimate their own delay, and would therefore gain from delay announcements. That is typically true with invisible queues, as occur in call centers; see Aksin et al. (2007) for background on call centers.

1.1. Delay-History-Based Estimators

In this paper, we examine alternative estimators based on recent customer delay history in the system. As in Armony et al. (2008), a candidate delay estimator based on recent customer delay history is the delay of the last customer to have entered service, prior to our customer's arrival at time t, denoted by LES. That is, letting w be the delay of the last customer to have entered service, the corresponding LES delay estimate is: $\theta_{LES}(t, w) \equiv w$. Armony et al. (2008) 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.

Closely related to LES 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. The HOL delay estimator was mentioned as a candidate delay announcement by Nakibly (2002). For a detailed discussion of the HOL and LES estimators, see Ibrahim and Whitt (2009a, b).

Experience indicates that the LES and HOL estimators have very similar performance. In complex systems, the LES delay is more likely to be observable than the HOL delay, because arrival and service completion times are more likely to be known than the experience of customers who have not yet completed their service; e.g., customers may have abandoned and that might not be known. Nevertheless, here we focus on HOL, because it is easier to analyze. However, we do so with the understanding that similar results will hold for LES.

1.2. Motivation For Delay-History-Based Estimators

We now briefly explain why it is important to study the performance of delay-history-based estimators; for more discussion, see §1 of Ibrahim and Whitt (2009a). First, delay-history-based estimators are currently used in service systems. For one example, the U.S. Citizenship and Immigration Service (USCIS) publishes the arrival time of the most recently completed application to give an idea about upcoming delays. For another example, the HOL estimator was used as an announcement in an Israeli bank studied by Mandelbaum et al. (2000).

Second, delay-history-based estimators are appealing for complicated service systems. For one example, there may be multiple customer classes with multiple service pools. For another example, with Web chat, servers typically 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 yet another example, in ticket queues studied by Xu et al. (2007). Upon arrival at a ticket queue, each customer is issued a numbered ticket. The number currently being served is displayed. The queue length is not known to ticket-holding customers or even to system managers, because they do not observe customer abandonments. Even in systems with no customer abandonment, we may not know the queue length in the system at a new arrival epoch. In a ticket queue (as at a supermarket), a ticketed customer may elect to go and do other shopping and plan to come back later to get in line. (Customers may also abandon, but that does not have to be the case.) Customers with tickets could return to the queue at some point in time and "preempt" customers who are already in line (e.g., if they have a lower numbered ticket). Now, suppose that there is a new arrival at the station.

It is unclear whether ticketed customers (currently doing some other shopping) will return quickly enough to be inserted before that new arrival. Therefore, the queue length cannot be determined at the new arrival epoch. Nevertheless, it is possible to determine who the LES (or HOL) customer is, and to know his/her delay.

Delay-history-based estimators are appealing, from a practical perspective, whenever the queue length is not known, but also because they do not depend on the model and use very little information about the system. They are robust because they respond automatically to changes in system parameters (e.g., number of servers, mean service time, and arrival rate).

To fully understand a complex service system, we need to study it in detail. However, to help develop a service science, we are systematically studying various delay estimators in controlled environments, i.e., in structured models, starting with $GI/M/s$ and extending to $GI/GI/s$ (non-exponential service times), $GI/GI/s + GI$ (abandonment with non-exponential patience distributions) in Ibrahim and Whitt (2009a, b) and now $M_t/GI/s$ and $M_t/GI/s + GI$ (time-varying arrival rates).

1.3. The Case of a Stationary Arrival Process

In Ibrahim and Whitt (2009a, b), we studied the performance of the LES and HOL delay estimators in many-server systems, both with and without customer abandonment, by studying conventional stationary queueing models. In Ibrahim and Whitt (2009a), we studied the performance of HOL in the $GI/M/s$ queueing model, which has a renewal arrival process, s homogeneous servers, an unlimited waiting room and the first-come-first-served (FCFS) service discipline. The service times are independent of the arrival process, and independent and identically distributed (i.i.d.) exponential random variables.

We showed that HOL is an effective estimator in the $GI/M/s$ model. As a frame of reference, we considered the classical delay estimator based on the queue length, denoted by QL, which multiplies the queue length plus one times the mean interval between successive service completions, ignoring customer abandonment. For this special idealized model with i.i.d. exponential service times and no customer abandonment, the QL estimator is provably the most effective estimator, under the mean squared error (MSE) criterion; see §4 below.

The HOL estimator performs worse than QL, because it does not exploit queue-length information. Nevertheless, we showed that the difference in performance need not be too great, particularly when the arrival process has low variability. Because the model is highly structured, we were able to obtain analytical results.

In Ibrahim and Whitt (2009b), we considered the $GI/GI/s + GI$ model, which includes independent sequences of i.i.d. service times and abandonment times with general distributions. As one would expect, QL can overestimate customer delay when there is significant customer abandonment in the system. We showed that QL performs poorly in a heavily loaded $GI/GI/s + GI$ model, while HOL remains an effective estimator.

When customer abandonment is a serious issue, it is possible to refine the queue-length-based delay estimator by using the exact expected conditional delay, given the queue length, in the $G/M/s+M$ model; we denote this by QL_m. However, for non-exponential service-time and abandonment distributions, the delay-history-based estimators can also outperform this refined queue-length-based estimator QL_m, even when the queue length and the model are known; e.g., see Figures 1-4 of Ibrahim and Whitt (2009b).

However, we do not mean to suggest that the queue length does not provide useful information when it is known. Indeed, our best estimator for the $GI/GI/s + GI$ model is an approximation-based estimator, referred to as QL_{ap}, which exploits the queue length as well as model parameters; we also will make use of QL_{ap} here for the $M_t/GI/s + GI$ model in §8.

1.4. Time-Varying Arrival Rates

In this paper, we study the performance of the HOL estimator with time-varying arrival rates. We do so primarily because arrival rates typically vary significantly over time in real-life service systems.

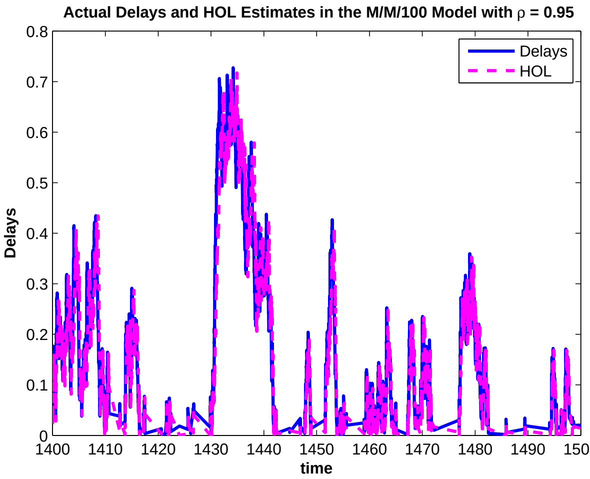
The HOL estimator can perform poorly when the delays vary systematically over time, as can occur when there are alternating periods of significant overload and underload. Then the delay of a new arrival may not be like the HOL delay. To demonstrate potential problems with the HOL estimator, we plot simulation sample paths of HOL delay estimates given, and actual delays observed, as a function of time, in simulation runs from two different heavily-loaded many-server systems. In Figure 1, we consider the stationary $M/M/100$ model with traffic intensity $\rho = 0.95$ and mean service time 5 minutes; in Figure 2, we consider the $M_t/M/100$ model with sinusoidal arrival rates, again with traffic intensity $\rho = 0.95$, but now defined as the long-run average, and mean service time 5 minutes. We consider a daily cycle, so that there is one peak during the day. We let the relative amplitude be $\alpha = 0.5$. (The ratio of the peak arrival rate to the average arrival rate is $1 + \alpha$.) We measure time and, thus, the delays in units of mean service times. The overall plotted time interval of length 500 mean service times is slightly less than two days, so we see two peaks.

For Figure 2, we deliberately chose an extreme case in which the system alternates between extreme overload and underload, while the number of servers remains fixed. In that setting, the maximum delays themselves are about 40 mean service times or 200 minutes, about 60 times greater than in the stationary environment. Delay estimation tends to be especially important with such large delays. Figure 2 shows that, with time-varying arrival rates, the HOL curve is clearly shifted to the right of the actual-delay curve; i.e., there is a time lag between the HOL estimates and the actual delays observed, leading to big errors.

Figure 2 also shows a third plot, the plot of a refined HOL estimator, denoted by HOL_r, which we develop in §4. Clearly, it eliminates the time lag; visually the HOL_r plot falls on top of the actual delays. The ratio of the average squared errors $\mathrm{ASE(HOL)}/\mathrm{ASE(HOL}_r)$, defined in §3 below, is about 95 in Figure 2. (If we would reduce the relative amplitude from 0.5 to 0.1, then the ratio would be only 1.3; it then requires careful analysis to see the improvement provided by HOL_r over HOL; see Ibrahim and Whitt (2009c) for the plot.)

In this paper, we not only show that HOL may not be an effective estimator with time-varying arrivals, particularly when the system alternates between phases of underload and overload, but we also develop refinements of the HOL estimator that remain effective for time-varying arrival rates. Through analysis and simulation, we show that these new estimators perform remarkably well with time-varying arrival rates, far better than HOL.

However, the improved performance of the refined HOL estimators comes at the expense



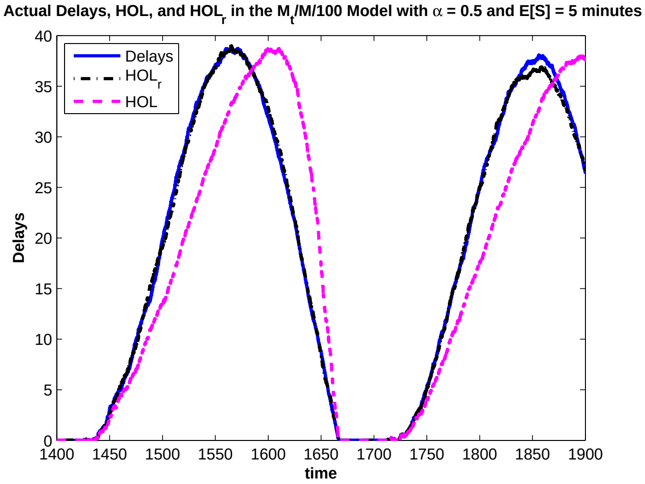


Figure 1: Sample paths of actual delays and HOL delay estimates with constant arrival rate

Figure 2: Sample paths of actual delays and delay estimates using HOL and HOL$_r$ with sinusoidal arrival rate

of exploiting more information about the system, such as the arrival rate, the number of servers and the mean service time. That requirement greatly reduces the advantage over queue-length-based delay estimators. Indeed, our strategy for obtaining the refined HOL estimators involves two steps: (i) representing or approximating the expected conditional delay given the queue length, and (ii) estimating the queue length, given the observed HOL delay and the model parameters. Hence, the refined HOL estimators are valuable only when the queue length is not known. However, such cases are not uncommon, as in Web chat and ticket queues, when we directly observe arrivals and service completions, but not the queue, because we do not observe customer abandonments.

Because our refined estimators exploit more information about the system, we also investigate (i) how our refined estimators perform if the extra information is known imperfectly, because it too must be estimated, and (ii) how this additional information can be estimated in real time. We propose estimation procedures for alternative system parameters, and quantify the estimation error resulting from those procedures. These additional experiments show that the refined estimators can be useful in practice.

1.5. Literature Review, Contributions, and Organization

The literature on delay announcements is large and growing. In broad terms, there are two main areas of research. The first area studies the effect of delay announcements on system dynamics; e.g., see Whitt (1999b), Armony and Maglaras (2004), Guo and Zipkin (2007), Armony et al. (2008), Allon et al. (2009), and references therein. The second area studies alternative ways of estimating customer delay in service systems; e.g., see Nakibly (2002), Whitt (1999a), Jouini et al. (2007), and Ibrahim and Whitt (2009a, b). For a more detailed review, see Section 2 of Jouini et al. (2007).

This paper falls in the second main area of research. Our main contributions are: (i) to show that time-varying arrival rates can cause estimation bias for delay-history-based delay estimators, (ii) to propose new and easily implementable delay estimators, based on the history of delays in the system, that effectively cope with time-varying arrivals and general service-time and abandon-time distributions, (iii) to provide analytical results quantifying the performance of some delay estimators, and (iv) to describe results of a wide range of simulation experiments evaluating alternative delay estimators, with time-varying arrivals.

The rest of this paper is organized as follows: In §2, we describe the modeling framework. In §3, we describe measures quantifying the performance of our candidate delay estimators. In §4, we introduce a new delay estimator for the $M_t/GI/s$ model. In §5, we provide analytical results for the performance of this estimator in the $M_t/M/s$ model. In §6, we present simulation results showing that it is effective in the $M_t/GI/s$ model. In §7, we propose ways of obtaining the additional system information required for implementing the new delay estimator of §4. In §8, we develop a new delay estimator for the $M_t/GI/s + GI$ model. In §9, we present simulation results showing that it is effective. We make concluding remarks in §10. Additional material appears in an online supplement, Ibrahim and Whitt (2009c).

2. The Framework

We consider many-server queueing models with time-varying arrival rates, both with and without customer abandonment. We model the arrival process as a nonhomogeneous Poisson process, which is the accepted model for capturing time-varying arrivals. It is completely characterized by its deterministic arrival-rate function $\lambda \equiv \{\lambda(u) : -\infty < u < \infty\}$. There is statistical evidence suggesting that a nonhomogeneous Poisson process is a good fit for the arrival process to a call center; see Brown et al. (2005). We adopt this model for arrivals, although we recognize its shortcomings. For example, this model does not reproduce an essential feature of call center arrivals, which is the over-dispersion of the number of arrivals relative to the Poisson distribution (i.e., the variance is larger than the mean); see Avramidis et al. (2004). Moreover, the arrival rate in a real-life system is often not known with certainty. Therefore, it could be assumed to be a random variable; see Jongbloed and Koole (2001). It is natural, however, to begin an investigation in a relatively tractable setting, for which we are able to obtain analytical results. Our results provide useful background for similar studies in even more complicated settings.

In §4-6, we consider the $M_t/GI/s$ model, which has a nonhomogeneous Poisson arrival process, i.i.d. service times distributed as a random variable S with a general distribution, having mean $E[S] = \mu^{-1}$ and no customer abandonment. Motivated by large service systems, we are primarily interested in the case of large s, which we take to be fixed. It is possible to choose appropriate time-varying staffing (making s a function of time) so that delays are stabilized at low levels; e.g., see Green et al. (2007). However, in practice there often is not adequate flexibility in setting staffing levels. Our fixed staffing assumption captures the spirit of such situations. We leave to future research the important extension to time-varying staffing levels.

Our delay estimators apply to arbitrary arrival rate functions, but to analyze the performance of these estimators we restrict attention to periodic arrival rate functions, under which the queueing system has a dynamic steady state, provided that the average arrival rate, denoted by $\bar{\lambda}$, is strictly less than the maximum possible service rate, $s\mu$; e.g., see Heyman and Whitt (1984). For our analysis, both analytically and by simulation, we further restrict attention to the special case of sinusoidal arrival rates. That is commonly done in studies of queues with time-varying arrivals; e.g., see Green et al. (2007) and references therein. Sinusoidal arrival rates capture the spirit of daily cycles.

In §8 and §9 we consider the $M_t/GI/s + GI$ model, which adds customer abandonment. The abandonment times are i.i.d. with mean ν^{-1} and a general cumulative distribution function (cdf) F. As in Ibrahim and Whitt (2009b), we see that the abandonment distribution has a significant impact.

3. Performance Measures for the Delay Estimators

In this section, we indicate how we evaluate the performance of our candidate delay estimators. We use computer simulation to do the actual estimation. In our simulation experiments, we quantify the performance of a delay estimator by computing the *average squared error* (ASE), defined by:

$$ASE \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - e_i)^2 \ , \qquad (3.1)$$

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

As discussed in Ibrahim and Whitt (2009a, b), the ASE approximates the expected *mean squared error* (MSE) for a system in steady state with a constant arrival rate, but the situation is more complicated with time-varying arrivals. We regard ASE as directly meaningful, but now we indicate how it relates to the MSE. Let $W_{HOL}(t, w)$ 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 elapsed delay of the customer at the head of the line at the time of his arrival, t, is equal to w. Let $\theta_{HOL}(t, w)$ be some given single-number delay estimate which is based on the HOL delay, w, and the time of arrival, t. Then, the MSE of the corresponding delay estimator is given by:

$$MSE(\theta_{HOL}(t, w)) \equiv E[(W_{HOL}(t, w) - \theta_{HOL}(t, w))^2] \ , \tag{3.2}$$

which is a function of w and t. In order to get the overall MSE of HOL at time t, we average with respect to the *unconditional* distribution of the HOL waiting time at time t, $W_{HOL}(t)$, i.e.,

$$MSE(t) \equiv E[MSE(\theta_{HOL}(t, W_{HOL}(t)))]. \tag{3.3}$$

Finally, in order to relate the ASE in (3.1) to the MSE, we need to average $MSE(t)$ defined in (3.3) appropriately over time, but since the ASE represents a customer average instead of a time average, we need to use a weighted time average of the time-dependent MSE in (3.2) in order to relate it to the ASE. In particular, if T is the cycle length, then

$$ASE \approx \frac{\int_0^T \lambda(u) MSE(u) du}{\int_0^T \lambda(u) du} \ , \tag{3.4}$$

where $MSE(t)$ is defined in (3.3); for supporting theory see the appendix of Massey and Whitt (1994).

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

$$RRASE \equiv \frac{\sqrt{ASE}}{(1/k)\sum_{i=1}^{k} p_i} \ , \tag{3.5}$$

using the same notation as in (3.1). The denominator in (3.5) is the average potential waiting time of customers who must wait. The RRASE is useful because it measures the effectiveness of an estimator relative to the average potential waiting time, given that the customer must wait.

4. Delay Estimators for the $M_t/GI/s$ Model

In this section, we propose a new refined HOL-based delay estimator, HOL$_r$, for the $M_t/GI/s$ model. Our idea is to use the refined estimator $\theta^r_{HOL}(t, w) \equiv E[W_{HOL}(t, w)]$ instead of the

HOL estimator $\theta_{HOL}(t,w) \equiv w$, because the mean necessarily minimizes the MSE based on this information. However, this mean is difficult to compute, so we propose an approximation. We approximate the mean in the given $M_t/GI/s$ model by its exact value in the corresponding $M_t/M/s$ model, with exponential service time having the given mean $E[S]$.

For the $M_t/M/s$ model, we have the representation:

$$W_{HOL}(t,w) \equiv \sum_{i=1}^{A(t)-A(t-w)+2} S_i/s \ , \tag{4.1}$$

where $\{A(t) : t \geq 0\}$ denotes the arrival (counting) process. We have division by s in (4.1) 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 rate μ, which makes the minimum exponential with rate $s\mu$. The random variable $A(t)-A(t-w)$ has a Poisson distribution with mean $\int_{t-w}^{t} \lambda(u)du$. Since $W_{HOL}(t,w)$ in (4.1) is a random sum of i.i.d. random variables, where $A(t) - A(t-w)$ is independent of the summands S_i/s, we can easily compute this mean. Hence our refined HOL estimator for the $M_t/GI/s$ model is this mean

$$\theta_{HOL_r}(t,w) \equiv E[W_{HOL,M_t/M/s}(t,w)] = \frac{1}{s\mu}\left(2 + \int_{t-w}^{t} \lambda(u)du\right) \ . \tag{4.2}$$

In general, with a non-exponential service-time distribution, $\theta_{HOL_r}(t,w)$ in (4.2) need not equal $E[W_{HOL}(t,w)]$, because many remaining service times at time t are residual service times for service times begun prior to time t. Consequently, these service times have a different distribution than the original service time. However, we can make stochastic comparisons. A cumulative distribution function (cdf) G of a nonnegative random variable is said to be new better (worse) than used - NBU (NWU) - if $G_t^c(x) \equiv G^c(t+x)/G^c(t) \leq (\geq) G^c(x)$ for all $t \geq 0$ and $x \geq 0$, where $G^c(x) \equiv 1 - G(x)$; see p. 159 of Barlow and Proschan (1975). In the parlance of survival analysis, a cdf is NBU (NWU) if the probability of surviving for an additional x time units, given survival up to time t, decreases (increases) with t.

Proposition 1. *If the service-time cdf is NBU (NWU), then* $\theta_{HOL_r}(t,w) \geq (\leq) E[W_{HOL}(t,w)]$.

Proof. The NBU and NWU condition means that the residual service times are stochastically ordered compared to the original service times. Intuitively, approximating an NBU (NWU) distribution by an exponential leads to overestimating (underestimating) the residual service times, and thus the overall delay. Given the elapsed times, the remaining service times are mutually independent. The minimum (the time until the next departure) is thus stochastically ordered compared to the minimum of mutually independent original service-time distributions. The random variable $W_{HOL}(t, w)$ is the sum of several of those intervals between successive departures. Even though those intervals may be dependent, the mean of the sum is the sum of the means. Hence the means are ordered, as claimed. ∎

More importantly, simulation shows that HOL_r provides a good approximation even when the service-time distribution is not nearly exponential; see §6.

We conclude this section by reviewing the QL estimator, previously considered in Ibrahim and Whitt (2009a, b). Let $W_Q(t, n)$ represent a random variable with the conditional distribution of the delay of an arriving customer, given that this customer must wait before starting service, and given that the queue-length seen upon arrival, at time t, is equal to n. Again, the QL estimator is obtained by using the exact expected value $E[W_Q(t, n)]$ for the corresponding $M_t/M/s$ model with the same mean service time.

In the $M_t/M/s$ model, $W_Q(t, n)$ is the sum of $n + 1$ i.i.d. exponential random variables, each with rate $s\mu$. The QL estimate given to a customer who finds n other customers in queue upon arrival is: $\theta_{QL}(t, n) \equiv E[W_Q(t, n)] = (n+1)/s\mu$, which depends on t only through n, which is directly observable. The optimal delay estimator, conditional on the number of customers, n, seen in line at time t, using the MSE criterion, is the one announcing the mean, $E[W_Q(t, n)]$. That is why the QL estimator is the optimal delay estimator, under the MSE criterion, in the $M_t/M/s$ model.

By essentially the same reasoning as for Proposition 1, we can obtain bounds for the mean delay compared to $\theta_{QL}(t, n)$ when the service-time cdf is NBU or NWU.

Proposition 2. *If the service-time cdf is NBU (NWU), then* $\theta_{QL}(t, n) \geq (\leq) E[W_Q(t, n)]$.

Fortunately, again simulation shows that QL remains effective in the $M_t/GI/s$ model, even when the service-time distribution is not nearly exponential; see §6. For the $M_t/M/s$ model, we obtain analytical results quantifying the difference in performance between QL and HOL$_r$ in the next section.

5. Analytical Expressions for the $M_t/M/s$ Model

The QL estimator has the desirable property that the estimation gets relatively more accurate as the observed queue length n increases. For the conditional waiting time at time t based on an observed queue length of n, we have the representation

$$W_Q(t,n) \equiv \sum_{i=1}^{n+1} S_i/s \ . \tag{5.1}$$

The expectation, variance, and squared coefficient of variation (SCV, equal to the variance divided by the square of the mean) of $W_Q(t,n)$ are given by:

$$E[W_Q(t,n)] = \frac{n+1}{s\mu}, \quad Var[W_Q(t,n)] = \frac{n+1}{s^2\mu^2}, \quad c^2_{W_Q(t,n)} \equiv \frac{Var[W_Q(t,n)]}{(E[W_Q(t,n)])^2} = \frac{1}{n+1} \ , \tag{5.2}$$

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

To treat HOL$_r$, we use the representation in (4.1), which allows us to characterize the probability distribution of the random variable $W_{HOL}(t,w)$, in the $M_t/M/s$ model.

Proposition 3. *For the $M_t/M/s$ model,*

$$Var[W_{HOL}(t,w)] = \frac{2}{s^2\mu^2}(1 + \int_{t-w}^{t} \lambda(u)du) \ , \tag{5.3}$$

which, combined with (4.2), yields

$$c^2_{W_{HOL}(t,w)} = \frac{Var[W_{HOL}(t,w)]}{(E[W_{HOL}(t,w)])^2} = 2 \times \frac{1 + \int_{t-w}^{t} \lambda(u)du}{(2 + \int_{t-w}^{t} \lambda(u)du)^2} \ . \tag{5.4}$$

Proof. Formula (5.3) follows from the conditional variance formula, e.g., p.51 of Ross (1996). Formula (5.4) immediately follows from (4.2) and (5.3). ∎.

Since $\theta_{HOL_r}(t,w) \equiv E[W_{HOL}(t,w)]$ and $\theta_{QL}(t,n) \equiv E[W_Q(t,n)]$, we can compare the performance of HOL$_r$ and QL by comparing the respective SCV's in (5.2) and (5.4). (When the delay estimate equals the conditional mean, the MSE coincides with the variance.)

To obtain further results, we consider a sinusoidal arrival-rate function

$$\lambda(u) = \bar{\lambda} + \beta \sin(\gamma u) \equiv \bar{\lambda} + \bar{\lambda}\alpha \sin(2\pi u/\Gamma), \quad \text{for } -\infty < u < \infty \ , \tag{5.5}$$

where $\bar{\lambda}$ is the average arrival rate, α is the relative amplitude and Γ is the cycle length. (We define $\beta \equiv \bar{\lambda}\alpha$ and $\gamma \equiv 2\pi/\Gamma$.) Given the cycle length, Γ, we can deduce the place where any time u falls within the cycle, in dynamic steady state. Henceforth, we focus solely on the interval $0 \leq u \leq \Gamma$, which describes a full cycle.

With sinusoidal arrival rates, we obtain analytical results comparing the performance of the QL and HOL$_r$ estimators. We determine the limit of the ratio of the SCV's as $n \to \infty$. Formula (5.6) below coincides with formula (4.25) of Ibrahim and Whitt (2009a) for the stationary $GI/M/s$ model. As before, the condition $n \to \infty$ arises naturally in heavy traffic, either with fixed s or as $s \to \infty$; e.g., see Garnett et al. (2002). (When $s \to \infty$ along with the arrival rate, the queue length is of order s and $\sqrt{s}$ in the ED and QED regimes.) Recall that $\rho \equiv \bar{\lambda}/s\mu$.

Proposition 4. *For the $M_t/M/s$ model with sinusoidal arrival rates,*

$$\frac{c^2_{W_{HOL(t,w)}}}{c^2_{W_{Q(n)}}} \to \frac{2}{\rho} \ \text{as } n \to \infty \ , \tag{5.6}$$

for all t, provided that $w/n \to 1/s\mu$.

Proof. Using Equations (4.2), (5.3), (5.4) and (5.5), we get the following expressions for the mean, variance, and SCV of $W_{HOL}(t,w)$, in the $M_t/M/s$ model with sinusoidal arrivals:

$$E[W_{HOL}(t,w)] = \frac{2 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))}{s\mu} \ , \tag{5.7}$$

and,

$$Var[W_{HOL}(t,w)] = 2 \times \frac{1 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t)))}{s^2\mu^2} \ , \tag{5.8}$$

which yields

$$c^2_{W_{HOL(t,w)}} = \frac{Var[W_{HOL}(t,w)]}{(E[W_{HOL}(t,w)])^2} = 2 \times \frac{1 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))}{[2 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))]^2} \ , \tag{5.9}$$

for $0 \leq t \leq \Gamma$. Using (5.9), and recalling that $-1 \leq \cos(u) \leq 1$ for all u, we obtain the following bounds for the SCV of $W_{HOL}(t, w)$:

$$\frac{2 + 2\bar{\lambda}w - 4\beta/\gamma}{(2 + \bar{\lambda}w + 2\beta/\gamma)^2} \leq c^2_{W_{HOL}(t,w)} \leq \frac{2 + 2\bar{\lambda}w + 4\beta/\gamma}{(2 + \bar{\lambda}w - 2\beta/\gamma)^2} \ . \tag{5.10}$$

Let $W(t)$ be the potential waiting time at time t, the time that an arrival at t would have to wait before beginning service. Since

$$W(t) = \sum_{i=1}^{Q(t)+1} S_i/s \ , \tag{5.11}$$

where $Q(t)$ is the number of customers waiting in queue upon arrival at t, the law of large numbers implies that $W(t)/Q(t) \to 1/s\mu$ as $Q(t) \to \infty$. Thus, when $Q(t)$ is large, we have $W(t) \approx Q(t)/s\mu$. Assuming that n in (5.2) is large with $w = n/s\mu + o(n)$ as $n \to \infty$, where $o(n)$ denotes a quantity that is asymptotically negligible when divided by n, and combining that with (5.10), for large n we get

$$\frac{(2 + 2\rho(n + o(n)) - 4\beta/\gamma)(n + 1)}{(2 + \rho(n + o(n)) + 2\beta/\gamma)^2} \leq \frac{c^2_{W_{HOL(t,w)}}}{c^2_{W_{Q(n)}}} \leq \frac{(2 + 2\rho(n + o(n)) + 4\beta\gamma)(n + 1)}{(2 + \rho(n + o(n)) - 2\beta/\gamma)^2} \ , \tag{5.12}$$

for all t. By a sandwiching argument, (5.12) yields (5.6) as $n \to \infty$. ■

6. Simulations Experiments for the $M_t/GI/s$ Model

In this section, we present simulation results for the $M_t/GI/s$ model, quantifying the performance of QL, HOL, and HOL$_r$ with sinusoidal arrival rates. For the service-time distribution, we consider M (exponential), D (deterministic), and $LN(1, 4)$ (lognormal with mean equal to 1 and variance equal to 4). The $LN(1, 4)$ (D) distribution exhibits high (low) variability, relative to M. We consider a lognormal distribution because there is statistical evidence suggesting a good fit of the service-time distribution to the lognormal distribution in call centers; see Brown et. al (2005).

Description of the Experiments. We fix the number of servers, $s = 100$, because we are interested in large service systems. We consider nonhomogeneous Poisson arrival processes with the sinusoidal arrival rate functions in (5.5). We vary $\bar{\lambda}$ to get alternative values of ρ, for fixed s. We consider values of ρ ranging from 0.90 to 0.98. These values of ρ are chosen to let our systems alternate between periods of overload and underload. We consider two values of the relative amplitude: $\alpha = 0.1$, and $\alpha = 0.5$. Simulation point and 95% confidence interval estimates are based on 10 independent replications of 5 million events each, where an event is either an arrival or a service completion. That is, each simulation run terminates when the sum of the number of arrivals and the number of service completions is equal to 5 million. Here, we show a sample of our simulation results; see Ibrahim and Whitt (2009c) for more.

The parameters of the arrival-rate intensity function, $\lambda(u)$ in (5.5), should be interpreted relative to the mean service time, $E[S]$. As in §1.4, we measure time in units of mean service times; hence $\mu = 1$. Then, we refer to γ in (5.5) as the relative frequency. Table 1 displays values of the relative frequency as a function of $E[S]$, assuming a daily cycle. For interpretation, we also will specify the associated mean service time in minutes, given a daily cycle.

Here, we consider two different values of γ. First, we consider $\gamma = 0.131$, which corresponds to $E[S] = 30$ minutes, assuming a daily cycle. This choice of $E[S]$ could be used to describe the experience of waiting customers in a call center, for example. Second, we consider $\gamma = 1.57$, which corresponds to $E[S] = 6$ hours. This choice of $E[S]$ could be used to describe the experience of waiting patients in a crowded hospital emergency department (ED). With $E[S] = 30$ minutes and $\alpha = 0.1$ ($E[S] = 6$ hours and $\alpha = 0.5$), and daily cycles, the arrival rate varies relatively slowly (rapidly) with respect to the service times.

In Table 2, we present simulation (point and 95% confidence interval estimates) quantifying the performance of QL, HOL$_r$, and HOL in the $M_t/GI/s$ model with M, $LN(1,4)$, and D service-time distributions. We discuss these results next.

Relative Frequency γ	Mean Service Time $E[S]$
0.0220	5 minutes
0.0436	10 minutes
0.131	30 minutes
0.262	1 hour
1.571	6 hours
3.14	12 hours
6.28	24 hours
12.6	48 hours

Table 1: The relative frequency, γ, as a function of the mean service time $E[S]$ for a daily cycle. The relative frequency is the frequency computed with measuring units so that $E[S] = 1$.

Comparing HOL$_r$ and HOL. Table 2 shows that, for $\alpha = 0.1$ and $E[S] = 30$ minutes, HOL$_r$ performs better than HOL, particularly for high values of ρ. We get consistent results with M, $LN(1,4)$, and D service times: ASE(HOL)/ASE(HOL$_r$) is roughly equal to 1 for $\rho = 0.9$, and roughly equal to 1.4 for $\rho = 0.98$. The case with high ρ corresponds to extreme fluctuations between phases of underload and overload, in which case HOL performs relatively poorly.

With $\alpha = 0.5$, and $E[S] = 6$ hours, the difference in performance between HOL and HOL$_r$ is significant, for all ρ considered. For example, with D service times, ASE(HOL)/ASE(HOL$_r$) ranges from about 1.8 for $\rho = 0.9$ to about 2.4 for $\rho = 0.98$. With M service times, ASE(HOL)/ASE(HOL$_r$) ranges from about 2.1 for $\rho = 0.9$ to about 4.8 for $\rho = 0.98$. The HOL$_r$ estimator is also relatively more accurate than HOL. For example, with $LN(1,4)$ service times, RRASE(HOL$_r$) ranges from about 27% for $\rho = 0.9$ to about 15% for $\rho = 0.98$. In this case, RRASE(HOL) ranges from about 38% for $\rho = 0.9$ to about 20% for $\rho = 0.98$.

Comparing HOL$_r$ and QL. In the $M_t/M/s$ model, QL is provably the optimal estimator given the observed queue length upon arrival, under the MSE criterion; see §4. With $\alpha = 0.1$, $E[S] = 30$ minutes, and M service times, Table 2 shows that RRASE(QL) ranges from about 21% for $\rho = 0.9$ to about 10% for $\rho = 0.98$. With non-exponential service times, QL remains the most effective estimator, under the MSE criterion. It is relatively accurate, in all models considered. For example, with $\alpha = 0.5$, $E[S] = 6$ hours, and $LN(1,4)$ service times, RRASE(QL) ranges from about 20% for $\rho = 0.9$ to about 12% for $\rho = 0.98$.

Consistent with §5, the approximation for the ratio of the SCV's in (5.6) provides a remarkably accurate approximation for the ratio of the ASE's with M service times, particularly for high values of ρ, as we would expect. (The distortion caused by the customer average in (3.4) is evidently minor,) For example, with $E[S] = 30$ minutes and $\alpha = 0.1$. Table 2 shows that the relative error between simulation point estimates for ASE(HOL$_r$)/ASE(QL) and numerical values given by (5.6), is less than 3% for $\rho = 0.98$.

With $LN(1,4)$ service times, $E[S] = 30$ minutes, and $\alpha = 0.1$, Table 2 shows that ASE(HOL$_r$)/ASE(QL) ranges from about 1.7 for $\rho = 0.9$ to about 1.5 for $\rho = 0.98$, which is less than predicted by (5.6). Similarly, with D service times, $E[S] = 6$ hours, and $\alpha = 0.5$, Table 2 shows that ASE(HOL$_r$)/ASE(QL) is approximately equal to 1.5 for all ρ.

7. Estimating the Required Additional Information for HOL$_r$

We have shown, both analytically and using simulation, that the HOL estimator can perform poorly when the arrival rate varies considerably over time while the staffing is fixed. We showed that the new refined HOL estimator, HOL$_r$, performs remarkably better than HOL in the $M_t/GI/s$ queueing model, with time-varying arrival rates; see §6.

However, the statistical accuracy of HOL$_r$ is obtained at the expense of ease of implementation. In addition to the HOL delay, w, HOL$_r$ depends on the arrival-rate function, $\lambda(t)$, and the mean time between successive service completions (which equals $1/s\mu$ with s simultaneously busy servers and i.i.d. exponential service times with rate μ); see (4.2). In practice, the implementation of HOL$_r$ requires knowledge of those system parameters, which may require estimation from data. Any estimation procedure inevitably produces some estimation error, which would affect the performance of HOL$_r$.

In this section, we propose estimation procedures for the arrival rate and the mean time

	$M_t/M/100$, $\alpha = 0.1$, $E[S]$ = *30 min*			$M_t/M/100$, $\alpha = 0.5$, $E[S]$ = *6 hrs*		
ρ	QL	HOL$_r$	HOL	QL	HOL$_r$	HOL
0.9	2.26	4.29	4.61	2.24	4.27	9.01
	±0.051	±0.088	±0.098	±0.023	±0.033	±0.15
0.93	3.77	7.29	8.04	2.83	5.45	14.1
	±0.10	±0.21	±0.26	±0.029	±0.063	±0.25
0.95	5.08	10.1	11.7	3.49	6.82	21.4
	±0.072	±0.15	±0.20	±0.033	±0.073	±0.28
0.97	7.16	14.1	17.5	4.82	9.46	39.0
	±0.098	±0.20	±0.24	±0.12	±0.22	±1.5
0.98	9.14	18.0	23.9	6.77	13.3	63.3
	±0.30	±0.59	±1.0	±0.32	±0.62	±3.9

	$M_t/LN(1,4)/100$, $\alpha = 0.1$, $E[S]$ = *30 min*			$M_t/LN(1,4)/100$, $\alpha = 0.5$, $E[S]$ = *6 hrs*		
ρ	QL	HOL$_r$	HOL	QL	HOL$_r$	HOL
0.9	4.36	7.30	7.78	2.08	3.60	7.79
	±0.25	±0.34	±0.36	±0.13	±0.19	±0.33
0.93	6.89	11.3	12.8	3.48	5.90	14.0
	±0.15	±0.34	±0.34	±0.18	±0.27	±0.49
0.95	9.82	15.9	19.0	5.70	9.52	22.5
	±0.28	±0.42	±0.56	±0.14	±0.22	±0.38
0.97	17.2	27.0	35.1	9.92	15.9	34.2
	±0.81	±1.3	±2.1	±0.60	±0.89	±1.1
0.98	23.2	35.8	48.9	20.1	31.0	52.1
	±0.94	±1.4	±2.4	±2.2	±3.3	±3.2

	$M_t/D/100$, $\alpha = 0.1$, $E[S]$ = *30 min*			$M_t/D/100$, $\alpha = 0.5$, $E[S]$ = *6 hrs*		
ρ	QL	HOL$_r$	HOL	QL	HOL$_r$	HOL
0.9	0.972	2.31	2.47	3.02	4.14	7.35
	±0.025	±0.034	±0.036	±0.023	±0.039	±0.054
0.93	1.23	3.84	4.18	3.71	5.01	8.91
	±0.024	±0.063	±0.078	±0.027	±0.026	±0.045
0.95	1.31	5.19	6.01	4.33	5.84	10.5
	±0.027	±0.041	±0.041	±0.038	±0.051	±0.068
0.97	1.35	7.26	9.29	5.41	7.54	15.5
	±0.026	±0.065	±0.038	±0.086	±0.075	±0.14
0.98	1.34	8.29	11.3	6.01	8.84	21.1
	±0.042	±0.057	±0.069	±0.075	±0.076	±0.49

Table 2: A comparison of the efficiency of QL, HOL$_r$, and HOL in the $M_t/GI/100$ model, as a function of the traffic intensity, ρ. Point and 95% confidence interval estimates of the average squared error (ASE) are shown (in units of mean service time squared per customer). Estimated ASE's are in units of 10^{-3}.

between successive service completions in real-life service systems. Further, we quantify the estimation error resulting from those procedures, and its impact on the performance of HOL_r; see Table 3. We show that the HOL_r estimator remains effective even with imperfect information about system parameters.

To estimate the arrival-rate function, $\lambda(t)$, we propose relying on forecasts relying on data from previous days, and observations over the current day, up to date. For $\theta_{HOL_r}(t, w)$ in (4.2), we need estimates of the arrival-rate function over the interval $[t - w, t]$. Here, we assume that the arrival process is a nonhomogeneous Poisson process with an integrable arrival-rate function. Since we observe customer arrival times, but not the arrival rates, we need to forecast future rates based on historical call volumes. For ways of forecasting future arrival rates, we refer the reader to recent work on forecasting arrival rates to service systems such as call centers. For one example, Shen and Huang (2008b) propose an approach to forecast the time series of an inhomogeneous Poisson process by first building a factor model for the arrival rates, and then forecasting the time series of factor scores. As another example, Aldor-Noiman et al. (2009) propose an arrival count model which is based on a mixed Poisson process approach incorporating day-of-week, periodic, and exogenous effects. For other related work, see Avramidis et al. (2004), Brown et al. (2005), and references therein.

We might also rely on historical data from previous days to estimate the mean time between successive service completions, combined with real-time data over the recent past. However, we consider a procedure based on real-time estimation alone, and investigate its feasibility. As a real-time estimator, we propose computing the sample average, $\hat{m}$, of (recent) time intervals between successive service completions in the system. In doing so, as an approximation, we assume (i) that all servers are simultaneously busy, and (ii) that the times between successive service completions are i.i.d. (Since we are interested in systems which are heavily loaded, the assumption of busy servers is not too restrictive. The second assumption is exact for exponential service times, but not more generally.) Given that assumption, we can apply elementary statistics to compute the sample size, $n(x)$, needed to obtain a desired margin of relative error, x, at a given confidence level. (Specifically, the half width of a confidence interval is a function of the number of observations used. Therefore, we can obtain a desired margin of relative error by changing the number of observations, thus leading to a different half width.) The error, x, measures the relative error between the actual mean and the sample mean.

To illustrate, consider the $M_t/M/100$ model with exponential service times. Then, $n(0.05) \approx 1540$ at the 95% confidence level. That is, the sample size required to get a relative error margin of $x = 0.05$ is roughly equal to 1540, at the 95% confidence level. It is important to get a sense of how long it would take to get a total of 1540 service completions in the system. For example, suppose that the mean service time is equal to 5 minutes. The length of the estimation interval is roughly equal to 77 minutes. Indeed, each service request requires, on average, 5 minutes to process, and there are 100 servers working in parallel. This numerical example illustrates that the computational burden of obtaining estimates of system parameters that are within a relative error margin of $x = 0.05$ of their actual values is not unreasonable.

There remains to study the effect of the estimation error, x, on the performance of the HOL$_r$ estimator. To that end, we consider modified HOL$_r$ delay estimators, denoted by HOL$_r(x)$, depending on the relative error, x, in estimating $1/s\mu$. That is, the HOL$_r(x)$ estimators use the following delay estimate:

$$\theta_{HOL_r}(t, x, w) = \frac{1+x}{s\mu}(2 + \int_{t-w}^{t} \lambda(u)du) \ ,$$

where $-1 < x < 1$, and $(1+x)/s\mu$ is our estimate of the mean time between successive service completions, including a relative error x. We study the performance of HOL$_r(x)$ for alternative small values of x. Clearly, the performance of HOL$_r(x)$ should degrade as $|x|$ increases, but we would like to know by how much.

In Table 3, we study the performance of HOL$_r(x)$ as a function of the traffic intensity, ρ, in the $M_t/M/100$ queueing model, with $\alpha = 0.5$ and $E[S] = 5$ minutes. We also include the sample sizes needed to obtain system parameter estimates within that error margin and, in parentheses, the corresponding required length of the estimation interval (under our model

$M_t/M/100$, $\alpha = 0.5$, $E[S]$ = *5 min*

	$HOL_r(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL$_r$	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	4.40 $\pm 5.3\times10^{-2}$	1.24 $\pm 2.53\times10^{-2}$	0.449 $\pm 1.21\times10^{-2}$	0.302 $\pm 6.4\times10^{-3}$	0.417 $\pm 9.3\times10^{-3}$	1.02 $\pm 2.1\times10^{-2}$	2.96 $\pm 4.1\times10^{-2}$	0.148 $\pm 6.8\times10^{-3}$	16.92 $\pm 1.4\times10^{-1}$
0.93	6.01 $\pm 5.0\times10^{-2}$	1.63 $\pm 2.9\times10^{-2}$	0.548 $\pm 1.5\times10^{-2}$	0.351 $\pm 8.8\times10^{-3}$	0.520 $\pm 1.5\times10^{-2}$	1.37 $\pm 3.4\times10^{-2}$	4.09 $\pm 7.2\times10^{-2}$	0.177 $\pm 6.0\times10^{-3}$	28.0 ±0.27
0.95	7.29 $\pm 9.3\times10^{-2}$	1.96 $\pm 3.7\times10^{-2}$	0.645 $\pm 1.7\times10^{-2}$	0.410 $\pm 1.8\times10^{-2}$	0.620 $\pm 2.8\times10^{-2}$	1.66 $\pm 4.5\times10^{-2}$	4.98 $\pm 7.1\times10^{-2}$	0.202 $\pm 7.4\times10^{-3}$	38.06 ±0.32
0.97	8.48 ±0.12	2.21 $\pm 5.5\times10^{-2}$	0.688 $\pm 2.4\times10^{-2}$	0.431 $\pm 1.4\times10^{-2}$	0.702 $\pm 2.7\times10^{-2}$	1.97 $\pm 5.7\times10^{-2}$	5.96 ±0.11	0.216 $\pm 6.6\times10^{-3}$	49.8 ±0.43
0.98	9.21 $\pm 8.2\times10^{-2}$	2.40 $\pm 3.5\times10^{-2}$	0.741 $\pm 2.3\times10^{-2}$	0.454 $\pm 2.3\times10^{-2}$	0.737 $\pm 3.0\times10^{-2}$	2.09 $\pm 4.4\times10^{-2}$	6.39 $\pm 7.4\times10^{-2}$	0.226 $\pm 6.9\times10^{-3}$	56.3 ±0.40
Sample size Est. interval	385 (20 min.)	1537 (77 min.)	9604 (480 min.)		9604 (480 min.)	1537 (77 min.)	385 (20 min.)		

Table 3: Performance of HOL$_r(x)$ delay estimators, as a function of the traffic intensity, ρ, and alternative x, in the $M_t/M/100$ queueing model with $\alpha = 0.5$ and $E[S] = 5$ minutes. Sample sizes needed and length of estimation intervals required are also included. Estimates of the ASE's are given in units of mean service time squared per customer.

assumptions). We consider values of x between -0.1 and 0.1. For these values, we find that HOL$_r$ still performs considerably better than HOL. For example, for $x = 0.05$, the ratio ASE(HOL)/ASE(HOL$_r(x)$) ranges from about 14 to about 23 for values of ρ between 0.9 and 0.98. For $x = -0.05$, ASE(HOL)/ASE(HOL$_r(x)$) ranges from about 16 to about 27 for ρ between 0.9 and 0.98. That is, simulation shows that HOL$_r$ remains remarkably more effective than HOL, even with imperfect information about system parameters, as would commonly occur in practice.

Additional simulation results are presented in the online supplement to the main paper. There, we consider lognormal and deterministic service times, and alternative arrival-rate parameters. We find that HOL$_r(x)$ usually performs better than HOL when the relative error, x, is at most 5%. For example, in the $M_t/H_2/100$ model with $\alpha = 0.5$, $E[S] = 6$ hours, and $x = -0.05$, the ratio ASE(HOL)/ASE(HOL$_r(x)$) ranges from 2.4 to 2.8.

8. Delay Estimators for the $M_t/GI/s + GI$ Model

In this section, we propose a new delay estimator for the $M_t/GI/s + GI$ model, based on the HOL delay observed upon arrival to the system. In §9 we show that this new estimator, QL_h, performs remarkably well. In particular, QL_h effectively copes with both time-varying arrivals and non-exponential abandonment-time distributions. As a frame of reference, we also consider a delay estimator based on the queue-length seen upon arrival to the system. This estimator, QL_m, was previously considered in Whitt (1999a) and Ibrahim and Whitt (2009b).

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

The Approximation-Based QL-Based Delay Estimator ($\mathbf{QL}_{ap}$). In Ibrahim and Whitt (2009b), we introduced an approximation-based queue-length-based delay estimator, QL_{ap}, which exploits established approximations for performance measures in the $M/GI/s+GI$ model, developed by Whitt (2005). We showed that QL_{ap} consistently outperforms all other estimators considered in the $GI/GI/s + GI$ model, with a stationary arrival process. Here, we propose an analog of QL_{ap} that uses the observed HOL delay, and effectively copes with time-varying arrival rates. We begin by briefly reviewing the QL_{ap} estimator for the

$GI/GI/s+GI$ model; a more complete description can be found in §3.5 of Ibrahim and Whitt (2009b) and Whitt (2005).

The QL$_{ap}$ estimator approximates the $GI/GI/s{+}GI$ model by the corresponding $GI/M/s+M(n)$ model, with state-dependent Markovian abandonment rates. In particular, we assume that a customer who is jth from the *end* of the queue has an exponential abandonment time with rate ψ_j, where ψ_j is given by

$$\psi_j \equiv h(j/\lambda), \quad 1 \leq j \leq k \; ; \tag{8.1}$$

k is the current queue length, λ is the arrival rate (assumed constant), and h is the abandonment-time hazard-rate function, defined as $h(t) \equiv f(t)/(1-F(t))$, $t \geq 0$, where f is the corresponding density function (assumed to exist). Here is how (8.1) is derived: If we knew that a given customer had been waiting for time t, then the rate of abandonment for that customer, at that time, would be $h(t)$. We therefore need to estimate the elapsed waiting time of that customer, given the available state information. Assuming that abandonments are relatively rare compared to service completions, it is reasonable to act as if there have been j arrival events since our customer arrived. Since a simple rough estimate for the time between successive arrival events is the reciprocal of the arrival rate, $1/\lambda$, the elapsed waiting time of is approximated by j/λ and the corresponding abandonment rate by (8.1).

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

$$W_Q(n) \approx \sum_{i=0}^{n} X_i \; , \tag{8.2}$$

where X_{n-i} is the time between the ith and $(i+1)$st departure events. Since the distribution of the X_i's is complicated, we assume that successive departure events are either service completions, or abandonments from the head of the line. We also assume that an estimate of the time between successive departures is $1/\lambda$. Under our first assumption, after each departure, all customers remain in line except the customer at the head of the line. The elapsed waiting time of customers remaining in line increases, under our second assumption, by $1/\lambda$. Let X_{n-l}, which is the time between the lth and $(l+1)$st departure events, have an exponential distribution with rate $s\mu + \delta_n - \delta_l$, where $\delta_k = \sum_{j=1}^{k} \psi_j = \sum_{j=1}^{k} h(j/\lambda)$, $k \geq 1$, and $\delta_0 \equiv 0$. That is the case because X_{n-l} is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and $n-l$ exponential random variables with rates ψ_i , $l+1 \leq i \leq n$ (corresponding to the abandonment times of the customers waiting in line).

The QL$_{ap}$ delay estimate given to a customer who finds n customers in queue upon arrival is

$$\theta_{QL_{ap}}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}} \; ; \tag{8.3}$$

that is, $\theta_{QL_{ap}}(n)$ approximates the mean of the potential waiting time, $E[W_Q(n)]$.

The QL$_h$ Estimator. We are now ready to propose a new delay estimator for the $M_t/GI/s+GI$ model, which we refer to as QL$_h$. This estimator requires knowledge of the abandonment-time hazard-rate function, h. That is convenient from a practical point of view, because it is relatively easy to estimate hazard rates from system data; see Brown et al. (2005).

We proceed in two steps: (i) we use the observed HOL delay, w, to estimate the queue length seen upon arrival, and (ii) we use this queue-length estimate to implement a new delay estimator, paralleling (8.3). Unlike QL$_{ap}$, QL$_h$ exploits the HOL delay, and does not assume knowledge of the queue length seen upon arrival.

For step (i), let $N_w(t)$ be the number of arrivals in the interval $[t-w, t]$ who do not abandon. That is, $N_w(t) + 1$ is the number of customers seen in the queue upon arrival at time t, given that the observed HOL delay at t is equal to w. It is significant that N_w has the structure of the number in system in a $M_t/GI/\infty$ infinite-server system, starting out empty in the infinite past, with arrival rate $\lambda(u)$ identical to the original arrival rate in $[t-w, t]$ (and equal to 0 otherwise). The individual service-time distribution is identical to the abandonment-time distribution in our original system. Thus, $N_w(t)$ has a Poisson distribution with mean

$$m(t,w) \equiv E[N_w(t)] = \int_{t-w}^{t} \lambda(s)(1 - F(t-s))ds \ , \tag{8.4}$$

where F is the abandonment-time cdf.

For step (ii), we use $m(t,w)+1$ as an estimate of the queue length seen upon arrival, at time t. In (8.1), we replace λ by $\hat{\lambda}$, where $\hat{\lambda}$ is defined as the average arrival rate over the interval $[t-w,t]$, i.e., $\hat{\lambda} \equiv (1/w)\int_{t-w}^{t} \lambda(s)ds$. We do so because we now have a nonstationary arrival process instead of a stationary arrival process. Paralleling (8.3), the QL$_h$ delay estimate given to a customer such that the observed HOL delay, at his time of arrival, t, is equal to w, is given by:

$$\theta_{QL_h}(t,w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s\mu + \hat{\delta}_n - \hat{\delta}_{n-i}} \ , \tag{8.5}$$

for $m(t,w)$ in (8.4), $\hat{\delta}_k = \sum_{j=1}^{k} h(j/\hat{\lambda})$, and $\hat{\delta}_0 = 0$. If we actually know the queue length, then we can replace $m(t,w)$ by $Q(t)$, i.e., we can use QL$_{ap}$. There remains to investigate ways of estimating the abandonment-time distribution needed to implement QL$_h$. We envision that such estimates will be based on long-term estimates of customer time-to-abandon distribution, instead of real-time information about customer abandonment times. Providing additional details relating to this estimation is outside the scope of this paper, and is left for future research.

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

In this section, we present simulation results for the $M_t/M/s + GI$ model with sinusoidal arrival rates. For the abandonment-time distribution, we considered M (exponential), E_{10} (Erlang, sum of 10 exponentials) and H_2 (hyperexponential with SCV equal to 4 and balanced means), but here we only discuss the first two cases; see Ibrahim and Whitt (2009c) for a discussion of the H_2 case. We consider the QL$_m$, QL$_h$, and HOL delay estimators. In this section, we show plots of the simulation results. Corresponding tables with estimates of 95%

confidence intervals, in addition to more simulation results, appear in Ibrahim and Whitt (2009c).

Description of the Experiments. We vary the number of servers, s, but consider only relatively large values ($s \geq 100$), because we are interested in large service systems. We let the service rate, μ, be equal to 1. For the arrival rate function, $\lambda(u)$ in (5.5), we fix the relative frequency, $\gamma = 1.571$. This value of γ corresponds to a mean service time $E[S] = 6$ hours, for daily arrival-rate cycles; see Table 1.

We consider a relative amplitude $\alpha = 0.5$, and an average arrival rate $\bar{\lambda} = 140$. The instantaneous offered load in the system, at time t, is given by $\lambda(t)/s\mu$. With $\alpha = 0.5$, the offered load varies between 0.7 and 2.1. Because of customer abandonment, the congestion is not extraordinarily high when the system is significantly overloaded. We let the abandonment rate, $\nu = 1$, because that seems to be a representative value. Simulation results for all models are based on 10 independent replications of length 1 month each, assuming a daily cycle.

Results for the $M_t/M/s+M$ model. Consistent with theory in §8, Figure 3 shows that QL_m is the best possible estimator, under the MSE criterion. The RRASE of QL_m ranges from about 14% for $s = 100$ to about 4% when $s = 1000$. Figure 3 shows that $s \times \text{ASE}(\text{QL}_m)$, the ASE of QL_m multiplied by the number of servers s, is nearly constant for all values of s considered. This shows that QL_m is asymptotically correct as s increases, i.e., $\text{ASE}(\text{QL}_m)$ approaches 0 as s increases.

The QL_h estimator is the second best estimator for this model. The RRASE of QL_h ranges from about 20% for $s = 100$ to about 6% for $s = 1000$. That is, QL_h is relatively accurate for this model. The difference in performance between QL_h and QL_m is not too great: $\text{ASE}(\text{QL}_h)/\text{ASE}(\text{QL}_m)$ is close to 1.6, for all s. Moreover, Figure 3 shows that QL_h is asymptotically correct: $s \times \text{ASE}(\text{QL}_h)$ is also roughly equal to a constant, for all s.

The HOL estimator performs much worse than QL_m and QL_h. For example, the ratio $\text{ASE}(\text{HOL})/\text{ASE}(\text{QL}_h)$ ranges from about 3 for $s = 100$ to about 20 for $s = 1000$. The RRASE of HOL ranges from about 33% for $s = 100$ to about 27% for $s = 1000$. That is, we

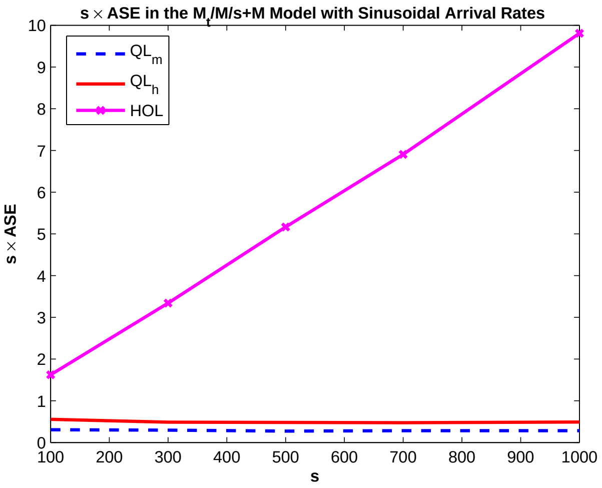


Figure 3: $E[S] = 6$ hours, $\alpha = 0.5$

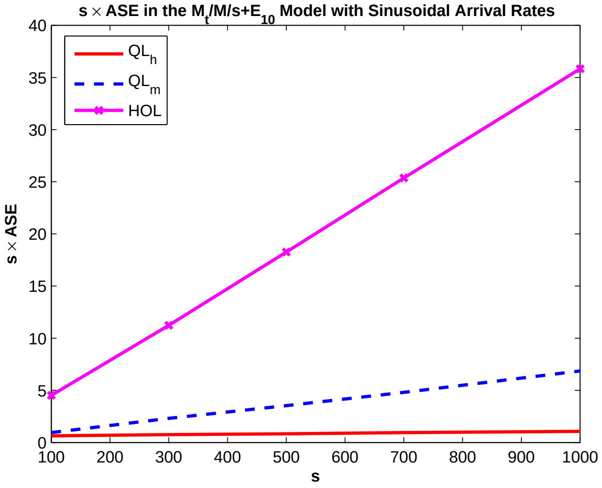


Figure 4: $E[S] = 6$ hours, $\alpha = 0.5$

do not see a considerable improvement in the performance of HOL, as s increases. That is confirmed by Figure 3, where we see that $s\times$ ASE(HOL) increases linearly, as s increases.

Results for the $M_t/M/s+E_{10}$ model. The QL$_h$ estimator is the most effective estimator, under the MSE criterion, for this model. The RRASE of QL$_h$ ranges from about 11% for $s = 100$ to about 4% for $s = 1000$. That is, QL$_h$ is relatively accurate for this model. Figure 4 shows that QL$_h$ is asymptotically correct: $s\times$ ASE(QL$_h$) is roughly equal to a constant for all values of s considered.

The QL$_m$ estimator performs significantly worse than QL$_h$, with E_{10} abandonment. The ratio ASE(QL$_m$)/ASE(QL$_h$) ranges from about 1.5 for $s = 100$ to about 6.5 for $s = 1000$. The RRASE of QL$_m$ ranges from about 13% for $s = 100$ to about 10% for $s = 1000$. Figure 4 shows that QL$_m$ is not asymptotically correct as s increases.

The least effective estimator is, yet again, the HOL estimator. The RRASE of HOL ranges from about 27% for $s = 100$ to about 25% for $s = 1000$. The difference in performance between HOL and QL$_h$ is remarkable: ASE(HOL)/ASE(QL$_h$) ranges from roughly 7 for $s = 100$ to roughly 33 for $s = 1000$. Figure 4 shows that $s\times$ ASE(HOL) increases linearly (and steeply) as s increases.

Results for Other Models. We consider general service-time and abandonment-time distributions in Ibrahim and Whitt (2009c). For the service-time distribution, we consider M, D, and H_2. For the abandonment-time distribution, we consider M, H_2, and E_{10}. We consider different combinations of service-time and abandonment-time distributions. These additional simulation results are consistent with those reported above: The QL_m estimator remains effective with M abandonment, even when the service-time distribution is not nearly exponential. With H_2 and E_{10} abandonment, QL_h outperforms QL_m, especially when the number of servers is large. The HOL estimator remains the least effective estimator, under the MSE criterion, in all models considered.

10. Conclusions

In this paper, we studied the performance of alternative delay estimators in the $M_t/GI/s$ and $M_t/GI/s + GI$ queueing models, which have a nonhomogeneous Poisson process. We concentrated on the HOL estimator, which is equal to the elapsed delay of the customer at the head of the line, at the time of arrival. We did so with the understanding, based on our previous work, that results for HOL should apply equally well to the delay of the last customer to enter service (LES). A main conclusion is that the performance of these delay-history-based delay estimators can degrade in face of time-varying arrivals, which often occurs in practice; that is dramatically shown in Figure 2.

As a consequence, we developed refinements of HOL, in particular, HOL_r in (4.2) for $M_t/GI/s$ and QL_h in (8.5) for $M_t/GI/s+GI$. Simulation experiments in §6 and §9 showed that these estimators effectively cope with both time-varying arrivals and non-exponential service-time and abandon-time distributions. We also established analytical results supporting HOL_r in §5. We quantified the difference in performance between QL and HOL_r and found that the ratio of their respective MSE's is roughly equal to 2, especially for high values of the traffic intensity, ρ; see (5.6).

However, the new refined estimators lose some of their appeal compared to the simple HOL and LES estimators, because they require information about the model, in particular, the arrival-rate function and the mean time between successive departures. Hence, in §7 we proposed ways to estimate the required information. Even if we rely on real-time estimation of the mean time between successive departures, we showed that we can obtain suitably accurate estimates without requiring that the observation interval be too long. Table 3 shows that the HOL$_r$ estimator remains effective even if the information is known imperfectly.

Our general strategy for creating the refined HOL estimators has been to approximate the mean conditional delay, given the observed HOL delay by (i) approximating the queue length, given the observed HOL delay, and (ii) approximating the expected delay given the queue length. As a consequence, direct queue-length-based delay estimators should be preferred if the queue length is known. However, in §1.2 we observed that there are complex service systems such as Web chat and ticket queues for which the queue length is not known.

Acknowledgments.

This research was supported by NSF grants DMI-0457095 and CMMI 0948190.

References

Aksin, O.Z., Armony, M. and Mehrotra, V. 2007. The Modern Call-Center: A multi-disciplinary perspective on operations management research, *Production and Operations Management*, 16:6, 665–688.

Aldor-Noiman, S. 2006. Forecasting demand for a telephone call center: Analysis of desired versus attainable precision. Unpublished masters thesis, Technion-Israel Institute of Technology, Haifa, Israel.

Allon, G, Bassambo, A. and I. Gurvich. 2009. We will be right with you: managing customer with vague promises, *Working Paper*, Northwestern Univ., Evanston, IL..

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

Armony, M., N. Shimkin and W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonments. *Operations Research.* 57: 66–81.

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

Barlow, R. E. , F. Proschan. 1975. *Statistical Theory of Reliability and Life Testing*, Holt, Rinehart and Winston, New York.

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

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

Green, L., Kolesar, P., and W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* (POMS), 16: 13–39.

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

Heyman, D. and W. Whitt. 1984. The asymptotic behavior of queues with time-varying arrival rates. *Journal of Applied Probability* 21: 143–156.

Ibrahim, R. and W. Whitt. 2009a. Real-time delay estimation based on delay history. *Manufacturing and Service Oper. Mgmt.* 11: 397–415.

Ibrahim, R. and W. Whitt. 2009b. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science.* 55: 1729–1742.

Ibrahim, R. and W. Whitt. 2009c. Supplement to "Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals", IEOR Department, Columbia University, New York, NY. Available at http://columbia.edu/∼rei2101.

Jongbloed, G., and G. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Appl. Stochastic Models Bus. Indust.* 17: 307–318.

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

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

Massey, W., and W. Whitt. 1994. A stochastic model to capture space and time dynamics in wireless communication systems. *Probability in the Engineering and Informational Sciences*, 8: 541–569.

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

Ross, S. 1996. *Stochastic Processes.* (2nd ed.), New York: Wiley.

Shen, H. and J. Huang. 2008a. Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Oper. Mgmt* 10:3.

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

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

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

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