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

by
Rouba Ibrahim and Ward Whitt

IEOR Department
Columbia University
500 West 120th Street 313 Mudd
New York, NY 10027
{rei2101, ww2040}@columbia.edu

Abstract

We develop new improved real-time delay estimators, based on recent customer delay history, in many-server service systems with time-varying arrivals, both with and without customer abandonment. These delay estimators may be used to make delay announcements. We model the arrival process by a nonhomogeneous Poisson process, which has a deterministic time-varying arrival-rate function. Our estimators effectively cope with time-varying arrivals 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 17, 2009

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. With invisible queues, such as in call centers, waiting customers are unable to estimate their own delay, and would therefore gain from additional delay estimates; see Gans et al. (2003) and Aksin et al. (2007) for background on call centers. Delay announcements may be especially helpful with emergency services, such as in a hospital's emergency department (ED).

The accepted model for capturing time-varying arrivals is a nonhomogeneous Poisson arrival process; such a process is completely characterized by its deterministic arrival-rate intensity function. 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, which is what we do in this paper. With a nonhomogeneous Poisson arrival process, we are able to obtain analytical results, and to propose simple and effective delay estimators. Our results provide useful background for similar studies in even more complicated settings.

When variability in the arrival process is slow over time, relative to the service times, it is customary to assume stationarity of the process in short (e.g., 30 minute), disjoint intervals of

time. In this case, the analysis of the system reduces to that of a stationary system; e.g., see Green et al. (2007). For an empirical study of the effectiveness of stationary approximations with sinusoidal arrival rates, see Green and Kolesar (1991). Here, we are interested in systems where the arrival rate is moderately or highly variable, so that stationary approximations perform poorly. In particular, we are interested in systems periodically alternating between phases of overload and underload, as is often encountered in real-life service systems.

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. 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 used as an announcement in an Israeli bank studied by Mandelbaum et al. (2000), and was mentioned as a candidate delay announcement by Nakibly (2002).

In previous work, 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 (but without considering customer response). Through analysis and extensive simulation experiments, we concluded that the LES and HOL estimators are very similar. Here, we only discuss HOL, and not LES, because the conditional distribution of the delay to be estimated is more tractable given the HOL information. Our results for HOL should apply equally well to LES.

The HOL estimator is appealing because it does not depend on the model and uses very little information about the system. It is robust because it responds automatically to changes in system parameters (e.g., number of servers, mean service time, and arrival rate). That is important because system parameters, in real-life systems, often change over time. Indeed, servers are humans who serve in different shifts and may well have different service-time distributions.

Changes in system parameters could also result from customer response to delay announcements: Customers typically respond to delay announcements, and their response alters system performance. For example, some customers may elect to balk, upon arrival, in response to a delay announcement. As a result, the arrival rate to the system would become state dependent. Changes in system performance in turn alter the delay estimates given. Delay-history-based estimators automatically account for customer response because they depend on the history of delays in the system, which in turn is affected by customer response. Therefore, delay-history-based estimators are appealing, from a practical point of view.

1.2. The HOL Estimator with Time-Varying Arrival Rates

In a first paper, Ibrahim and Whitt (2009a), we studied the performance of the HOL delay estimator in the $GI/M/s$ queueing model. That model has a renewal arrival process, s homogeneous servers, and an unlimited waiting room. Service times are independent and identically distributed (i.i.d.) exponential random variables. Customers are served in order of arrival, i.e., according to the first-come-first-served (FCFS) service discipline.

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, QL, which multiplies the queue length plus one times the mean interval between successive service completions, ignoring customer abandonment. The QL estimator is provably the most effective estimator, under the mean squared error (MSE) criterion (see (2.2)), with i.i.d. exponential service times, and no customer abandonment; see §3 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.

In a second paper, Ibrahim and Whitt (2009b), we considered the $GI/GI/s+GI$ model, which includes customer abandonment. This model has i.i.d. service times and abandonment times with general distributions. Intuitively, we should expect that QL will overestimate customer delay when there is significant customer abandonment in the system. Consistent with intuition, we showed that QL performs poorly in a heavily loaded $GI/GI/s+GI$ model. We also showed that HOL remains an effective estimator in this more general setting.

With time-varying arrivals, the HOL estimator may not be an effective estimator. Intuitively, we should expect that HOL can perform poorly when the arrival rate changes rapidly over time, because the delays may vary systematically over time. To illustrate the potential deficiency of the HOL estimator, we plot simulation sample paths of HOL delay estimates given, and actual delays observed, as a function of time, in two given simulation runs. In Figure 1, we consider the stationary $M/M/100$ model. In Figure 2, we consider the $M_t/M/100$ with sinusoidal arrival rates. (The model will be fully specified later.) We deliberately choose an extreme case where the arrival rate varies significantly with respect to the service times, while the number of servers remains fixed. This case serves to illustrate how poorly the HOL estimator can perform.

Figure 1 shows that, with a stationary arrival process, the HOL delay estimates agree closely with the actual delays observed in the system. In contrast, Figure 2 shows that, with time-varying arrival rates, the HOL curve is clearly shifted to the right, compared to the actual-delays curve. That is, there is a time lag between the HOL estimates and the actual delays observed. Figures 1 and 2 nicely illustrate the deterioration in performance of the HOL estimator that may occur with time-varying arrival rates.

In this paper, we show that HOL may not be an effective estimator with time-varying arrivals, particularly when the system alternates between phases of underload and overload. In this paper, we develop refinements of the HOL estimator that remain effective for time-varying arrival rates. These refinements exploit knowledge of the arrival-rate function.

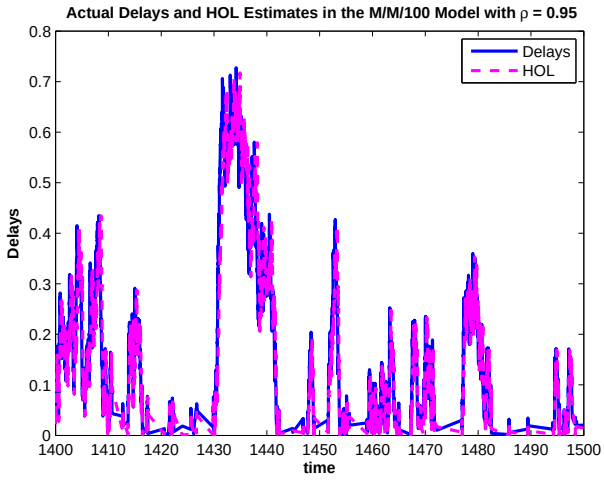

Figure 1: Sample paths of actual delays and HOL delay estimates

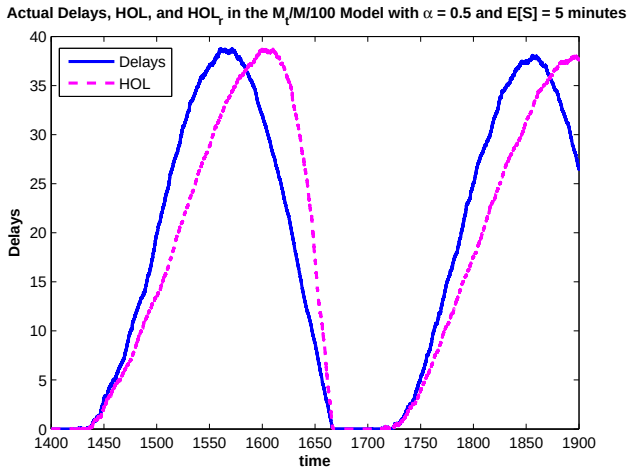

Figure 2: Sample paths of actual delays and HOL delay estimates

Through analysis and simulation, we show that these new estimators perform remarkably well with time-varying arrival rates, far better than HOL.

In on-going work, we consider alternative delay estimators based on the queue length seen upon arrival, in many-server systems with time-varying arrivals and customer abandonment.

1.3. The Queueing Models

In §§3-5, we consider the $M_t/GI/s$ model, not allowing customer abandonment. This model has a nonhomogeneous Poisson arrival process with an arrival-rate intensity function $\lambda \equiv \{\lambda(u) : -\infty < u < \infty\}$. Let $\bar{\lambda}$ denote the average arrival rate, defined as

$$\bar{\lambda} = \lim_{t\to\infty} \frac{1}{t} \int_0^t \lambda(u)du \ ,$$

which we assume is well defined. In this work, we focus on sinusoidal arrival rates, but our results hold for general arrival rates. The case of sinusoidal arrivals is interesting to understand queues with periodic arrival rates (e.g., daily cycles), which are familiar in practice; for previous investigations of queues with sinusoidal arrival rates, see Eick et al. (1993) and references therein.

In the $M_t/GI/s$ model, the service times S_n are i.i.d. with a general distribution and mean $E[S] = \mu^{-1}$ (we omit the subscript from S when the specific index is not important). Motivated by large service systems, we are primarily interested in the case of large s. The arrival and service processes are independent. We use the FCFS service discipline. The traffic intensity, ρ, is given by $\rho \equiv \bar{\lambda}/s\mu$. We emphasize that, in our models, the number of servers s is fixed. That is, we focus on the scenario where the service provider does not have the resources nor the flexibility to adjust staffing to meet unexpected high (or low) demand, during the day. Considering a time-varying number of servers is an interesting topic, which we leave for future research.

Motivated by applications to real-life service systems, which rarely are as simple as the $M_t/GI/s$ model, we consider the $M_t/GI/s+GI$ model in §6 and §7. Abandonment times are i.i.d. with mean ν^{-1} and a general cumulative distribution function (cdf) F. To capture a wide range of possible of abandonment-time distributions, we consider M (exponential), H_2 (hyperexponential, mixture of two exponentials), and E_{10} (Erlang, sum of ten exponentials) distributions. The H_2 (E_{10}) distribution exhibits high (low) variability, relative to M.

1.4. Actual and Potential Waiting Times

As in Baccelli et al. (1984) and Garnett et al. (2002), we need to distinguish between the *actual* and *potential* waiting times of a given delayed customer in a queueing model with customer abandonment. A customer's actual waiting time is the amount of time that this customer spends in queue, until he either abandons or joins service, whichever comes first. A customer's potential waiting time is the delay he would experience, if he had infinite patience (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.) In this study, we estimate the potential waiting times of delayed customers.

1.5. Literature Review and Main Contributions

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 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, (ii) to provide analytical results quantifying the performance of some delay estimators with time-varying arrivals, and (iii) to describe results of a wide range of simulation experiments evaluating alternative delay estimators, with time-varying arrivals.

1.6. Organization of the Paper

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

2. Performance Measures of 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.

2.1. Quantifying Performance: Average Squared Error (ASE)

In our simulation experiments, we quantify the performance of a delay estimator by computing the *average squared error* (ASE), defined by:

$$ASE \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - e_i)^2 \ , \tag{2.1}$$

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

2.2. Mean Squared Error (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:

$$MSE \equiv MSE(\theta_{HOL}(t,w)) \equiv E[(W_{HOL}(t,w) - \theta_{HOL}(t,w))^2] \ . \tag{2.2}$$

Note that the MSE of a delay-history-based estimator is a function of w and t. Here, we consider a periodic arrival-rate function, as occurs with daily demand cycles. We assume that the system is initially empty. In this setting, we think of the system as being in dynamic steady state, as occurs if the system has been operating for a long period of time; e.g., see Heyman and Whitt (1984). When the cycle length is specified, we can deduce the place where any time t falls within the cycle. By looking at the ASE, we are looking at the

expected MSE averaging over all w where the arrival must wait, and over time t, in dynamic steady state.

2.3. Root Relative Squared Error

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

$$RRASE \equiv \frac{\sqrt{ASE}}{(1/k)\sum_{i=1}^{k} p_i} \ , \tag{2.3}$$

using the same notation as in (2.1). The denominator in (2.3) is the average potential waiting time of customers who must wait. For large samples, the RRASE should agree with the expected *root relative mean squared error* (RRMSE). The RRASE (RRMSE) is useful because it measures the effectiveness of an estimator relative to the mean potential waiting time, given that the customer must wait. It is thus easy to interpret.

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

In this section, we consider the $M_t/GI/s$ model. We propose a new refined HOL-based delay estimator, HOL_r, for this model. In §5, we show that HOL_r is effective. As a frame of reference, we also consider the standard QL delay estimator, briefly discussed in the introduction.

3.1. A Refined HOL (HOL_r) Delay Estimator

As an alternative to the direct HOL estimator, $\theta_{HOL}(t,w) \equiv w$, we want to use the refined estimator $\theta^r_{HOL}(t,w) \equiv E[W_{HOL}(t,w)]$, because the mean necessarily minimizes the MSE. Since we do not have a convenient formula for the mean, in the $M_t/GI/s$ model, we propose the following approximation.

We approximate the $M_t/GI/s$ model by the corresponding $M_t/M/s$ model, with the

same service-time mean. For the $M_t/M/s$ model, we have the representation:

$$W_{HOL}(t,w) \equiv \sum_{i=1}^{A(t)-A(t-w)+2} S_i/s \ , \tag{3.1}$$

where $\{A(t) : t \geq 0\}$ denotes the arrival (counting) process. We have division by s in (3.1) 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 $s\mu$. Since the arrival process is a nonhomogeneous Poisson process, the random variable $A(t) - A(t-w)$ is distributed as a Poisson random variable with mean given by

$$E[A(t) - A(t-w)] = \int_{t-w}^{t} \lambda(u)du \ . \tag{3.2}$$

Since $W_{HOL}(t,w)$ in (3.1) 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 have, for the $M_t/M/s$ model,

$$E[W_{HOL}(t,w)] = \frac{1}{s\mu}(2 + \int_{t-w}^{t} \lambda(u)du) \ . \tag{3.3}$$

The HOL_r delay estimate given to a customer who arrives to the system at time t, such that the elapsed waiting time of the customer at the head of the line is w, is

$$\theta_{HOL_r}(t,w) = \frac{1}{s\mu}(2 + \int_{t-w}^{t} \lambda(u)du) \ , \tag{3.4}$$

which coincides with $E[W_{HOL}(t,w)]$ with exponential service times, but not otherwise. With a non-exponential service-time distribution, θ_{HOL_r} in (3.4) either overestimates or underestimates $E[W_{HOL}(t,w)]$, depending on the stochastic variability of the service-time distribution, relative to the exponential distribution; for background on stochastic order relations between random variables, see Chapter 9 of Ross (1996).

The time between successive departures from service in an $M_t/GI/s$ model is the minimum of remaining service times of customers currently in service. With hyperexponential service times (or service times with any NWUE distribution), the minimum of remaining service times is stochastically more variable than an exponential, so (3.4) underestimates

the mean of the actual delay experienced. With deterministic (or Erlang) service times (or service times with any NBUE distribution), the minimum of remaining service times is stochastically less variable than an exponential, so (3.4) underestimates the mean of the actual delay experienced. It is significant that (3.4) constitutes a bound on the mean actual delay experienced, in these cases. Simulation shows, however, that these bounds are good approximations; HOL_r performs considerably better than HOL, in the $M_t/GI/s$ model, even when the service-time distribution is not nearly exponential; see §5.

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

We now review 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 potential 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. The QL estimator approximates the $M_t/GI/s$ model by the corresponding $M_t/M/s$ model, with the same mean service time. In the $M_t/M/s$ model, as in (3.1), $W_Q(t,n)$ 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 QL estimator is the optimal delay estimator, under the MSE criterion, in the $M_t/M/s$ model. Simulation shows that QL remains effective in the $M_t/GI/s$ model, even when the service-time distribution is not nearly exponential; see §5. In the $M_t/M/s$ model, we obtain analytical results quantifying the difference in performance between QL and HOL_r. We derive these results next.

4. Heavy-Traffic Limits for the $M_t/M/s$ Model

For the $M_t/M/s$ model, we obtain analytical results characterizing the performance of the QL and HOL_r delay estimators. In §5, we describe results of simulation experiments for the $M_t/GI/s$ model, quantifying the performance of QL, HOL, and HOL_r. More simulation

results appear in Ibrahim and Whitt (2009c).

4.1. Analysis of $W_Q(t,n)$ and $W_{HOL}(t,w)$

We have the representation

$$W_Q(t,n) \equiv \sum_{i=1}^{n+1} S_i/s \ . \tag{4.1}$$

As discussed in Ibrahim and Whitt (2009a), $W_Q(t,n)$ has the desirable property that the estimation gets relatively more accurate as the observed queue length n increases. 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:

$$E[W_Q(t,n)] = \frac{n+1}{s\mu}, \ \ Var[W_Q(t,n)] = \frac{n+1}{s^2\mu^2}, \quad \text{and} \quad c^2_{W_Q(t,n)} \equiv \frac{Var[W_Q(t,n)]}{(E[W_Q(t,n)])^2} = \frac{1}{n+1} \ , \tag{4.2}$$

so that $c^2_{W_Q(t,n)} \to 0$ as $n \to \infty$. As discussed in §3, the QL estimator is the best possible estimator, under the MSE criterion, in the $M_t/M/s$ model. To treat HOL_r, we use the representation in (3.1), which allows us to characterize the probability distribution of the random variable $W_{HOL}(t,w)$, in the $M_t/M/s$ model.

Proposition 1. *For the $M_t/M/s$ model,*

$$Var[W_{HOL}(t,w)] = \frac{2}{s^2\mu^2}(1 + \int_{t-w}^{t} \lambda(u)du) \ , \tag{4.3}$$

which, combined with (3.3), yields

$$c^2_{W_{HOL}(t,w)} = \frac{Var[W_{HOL}(t,w)]}{E[W_{HOL}(t,w)^2]} = 2 \times \frac{1 + \int_{t-w}^{t} \lambda(u)du}{(2 + \int_{t-w}^{t} \lambda(u)du)^2} \ . \tag{4.4}$$

Proof. Formula (4.3) follows from the conditional variance formula, e.g., p.51 of Ross (1996). Formula (4.4) immediately follows from (3.3) and (4.3). ∎.

Since $\theta_{HOL_r}(t,w) \equiv E[W_{HOL}(t,w)]$ and $\theta_{QL}(t,n) \equiv E[W_{QL}(t,n)]$, we can compare the performance of HOL_r and QL by comparing the respective SCV's in (4.2) and (4.4). (When the delay estimate equals the conditional mean, the MSE coincides with the variance.) To this end, we need to specify the behavior of the arrival-rate intensity function, λ. In this paper, we consider sinusoidal arrivals.

4.2. The $M_t/M/s$ model with sinusoidal arrival rates

We consider a sinusoidal arrival-rate intensity function

$$\lambda(u) = \bar{\lambda} + \beta \sin(\gamma u) \equiv \bar{\lambda} + \bar{\lambda}\alpha \sin(2\pi u/\Gamma), \quad \text{for } -\infty < u < \infty \ , \tag{4.5}$$

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 $0 \le u \le \Gamma$, which describes a full cycle. With sinusoidal arrival rates, we obtain analytical results comparing the performance of the QL and HOL_r estimators.

Proposition 2. *For the $M_t/M/s$ model with sinusoidal arrival rates,*

$$\frac{c^2_{W_{HOL(t,w)}}}{c^2_{W_{QL}(n)}} \to \frac{2}{\rho} \ as \ n \to \infty \ , \tag{4.6}$$

for all t and w.

Proof. Using Equations (3.3), (4.3), and (4.4), together with (4.5), 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:

$$E[W_{HOL}(t,w)] = \frac{2 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))}{s\mu} \ , \tag{4.7}$$

and,

$$Var[W_{HOL}(t,w)] = 2 \times \frac{1 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t)))}{s^2\mu^2} \ , \tag{4.8}$$

which yields

$$c^2_{W_{HOL}(t,w)} = \frac{Var[W_{HOL}(w)]}{E[W_{HOL}(w)]^2} = 2 \times \frac{1 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))}{[2 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))]^2} \ , \qquad (4.9)$$

for $0 \leq t \leq \Gamma$. Using (4.9), and recalling that $-1 \leq \cos(u) \leq 1$ for all u, we obtain the following bounds for the SCV of $W_{HOL}(t, w)$:

$$\frac{2 + 2\bar{\lambda}w - 4\beta/\gamma}{(2 + \bar{\lambda}w + 2\beta/\gamma)^2} \leq c^2_{W_{HOL}(t,w)} \leq \frac{2 + 2\bar{\lambda}w + 4\beta/\gamma}{(2 + \bar{\lambda}w - 2\beta/\gamma)^2} \ . \qquad (4.10)$$

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

$$W(t) = \sum_{i=1}^{Q(t)+1} S_i/s \ , \qquad (4.11)$$

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 $W(t) \approx Q(t)/s$. Assuming that n in (4.2) is large with $w \approx n/s\mu$, and combining that with (4.10), we get that, for large n

$$\frac{(2 + 2\rho n - 4\beta/\gamma)(n+1)}{(2 + n\rho + 2\beta/\gamma)^2} \leq \frac{c^2_{W_{HOL}(t,w)}}{c^2_{W_{QL}(n)}} \leq \frac{(2 + 2\rho n + 4\beta\gamma)(n+1)}{(2 + \rho n - 2\beta/\gamma)^2} \ , \qquad (4.12)$$

for all t. By a sandwiching argument, (4.12) yields (4.6) as $n \to \infty$. ∎

Equation (4.6) quantifies the difference in performance between HOL_r and QL in the $M_t/M/s$ model, with sinusoidal arrival rates. We note that (4.6) coincides with formula (4.25) of Ibrahim and Whitt (2009a), for the stationary $GI/M/s$ model.

5. 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).

5.1. Description of the Experiments

We fix the number of servers, $s = 100$, because we are interested in large service systems. 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 heavy load and underload, which is a common case in practice. 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.

As pointed out by Eick et al. (1993), the parameters of the arrival-rate intensity function, $\lambda(u)$ in (4.5), should be interpreted relative to the mean service time, $E[S]$. We let the service rate, μ, be equal to 1. We do this without loss of generality, since we are free to choose the time units in our system, and this assumption amounts to measuring time in units of mean service time. Then, we speak of γ as the relative frequency. Small (large) values of γ correspond to slow (fast) time-variability in the arrival process, relative to the service times. Table 1 displays values of the relative frequency as a function of $E[S]$, assuming 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 hospital's ED. Indeed, treatment times for less critically ill patients in the ED should be in the order of a few hours. Note that with

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 $E[S] = 1$.

$E[S] = 30$ 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.

5.2. Revisiting Figure 2

To illustrate the difference in performance between HOL and HOL$_r$, we consider the $M_t/M/100$ model with $E[S] = 5$ minutes, and $\bar{\lambda} = 95$. We consider both $\alpha = 0.1$, and $\alpha = 0.5$. (The case with $\alpha = 0.5$ was previously considered in Figure 2.) The instantaneous offered load in the system, at time t, is given by $\lambda(t)/s\mu$. With $\alpha = 0.1$, the offered load varies roughly between 0.85 and 1.0. With $\alpha = 0.5$, the offered load varies roughly between 0.5 and 1.4. The case with $\alpha = 0.5$ is extreme, and we consider it to show that the difference in performance between HOL and HOL$_r$ can be dramatic.

In Figures 3 and 4, we plot simulation sample paths of the HOL and HOL$_r$ delay estimates given, and actual delays observed, as a function of time, in two given simulation runs. (With $E[S] = 5$ minutes, an interval of length 500 corresponds to approximately 2 hours.) Figure 3 shows that, with $\alpha = 0.1$ and $E[S] = 5$ minutes, both the HOL and HOL$_r$ delay estimates coincide with the actual delays observed in the system. In this case, ASE(HOL)/ASE(HOL$_r$) is roughly equal to 1.3. Figure 4 shows that, with $\alpha = 0.5$, the HOL curve is clearly shifted

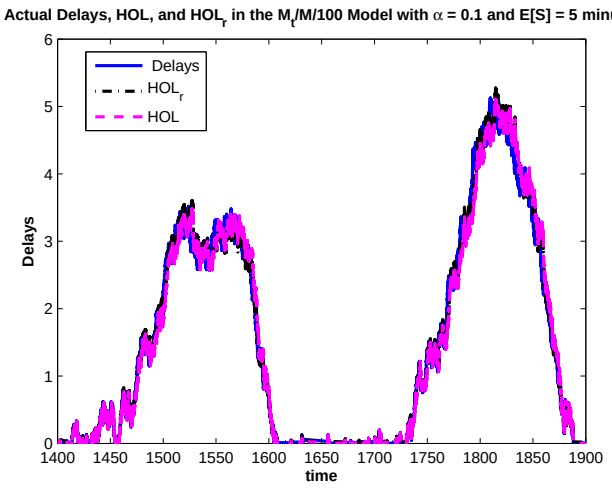

Figure 3: Sample paths of actual delays and corresponding delay estimates

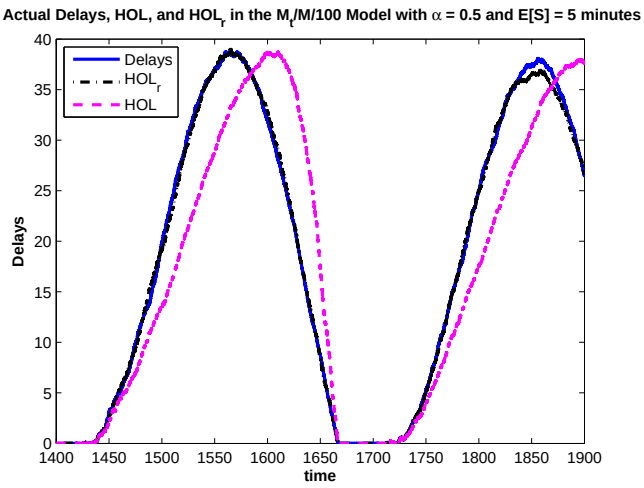

Figure 4: Sample paths of actual delays and corresponding delay estimates

to the right, compared to the actual delays curve, which still coincides with the HOL$_r$ curve. That is, the proposed refinement in (3.4) corrects the estimation error in HOL: ASE(HOL)/ASE(HOL$_r$) is roughly equal to 95 in this case.

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.

5.3. The Difference in Performance Between 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 $\rho = 0.98$.

5.4. The Difference in Performance Between HOL_r and QL

In the $M_t/M/s$ model, QL is provably the optimal estimator, under the MSE criterion; see §3. 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 §4, the approximation in (4.6) is remarkably accurate with M service times, particularly for high values of ρ. 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 (4.6), 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 (4.6). 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.

The HOL_r estimator is appealing because it uses the observed HOL delay, but performs remarkably better than the HOL estimator. That is substantiated by mathematical analysis and multiple simulation experiments. Unfortunately, direct mathematical analysis is substantially harder for the more general $M_t/GI/s + GI$ model, which is more interesting from a practical perspective. Nevertheless, we are able to obtain a new and effective delay estimator, for this more general case. We present this new estimator next.

	$M_t/M/100$, $\alpha = 0.1$, $E[S]$ = *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. Estimated ASE's are in units of 10^{-3}.

6. 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. We show in §7 that this new estimator, QL_h, performs remarkably well. In particular, QL_h effectively copes with time-varying arrivals, and non-exponential abandonment-time distributions. As a frame of reference, we also consider a classical 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).

6.1. The Markovian Queue-Length-Based Delay Estimator (QL_m)

As in Ibrahim and Whitt (2009b), this estimator, QL_m, approximates the $M_t/GI/s+GI$ model by the corresponding $M_t/M/s+M$ model with the same service-time and abandon-time means. For the $M_t/M/s+M$ model, we have the representation:

$$W_Q(t,n) \equiv \sum_{i=0}^{n} Y_i \ , \tag{6.1}$$

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

$$E[W_Q(t,n)] = \sum_{i=0}^{n} E[Y_i] = \sum_{i=0}^{n} \frac{1}{s\mu+i\nu} \ . \tag{6.2}$$

The QL_m estimator given to a customer who finds n customers in queue upon arrival is $\theta_{QL_m}(t,n) \equiv E[W_Q(t,n)]$. Under the MSE criterion, QL_m is the best possible estimator in the $M_t/M/s+M$ model, but we find that it is not always so good for the more general $M_t/GI/s+GI$ model; see §7, and Ibrahim and Whitt (2009c). Thus, there is a need to go beyond QL_m, in practice.

6.2. A New HOL-Based Delay Estimator: QL_h

In Ibrahim and Whitt (2009b), we introduced an approximation-based delay estimator, QL_{ap}, which exploits established approximations for performance measures in the $GI/GI/s+GI$ 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} which uses the observed HOL delay, and effectively copes with time-varying arrival rates. Let QL_h denote this new delay estimator. 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).

6.2.1. The approximation-based queue-length (QL_{ap}) delay estimator

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

$$\psi_j \equiv h(j/\lambda), \quad 1 \le j \le k \; ; \tag{6.3}$$

k is the current queue length, λ is the arrival rate (assumed constant), and h is the abandonment-time hazard-rate function, defined as $h(t) \equiv f(t)/(1-F(t))$, $t \ge 0$, where f is the corresponding density function (assumed to exist). Here is how (6.3) 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, a reasonable estimate is that 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 (6.3).

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:

$$W_Q(n) \approx \sum_{i=0}^{n} X_i \ , \tag{6.4}$$

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 $1/\lambda$. 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 $\delta_0 \equiv 0$. The QL_{ap} delay estimate given to a customer who finds n customers in queue upon arrival is

$$\theta_{QL_{ap}}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}} \ . \tag{6.5}$$

6.2.2. The QL_h estimator

We are now ready to propose a new delay estimator for the $M_t/GI/s + GI$ model. In particular, 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 (6.5). Step (i) is important because the queue length seen upon arrival in the system may not be directly observable. That is nicely illustrated by the 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. It is significant that, 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 $[t-w, w]$ who do not abandon. That is, $N_w(t)+1$ is the number of customers seen in the system 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. As in Eick et al. (1993), $N_w(t)$ has a Poisson distribution with mean

$$m(t,w) \equiv E[N_w(t)] = \int_{t-w}^{t} \lambda(s)(1-F(t-s))ds \ , \tag{6.6}$$

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 (6.3), 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 approximating the arrival process by a stationary process, with constant rate λ, leads to estimation error. Paralleling (6.5), 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:

$$\theta_{QL_h}(t,w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s\mu + \hat{\delta}_n - \hat{\delta}_{n-i}} \ , \tag{6.7}$$

for $m(t,w)$ in (6.6), $\hat{\delta}_k = \sum_{j=1}^{k} h(j/\hat{\lambda})$, and $\hat{\delta}_0 = 0$.

Simulation shows that QL_h performs consistently better than HOL and QL_m in the $M_t/GI/s+GI$ model; see §7. It effectively copes with time-varying arrival rates, and non-exponential service-time and abandonment-time distributions. That is why QL_h is appealing from a practical point of view.

7. 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 consider M (exponential), H_2

(hyperexponential with SCV equal to 4 and balanced means), and E_{10} (Erlang, sum of 10 exponentials). 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).

7.1. Description of the Experiments

We vary the number of servers, s, but consider only relatively large values ($s \geq 100$), because we are interested in large service systems. We let the service rate, μ, be equal to 1. For the arrival rate function, $\lambda(u)$ in (4.5), 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.

7.2. Results for the $M_t/M/s + M$ model

Consistent with theory in §6, Figure 5 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 6 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 6 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

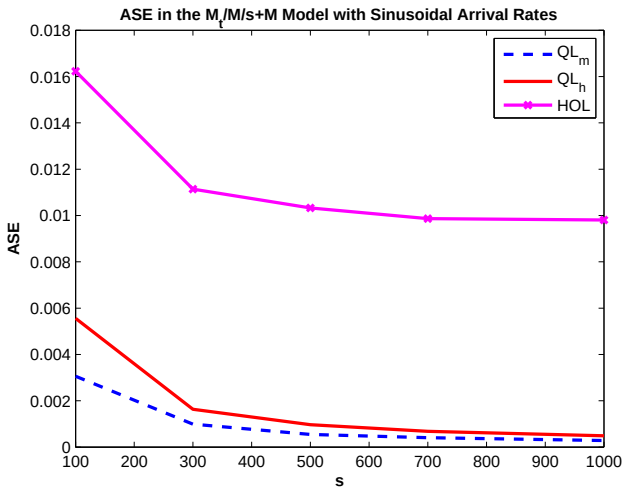

Figure 5: $E[S] = 6$ hours, $\alpha = 0.5$

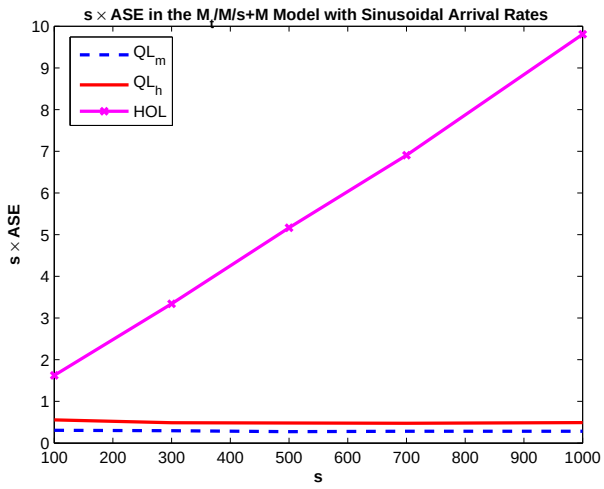

Figure 6: $E[S] = 6$ hours, $\alpha = 0.5$

ASE(HOL)/ASE(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 do not see a considerable improvement in the performance of HOL, as s increases. That is confirmed by Figure 6, where we see that $s\times$ ASE(HOL) increases roughly linearly, as s increases.

7.3. Results for the $M_t/M/s + H_2$ model

With H_2 abandonment, Figure 7 shows that QL_h is the best possible estimator, under the MSE criterion, for large values of s. In particular, QL_h outperforms QL_m for $s \geq 300$. This confirms the need to go beyond QL_m when the abandonment-time distribution is not exponential. The ratio ASE(QL_m)/ASE(QL_h) ranges from about 0.9 for $s = 100$ to about 3 for $s = 1000$. The RRASE of QL_h ranges from about 20% for $s = 100$ to about 6% for $s = 1000$. However, the QL_m estimator remains relatively accurate for this model: RRASE(QL_m) ranges from about 20% for $s = 100$ to about 11% for $s = 1000$.

Figure 7 also shows that QL_h is asymptotically correct as s increases: $s\times$ ASE(QL_h) is roughly constant for all s. In contrast, $s\times$ ASE(QL_m) increases roughly linearly, as s increases, which shows that the performance of QL_m deteriorates as s increases.

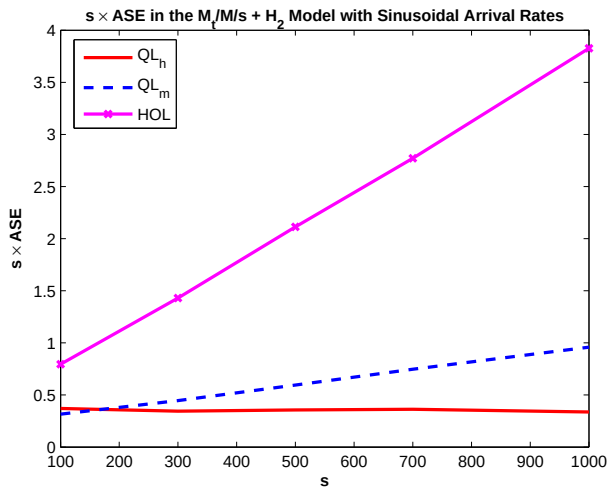

Figure 7: $E[S] = 6$ hours, $\alpha = 0.5$

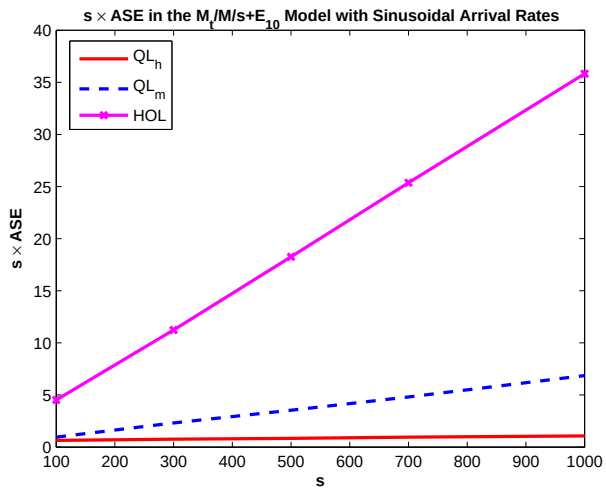

Figure 8: $E[S] = 6$ hours, $\alpha = 0.5$

The HOL estimator is, once more, the least effective estimator for this model. The RRASE of HOL ranges from about 31% for $s = 100$ to about 22% for $s = 1000$. The HOL estimator performs slightly better than with M abandonment: ASE(HOL)/ASE(QL_h) ranges from about 2 for $s = 100$ to about 11 for $s = 1000$. Once more, we do not see an improvement in the performance of HOL, as s increases: Figure 7 shows that $s\times$ ASE(HOL) increases roughly linearly as s increases. The slope of the $s\times$ ASE(HOL) curve is substantially greater than that of the $s\times$ ASE(QL_m) curve.

7.4. 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 as well. 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 8 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 8 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 8 shows that $s\times$ ASE(HOL) increases linearly (and steeply) as s increases.

7.5. 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 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, particularly when the number of servers is large. The HOL estimator remains the least effective estimator, under the MSE criterion, in all models considered.

8. Conclusions

In this paper, we studied the performance of alternative delay-history-based estimators, in the $M_t/GI/s$ and $M_t/GI/s + GI$ queueing models with a nonhomogeneous Poisson process. As a frame of reference, we considered two classical estimators based on the queue length seen upon arrival, QL and QL_m, previously studied in Ibrahim and Whitt (2009b). We proposed estimators that effectively cope with time-varying arrivals, and with non-exponential service-time and abandon-time distributions.

8.1. The HOL Estimator with Time-Varying Arrivals

We considered the HOL estimator, which is equal to the elapsed delay of the customer at the head of the line, at the time of arrival. The HOL estimator is appealing because it is robust: It does not depend on system parameters, and is easy to implement in practice.

When variability in the arrival process is slow over time, relative to the service times, HOL is an effective estimator, as shown previously in Ibrahim and Whitt (2009a, b). However, when the system alternates rapidly between phases of underload and overload, then HOL may perform poorly; see Figure 2. Motivated by the practical appeal of HOL, we developed refined estimators that use the HOL delay, and perform significantly better in systems with time-varying arrivals.

8.2. New HOL-Based Estimator for the $M_t/GI/s$ Model

In §3.1, we developed a new delay estimator for the $M_t/GI/s$ model. This estimator approximates the $M_t/GI/s$ model by the corresponding $M_t/M/s$ model, with the same service-time mean. We characterized the distribution of the conditional delay given the HOL observation, and proposed the mean of this delay as a refined-HOL-based delay estimator, HOL_r.

In §4, we obtained analytical results for the $M_t/M/s$ model. 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, particularly for high values of the traffic intensity, ρ; see (4.6). That is consistent with analytical results for the HOL estimator in the $GI/M/s$ model, established in Ibrahim and Whitt (2009a). Simulation shows that HOL performs consistently worse than HOL_r. The difference in performance is remarkable for high values of ρ. The HOL_r estimator is appealing because of its simple form, and its good performance.

8.3. New HOL-Based Estimator for the $M_t/GI/s + GI$ Model

We also developed a new delay estimator for the $M_t/GI/s+GI$ model. We used the observed HOL delay to estimate the queue length seen upon arrival in the system. We then used this queue-length estimate, together with established approximations in Whitt (2005), to develop a new estimator, QL_h.

As a frame of reference, we compared QL_h to the classical queue-length-based estimator, QL_m. The QL_m estimator approximates the service-time and abandonment-time distributions by corresponding exponential distributions. The QL_m estimator is provably the most effective estimator, under the MSE criterion, in the $M_t/M/s+M$ model. We showed, via sim-

ulation, that QL_h performs significantly better than both QL_m and HOL with time-varying arrivals, and non-exponential service-time and abandon-time distributions.

8.4. Managerial Insights

Our estimators exploit the history of delays in the system. They are therefore especially appealing in real-life systems where this information is easy to obtain. The HOL estimator is appealing in real-life systems where time-variability in the arrival process is negligible, so that it can be approximated by a stationary process.

The HOL_r estimator is effective in systems where customer abandonment is negligible, and the arrival process is highly time-varying, in which case the HOL estimator performs very poorly. Simulation shows that HOL_r performs well even when the service-time distribution is not nearly exponential, which is an important case to consider in practice.

The QL_h estimator is appealing in real-life systems where the arrival process is highly time-varying, and there is significant customer abandonment. Customer abandonment is an important phenomenon in practice, because it significantly impacts the performance measures of the system. The QL_h estimator is particularly effective in systems where the service and abandonment times are not nearly exponential, as often occurs in practice.

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.

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

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(1): 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.

Baccelli, Boyer, and Hebuterne. 1984. Single-sever queues with impatient customers. *Adv. Appl., Prob.* 16: 887–905.

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.

Eick, S., W.A. Massey, W. Whitt. 1993. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management. Sci.* 39(2): 241–252.

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

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

Green, L. and P. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.*, 37: 84-97.

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), vol. 16, No. 1, January-February 2007, pp. 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*, vol. 21, No. 1, pp. 143-156

Ibrahim, R. and W. Whitt. 2009a. Real-time delay estimation based on delay history. *Manufacturing and Service Oper. Mgmt.* Forthcoming. Articles in Advance available online.

Ibrahim, R. and W. Whitt. 2009b. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*. Forthcoming.

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 307318.

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.

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.

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.