Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

DELAY PREDICTORS FOR CUSTOMER SERVICE SYSTEMS WITH TIME-VARYING PARAMETERS

Rouba Ibrahim
Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University
New York, 10027, USA

ABSTRACT

Motivated by interest in making delay announcements in service systems, we develop new real-time delay predictors that effectively cope with customer abandonment and time-varying parameters. First, we focus on delay predictors exploiting recent customer delay history. We show that time-varying arrival rates can introduce significant prediction bias in delay-history-based predictors when the system experiences alternating periods of overload and underload. We then introduce a new delay-history-based predictor that effectively copes with time-varying arrival rates. Second, we consider a time-varying number of servers. We develop two new predictors which exploit an established deterministic fluid approximation for a many-server queueing model with time-varying demand and capacity. The new predictors effectively cope with those features, often observed in practice. Throughout, we use computer simulation to quantify the performance of the alternative delay predictors.

1 INTRODUCTION

We investigate alternative ways to predict, in real time, the delay (before entering service) of an arriving customer in a service system such as a hospital emergency department (ED) or a call center. We model such a service system by a queueing model with a time-varying arrival rate, a time-varying number of servers, and customer abandonment. In this paper, we present initial results of ongoing research extending Ibrahim and Whitt (2009a, b, c) where we investigated real-time delay prediction in conventional stationary queueing models both with and without customer abandonment. Our main contribution here is to propose new real-time delay predictors that effectively cope with the time variation often observed in practice; e.g., see Brown et al. (2005).

Our delay predictors may be used to make delay announcements. Delay announcements may be especially helpful when delays are sometimes long, as in a hospital ED. In many cases, waiting customers are unable to accurately predict 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 Alternative Delay Predictors

Alternative delay predictors differ in the type and amount of information that their implementation requires. In broad terms, we consider two families of delay predictors: (i) delay-history-based predictors, and (ii) queue-length-based predictors. Delay-history-based predictors exploit information about recent customer delay history in the system. Queue-length-based predictors exploit knowledge of the queue length (number of customers) seen upon arrival.

Delay-history-based predictors are appealing because they rely solely on information about recent customer delay history and thus need not assume knowledge of system parameters. Therefore, they are robust and respond automatically to changes in those parameters. A standard delay-history-based predictor 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. That is, $\theta_{HOL}(t,w) \equiv w$, where w is the elapsed delay of the HOL customer at the time of a new arrival, t. Delay-history-based predictors are natural candidates when the queue length in the system cannot be observed. For one 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. 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.

Queue-length-based predictors exploit system-state information including the queue length seen upon arrival. Additionally, they exploit information about various system parameters such as the arrival rate, the abandonment rate, and the number of servers. In general, queue-length-based predictors are more accurate than delay-history-based predictors because they exploit additional information about the state of the system at the time of prediction. A standard predictor is the simple queue-length-based predictor, QL_s, commonly used in practice, which multiplies the queue length plus one times the mean interval between successive service completions, ignoring customer abandonment. That is, a system having s agents each of whom, on average, completes one service request in m time units, may predict that a customer arriving at time t and finding n customers in queue upon arrival will be able to begin service in $\theta_{QL_s}(t,n) \equiv (n+1)m/s$ time units.

1.2 Queueing Models

We consider queueing models with customer abandonment and time-varying parameters. In the first part of this paper (Sections 3-5), we consider the $M(t)/M/s+GI$ queueing model which has a nonhomogeneous Poisson arrival process with an arrival-rate function $\lambda \equiv \{\lambda(t) : -\infty < t < \infty\}$. The number of servers, s, is fixed. Service times, S_n, are independent and identically distributed (i.i.d.) exponential random variables with mean $E[S] = \mu^{-1}$ (we omit the subscript when the specific index is not important). Abandonment times, T_n, are i.i.d. with a general distribution and mean $E[T] = \nu^{-1}$. The arrival, service, and abandonment processes are assumed to be independent. Customers are served according to the first-come-first-served (FCFS) service discipline. In the second part of this paper (Sections 6-9), we consider the $M(t)/M/s(t)+GI$ queueing model which has a time-varying number of servers. In particular, we assume that the number of servers varies according to the staffing function: $s \equiv \{s(t) : -\infty < t < \infty\}$.

1.3 Performance Measures

We quantify the accuracy of a delay predictor by the mean-squared error (MSE), which is defined as the expected value of the square of the difference between delay prediction and corresponding actual delay. It is usually difficult to determine the MSE of a delay predictor. Therefore, we rely in this paper on computer simulation to quantify the accuracy of the alternative predictors. In our simulation experiments, we quantify the accuracy of a delay predictor by computing the *average squared error* (ASE), defined by:

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

where p_i is the delay prediction for customer i, $a_i > 0$ is the potential waiting time of delayed customer i, and k is the number of customers in our sample. 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. We regard ASE as directly meaningful, but now we indicate how it relates to the MSE.

Let $W_{QL}(t,n)$ 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 number of customers seen in line at the time of his arrival, t, is equal to n. Let $\theta_{QL}(t,n)$ be some given single-number delay estimate which is based on n and t. Then, the MSE of the corresponding delay predictor is given by: $MSE(\theta_{QL}(t,n)) \equiv E[(W_{QL}(t,n) - \theta_{QL}(t,n))^2]$, which is a function of t and n. In order to get the overall MSE of the predictor at time t, we average with respect to the unconditional distribution of the number of customers $Q(t) = n$, seen in queue at time t, i.e.,

$MSE(t) \equiv E[MSE(\theta_{QL}(t,Q(t)))]$. Finally, to obtain an average "per-customer" perspective, we consider a weighted MSE (WMSE), defined by

$$WMSE \equiv \frac{\int_0^T \lambda(t) MSE(t) dt}{\int_0^T \lambda(t) dt} \ .$$

Our ASE is an estimate of the WMSE.

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

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

using the same notation as in (1). The RRASE is useful because it measures the effectiveness of a predictor relative to the average potential waiting time, given that the customer must wait.

1.4 Organization

The rest of this paper is organized as follows. In Section 2, we discuss previous research. In Section 3, we demonstrate potential problems with the HOL predictor with time-varying arrivals. In Section 4, we propose a new HOL-based predictor for the $M(t)/M/s+GI$ model, HOL_a, which effectively copes with time-varying arrivals. The HOL_a predictor is similar to a queue-length-predictor, QL_a, proposed in Ibrahim and Whitt (2009b). In Section 5, we describe simulation results for QL_a, HOL_a, and HOL in the $M(t)/M/s+GI$ model. In Section 6, we demonstrate potential problems with both QL_a and HOL_a with a time-varying number of servers. In Section 7, we review a deterministic fluid approximation for the $M(t)/M/s(t)+GI$ model developed in Liu and Whitt (2010). In Section 8, we propose new delay predictors based on those fluid approximations, which effectively cope with a time-varying number of servers. In Section 9, we describe simulation results for all predictors in the $M(t)/M/s(t)+GI$ model. In Section 10, we make concluding remarks.

2 PREVIOUS RESEARCH

In Ibrahim and Whitt (2009a), we studied the performance of QL_s and HOL in the $GI/M/s$ queueing model, which has a renewal arrival process and no abandonment. We showed that QL_s is the most effective predictor, under the MSE criterion, in the $GI/M/s$ model. We also showed that HOL is an effective predictor, and that the difference in performance between QL_s and HOL need not be too great. In particular, we showed that the ratio ASE(HOL)/ASE(QL_s) should be approximately equal to $(1+c_a^2)$, where c_a^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of the interarrival-time distribution. (This relation was shown to hold especially in large systems.) That is, HOL performs nearly the same as QL_s when the arrival process has low variability.

In Ibrahim and Whitt (2009b, c), we considered the $GI/GI/s+GI$ model, including customer abandonment. As one would expect, QL_s can overestimate customer delay when there is significant customer abandonment in the system. We showed that QL_s performs poorly in a heavily loaded $GI/GI/s+GI$ model, while HOL remains an effective estimator. For systems where customer abandonment is a serious issue, we proposed alternative queue-length-based predictors which effectively cope with non-exponential abandonment-time distributions. One such predictor is the approximation-based queue-length predictor, QL_a, which we describe in Section 4. The QL_a predictor assumes that the number of servers in the system is constant. In Section 6, we show that QL_a is not an effective predictor with a time-varying number of servers. Therefore, there is a need to propose new delay predictors which effectively cope with time-varying demand and capacity.

3 TIME-VARYING ARRIVAL RATES

In this section, we study the performance of the HOL predictor with time-varying arrival rates. When the delays vary rapidly over time, as can occur when there are alternating periods of significant overload and underload, we expect that the delay of a new arrival will not be like the HOL delay. To demonstrate potential problems with

the HOL predictor, we plot simulation sample paths of HOL predictions given, and actual delays observed, as a function of time, in a simulation run for a heavily-loaded $M(t)/M/100+M$ model.

In Figure 1, we consider the $M(t)/M/100+M$ model with sinusoidal arrival rates, a traffic intensity $\rho = 1.4$ (defined as the long-run average arrival rate divided by the total service rate), and mean service time $E[S] = 30$ minutes. We let the relative amplitude be $\alpha_a = 0.5$. (The ratio of the peak arrival rate to the average arrival rate is 1 + α_a.) We let the abandonment rate be $\nu = 0.1$ which corresponds to slow customer abandonment. With a small value of ν, waiting times are typically long. Delay prediction tends to be especially important with such long waiting times. We measure time and, thus, the delays in units of mean service times. Figure 1 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 predictions and the actual delays observed, leading to big errors. Figure 1 also shows a third plot of an new approximation-based HOL predictor, denoted by HOL_a, which we develop in Section 4. Clearly, it eliminates the time lag; visually the HOL_a plot falls nearly on top of the actual delays.

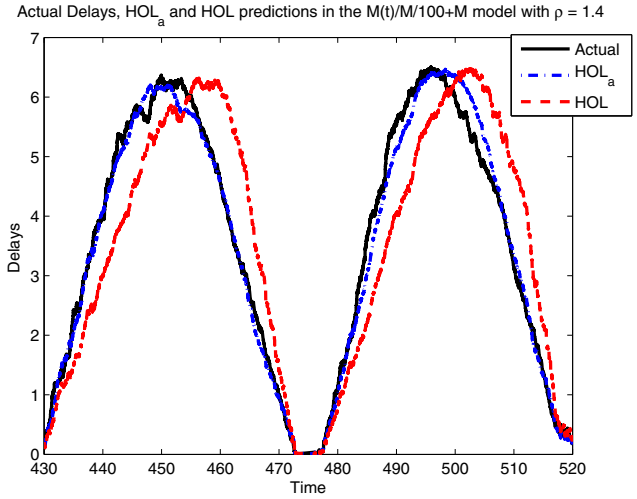

Figure 1: ASE of the alternative predictors in the $M(t)/M/100+M$ model for sinusoidal arrival rates and a relative amplitude $\alpha_a = 0.5$. We let $\gamma_a = 0.131$ which corresponds to $E[S]$ = 30 minutes with a 24 hour arrival-rate cycle. We let the abandonment rate $\nu = \mu/10 = 0.1$ which corresponds to slow abandonment.

4 A NEW HOL-BASED PREDICTOR WITH TIME-VARYING ARRIVALS

In this section, we propose a new HOL-based predictor for the $M(t)/M/s+GI$ model. We denote this new predictor by HOL_a because it is based on approximations for the $M/M/s+GI$ model developed in Whitt (2005). The HOL_a predictor is similar to a previous queue-length-based predictor (QL_a) for the $M/M/s+GI$ model, proposed in Ibrahim and Whitt (2009b). However, unlike QL_a, HOL_a exploits the HOL delay and does not assume knowledge of the queue length seen upon arrival. We begin with a description of QL_a.

We have the representation $W_{QL}(t,n) \equiv \sum_{i=0}^{n} Y_i$, where Y_{n-i} is the time between the ith and $(i+1)$st departure epochs.

For QL_a, we draw on the approximations in Whitt (2005). That is, we approximate the $M/M/s+GI$ model by the $M/M/s+M(n)$ model, with state-dependent Markovian abandonment rates. We begin by describing the Markovian approximation for abandonments, as in Section 3 of Whitt (2005). 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

$$\psi_j \equiv h(j/\lambda), \quad 1 \le j \le k \ ; \tag{3}$$

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

Here is how (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 $h(t)$. Therefore, we 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. With a stationary arrival process, a simple rough estimate for the time between successive arrival events is the reciprocal of the arrival rate, $1/\lambda$. Therefore, the elapsed waiting time of our customer is approximated by j/λ, and the corresponding abandonment rate by (3).

With time-varying arrival rates, we replace λ by $\hat{\lambda}$, where $\hat{\lambda}$ is defined as the average arrival rate over some recent time interval. For example, assuming that we know w, the elapsed delay of the customer at the HOL at the time of estimation, then we could define $\hat{\lambda}$ as the average arrival rate over the interval $[t-w,t]$, i.e., $\hat{\lambda} \equiv (1/w)\int_{t-w}^{t} \lambda(s)ds$.

For the $M(t)/M/s+M(n)$ model, we need to make further approximations in order to describe $W_{QL}(t,n)$: 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 $1/\hat{\lambda}$. 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 $1/\hat{\lambda}$. Then, Y_i has an exponential distribution with rate $s\mu+\delta_n-\delta_{n-i}$, where $\delta_k = \sum_{j=1}^{k} \psi_j = \sum_{j=1}^{k} h(j/\hat{\lambda})$, $k \geq 1$, and $\delta_0 \equiv 0$. That is the case because Y_i is 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 rates ψ_l, $n-i+1 \leq l \leq n$ (corresponding to the abandonment times of the customers waiting in line). The QL_a delay prediction given to a customer who finds n customers in queue upon arrival is

$$\theta_{QL_a}(n) = \sum_{i=0}^{n} \frac{1}{s\mu+\delta_n-\delta_{n-i}} \ ; \tag{4}$$

that is, $\theta_{QL_a}(n)$ approximates the mean of the potential waiting time, $E[W_{QL}(t,n)]$. With a time-varying number of servers, we replace s in (4) by $\bar{s}$, defined as the average number of servers in the system.

We now propose a new predictor, HOL_a, which is based on the HOL delay and does not assume knowledge of the queue length seen upon arrival. 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 predictor, paralleling (4). The HOL_a predictor is valuable only when the queue length is not known. This case is 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. An important point is that HOL_a is based on the HOL delay but is much more accurate than the direct HOL predictor with time-varying arrivals; see Section 5.

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 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

$$m(t,w) \equiv E[N_w(t)] = \int_{t-w}^{t} \lambda(s)(1-F(t-s))ds \ , \tag{5}$$

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. Paralleling (4), the HOL_a delay prediction given to a customer such that the observed HOL delay, at his time of arrival, t, is equal to w, is given by:

$$\theta_{HOL_a}(t,w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s\mu+\delta_n-\delta_{n-i}} \ , \tag{6}$$

for $m(t,w)$ in (5). If we actually know the queue length, then we can replace $m(t,w)$ by $Q(t)$, i.e., we can use QL_a. With a time-varying number of servers, we replace s in (6) by $\bar{s}$.

5 SIMULATION RESULTS WITH TIME-VARYING ARRIVALS

In this section, we present a sample of our simulation results for the $M(t)/M/s+GI$ model. Our methods apply to general time-varying arrival functions. To illustrate, we consider a sinusoidal function which is similar to what is observed with daily cycles. In Figures 2 and 3, we plot $s\times$ ASE (average number of servers times the ASE) for QL_a, HOL_a, and HOL in the $M(t)/M/s+M$ and $M(t)/M/s+E_{10}$ (Erlang abandonment, sum of 10 exponentials) models, respectively.

5.1 Description of the Experiments.

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

$$\lambda(t) = \bar{\lambda} + \bar{\lambda}\alpha_a \sin(\gamma_a t) \ , \tag{7}$$

where $\bar{\lambda}$ is the average arrival rate, α_a is the amplitude, and γ_a is the frequency. As pointed out by Eick et al. (1993), the parameters of $\lambda(t)$ in (7) should be interpreted relative to $E[S]$. Then, we speak of γ_a as the *relative* frequency. Small (large) values of γ_a 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 (24 hour) cycle. Here, we fix $\gamma_a = 1.57$ which corresponds to $E[S] = 6$ hours; see Table 1. For example, this value could be used to describe the experience of waiting patients in a hospital ED. We let $\alpha_a = 0.5$ and the traffic intensity $\rho = \bar{\lambda}/s\mu = 1.4$. We let the abandonment rate be $\nu = 1$ 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. We present additional simulation results in Ibrahim and Whitt (2010a).

5.2 Results for the $M(t)/M/s+M$ Model.

Figure 2 shows that QL_a is the most accurate predictor. In Ibrahim and Whitt (2009b), we provided theoretical support for this observation. The RRASE of QL_a ranges from about 14% for $s = 100$ to about 4% when $s = 1000$. From Section 4 of Ibrahim and Whitt (2009a) and Section 5 of Ibrahim and Whitt (2009b), we have theoretical results that provide useful perspective for the more complicated models we consider here. For example, we anticipate that the ASE of QL_a and HOL_a should be inversely proportional to the number of servers. Indeed, Figure 2 shows that $s\times\text{ASE}(\text{QL}_a)$ is nearly constant for all values of s considered. This shows that QL_a is asymptotically correct, i.e., $\text{ASE}(\text{QL}_a)$ approaches 0 as s increases.

As expected, HOL_a is the second most accurate predictor for this model. The RRASE of HOL_a ranges from about 20% for $s = 100$ to about 6% for $s = 1000$. The difference in performance between HOL_a and QL_a is not too great: $\text{ASE}(\text{HOL}_a)/\text{ASE}(\text{QL}_a)$ is close to 1.6, for all s. Moreover, Figure 2 shows that HOL_a is asymptotically correct: $s\times\text{ASE}(\text{HOL}_a)$ is roughly equal to a constant, for all s.

The HOL predictor performs much worse than both QL_a and HOL_a. For example, the ratio $\text{ASE}(\text{HOL})/\text{ASE}(\text{HOL}_a)$ 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 considerable improvement in the performance of HOL as s increases. That is confirmed by Figure 2, which shows that $s\times$ ASE(HOL) increases roughly linearly as s increases.

Similar conclusions hold with smaller service times, but the difference in performance between the alternative predictors is less extreme in that case. For example, with $E[S] = 5$ minutes and $s = 100$ (and all other parameters as above), the ratio $\text{ASE}(\text{HOL})/\text{ASE}(\text{HOL}_a)$ is roughly equal to 1.5, whereas it is roughly equal to 3 with $E[S] = 6$ hours.

5.3 Results for the $M(t)/M/s+GI$ Model.

In Figure 3, we consider E_{10} abandonment. For simulation results corresponding to other abandonment-time distributions, see Ibrahim and Whitt (2010a). Figure 3 shows that QL_a and HOL_a perform nearly the same for this model, so we will only discuss HOL_a. The RRASE of HOL_a ranges from about 11% for $s = 100$ to about 4% for $s = 1000$. That is, HOL_a is relatively accurate for this model. Figure 3 also shows that HOL_a is asymptotically correct.

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 $E[S] = 1$.

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

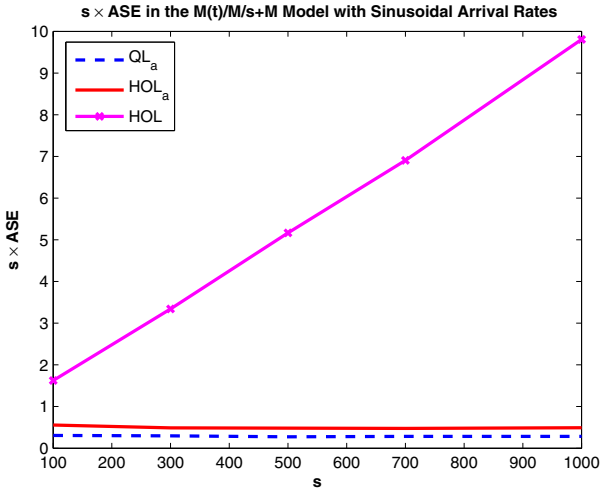

Figure 2: $E[S] = 6$ hours, $\alpha_a = 0.5$, $\rho = 1.4$

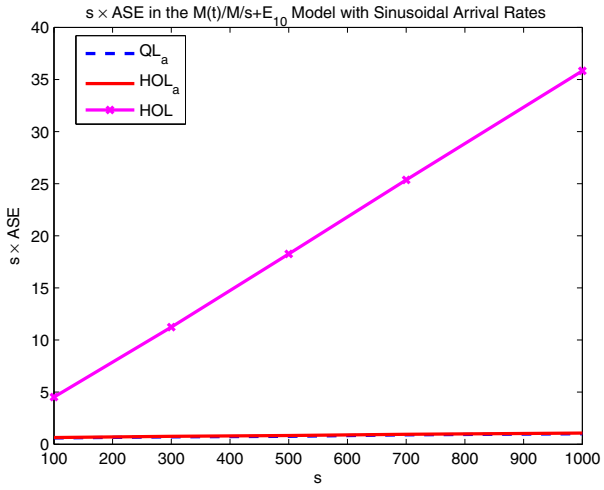

Figure 3: $E[S] = 6$ hours, $\alpha_a = 0.5$, $\rho = 1.4$

The least effective predictor is, yet again, the HOL predictor. The RRASE of HOL ranges from about 27% for $s = 100$ to about 25% for $s = 1000$. The difference in performance between HOL and HOL$_a$ is remarkable: ASE(HOL)/ASE(HOL$_a$) ranges from roughly 7 for $s = 100$ to roughly 33 for $s = 1000$. Figure 3 shows that $s\times$ ASE(HOL) increases roughly linearly (and steeply) as s increases.

6 TIME-VARYING NUMBER OF SERVERS

In this and the following sections, we consider a time-varying number of servers as well as a time-varying arrival rate. In Figure 4, we demonstrate potential problems with both HOL$_a$ and QL$_a$ in that setting. In particular, we consider the $M(t)/M/s(t)+M$ model with a sinusoidal arrival-rate intensity function, $\lambda(t)$, and a sinusoidal number of servers, $s(t)$, where there are periods of overloading leading to significant delays. We assume that $\lambda(t)$ and $s(t)$ have a period equal to 4 times the mean service time. With daily (24 hour) arrival-rate cycles, this assumption is equivalent to having a mean service time $E[S] = 6$ hours. We let the relative amplitude, α_a, for $\lambda(t)$ be equal to 0.5. We let the relative amplitude, α_s, for $s(t)$ be equal to 0.3; see Figure 4.

The HOL$_a$ and QL$_a$ predictors assume that the number of servers seen upon arrival is constant throughout the waiting time of the arriving customer, and equal to the average number of servers in the system. In the second (third) subplot of Figure 4, we plot simulation estimates of the average differences between HOL$_a$ (QL$_a$) delay predictions and actual delays observed in the system, as a function of time (dashed curves). These simulation

estimates are based on averaging 100 independent simulation replications. It is apparent that both HOL_a and QL_a are systematically biased in the $M(t)/M/s(t)+M$ model.

In Section 8, we propose a refined HOL-based predictor, HOL_r, and a refined queue-length-based predictor, QL_r. Both HOL_r and QL_r are based on a fluid approximation for the $M(t)/M/s(t)+GI$ queue, developed in Liu and Whitt (2010). Figure 4 nicely illustrates the improvement in performance resulting from our proposed refinements: We plot simulation estimates of the average differences between HOL_r (QL_r) delay predictions and actual delays observed in the system, as a function of time (solid curves).

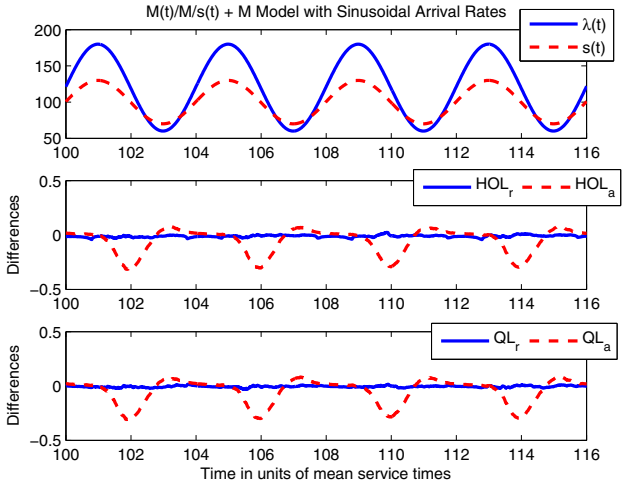

Figure 4: Bias of standard and refined delay predictors in the $M(t)/M/s(t)+M$ model with sinusoidal arrival rates. The differences between delay predictions and actual (potential) delays observed are based on averaging 100 independent simulation replications.

7 FLUID MODEL

In this section, we review fluid approximations for the $M(t)/M/s(t)+GI$ queueing model developed by Liu and Whitt (2010). Let $Q(t,x)$ denote the quantity of fluid in queue (but not in service), at time t, that has been in queue for time less than or equal to x time units. Similarly, let $B(t,x)$ denote the quantity of fluid in service, at time t, that has been in service for time less than or equal to x time units. We assume that functions Q and B are integrable with densities q and b, i.e.,

$$Q(t,x) = \int_0^x q(t,y)dy \quad \text{and} \quad B(t,x) = \int_0^x b(t,y)dy \ ,$$

where we define $q(t,x)$ ($b(t,x)$) as the rate at which quanta of fluid that has been in queue (service) for exactly x time units, is created at time t. Let $Q_f(t) \equiv Q(t,\infty)$ be the total fluid content in queue at time t, and let $B_f(t) \equiv B(t,\infty)$ be the total fluid content in service at time t. We require that $(B_f(t)-s(t))Q_f(t) = 0$ for all t, i.e., $Q_f(t)$ is positive only if all servers are busy at t. Under the FCFS service discipline, we can define a boundary waiting time at time t, $w(t)$, such that $q(t,x) = 0$ for all $x > w(t)$:

$$w(t) = \inf\{x > 0 : q(t,y) = 0 \text{ for all } y > x\} \ . \tag{8}$$

In other words, $w(t)$ is the waiting time experienced by quanta of fluid that enter service at time t (and have arrived to the system at time $t - w(t)$). We assume that the system alternates between intervals of overload ($Q_f(t) > 0, B_f(t) = s(t)$, and $w(t) > 0$) and underload ($Q_f(t) = 0, B_f(t) < s(t)$, and $w(t) = 0$). For simplicity, we assume that the system is initially empty.

Let $\bar{F}$ denote the complementary cumulative distribution function (ccdf) of the abandon-time distribution; i.e., $\bar{F}(x) = 1 - F(x)$. Let $\bar{G}$ denote the ccdf of the service-time distribution. The dynamics of the fluid model are defined

in terms of $(q, b, \bar{F}, \bar{G}, w)$ as follows:

$$q(t+u, x+u) = q(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)} \ , 0 \le x \le w(t) \ , \text{ and,} \tag{9}$$

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)} \ . \tag{10}$$

The queue length in the fluid model, at time t, is therefore given by

$$Q_f(t) = \int_0^{w(t)} q(t,y)dy = \int_0^{w(t)} \lambda(t-x)\bar{F}(x)dx \ , \tag{11}$$

where we use (9) to write $q(t,x) = q(t-x,0)\bar{F}(x) = \lambda(t-x)\bar{F}(x)$.

Let $v(t)$ denote the potential waiting time in the fluid model at time t. That is, $v(t)$ is the waiting time of infinitely patient quanta of fluid arriving to the system at t. Recalling that the waiting time of fluid entering service at t is equal to $w(t)$, it follows that this fluid must have arrived to the system $w(t)$ time units ago, and that

$$v(t - w(t)) = w(t) \ . \tag{12}$$

Therefore, for a given feasible boundary waiting time process, $\{w(t) : t \ge 0\}$, we can determine the associated potential waiting time process, $\{v(t) : t \ge 0\}$, using (12).

Liu and Whitt (2010) show that, under some regulatory conditions, if $Q_f(t) > 0$, then $w(t)$ must satisfy the following ordinary differential equation (ODE): $w'(t) = 1 - b(t,0)/q(t,w(t))$, for some initial boundary waiting time. With exponential service times, $b(t,0) = s(t)\mu + s'(t)$ whenever $Q_f(t) > 0$, where $s'(t)$ denotes the derivative of $s(t)$ with respect to t. Note that this implies the following *feasibility condition* on $s(t)$ when all servers are busy (i.e., during an overload phase): $s(t)\mu + s'(t) \ge 0$ for all t . This feasibility condition is possible because there is no randomness in the fluid model. For the stochastic system, there would always be some probability of infeasibility. Here, we assume that the staffing function is feasible.

Using (9), we can write that $q(t, w(t)) = \lambda(t - w(t))\bar{F}(w(t))$. As a result, with exponential service times,

$$w'(t) = 1 - \frac{s(t)\mu + s'(t)}{\lambda(t - w(t))\bar{F}(w(t))} \ . \tag{13}$$

Note that (13) is only valid for t such that $Q_f(t) > 0$ (i.e., during an overload phase). During underload phases, quanta of fluid is served immediately upon arrival, without having to wait in queue, i.e., $w(t) = 0$. Using the dynamics of the fluid model in (9) and (10), together with (13), we can determine $w(t)$ for all t, with exponential service times.

We now specify how to compute $w(t)$ by describing fluid dynamics in underload and overload phases. Assume that t_0 is the beginning of an underload phase, and let $B_f(t_0)$ be the fluid content in service at time t_0. (We assume that $Q_f(t_0) = 0$.) Let t_1 denote the first time epoch after t_0 at which $Q_f(t) > 0$. That, the system switches to an overload period at time t_1. For all $t \in [t_0, t_1]$, the fluid content in service is given by

$$B_f(t) = B_f(t_0)e^{-\mu(t-t_0)} + \int_{t_0}^{t} \lambda(t-x)e^{-\mu x}dx \ . \tag{14}$$

The first term in (14) is the remaining quantity of fluid, in service, that had already been in service at time t_0. The second term is the remaining fluid in service, at time t, that entered service in the interval $(t_0, t_1]$. We define t_1 as follows: $t_1 = \inf\{t > 0 : B_f(t) \ge s(t)\}$, for $B_f(t)$ in (14). Note that $w(t) = 0$ for all $t \in (t_0, t_1]$. Let t_2 denote the first time epoch after t_1 at which $Q_f(t) = 0$. That is, $[t_1, t_2]$ is an overload phase. For all $t \in (t_1, t_2]$, we compute $w(t)$ by solving (13). We define t_2 as follows: $t_2 = \inf\{t > t_1 : w(0) = 0\}$. At time t_2 a new underload period begins and we proceed as above to calculate $w(t)$. As such, we obtain $w(t)$ for all values of t. Using $w(t)$, we obtain $v(t)$ via (12), and $Q_f(t)$ via (11), for all t.

8 FLUID-BASED PREDICTORS: QL_r AND HOL_r

In this section, we use fluid approximations for $w(t)$, $v(t)$, and $Q_f(t)$ to develop new fluid-based delay predictors for the $M(t)/M/s(t)+GI$ model, which effectively cope with time-varying arrivals, a time-varying number of servers, and customer abandonment. Those new fluid-based predictors are consistent with previous ones, proposed in Ibrahim and Whitt (2009b), based on fluid approximations for the stationary $GI/GI/s+GI$ model.

8.1 The Refined-Queue-Length-Based (QL_r) Delay Predictor

We propose a simple refinement of QL_s, QL_r, which makes use of the fluid model in Section 7. Consider a customer who arrives to the system at time t, and who must wait before starting service. In the fluid approximation, the associated queue length, $Q_f(t)$, seen upon arrival at time t, is given by (11). As a result, $\text{QL}_{s,f}$ predicts the delay of a customer arriving to the system at time t, in the fluid model, as the deterministic quantity $\theta_{QL_{s,f}}(Q_f(t)) \equiv (Q_f(t)+1)/(s(t)\mu)$. The fluid approximation for the potential waiting time, $v(t)$, is given by (12). For QL_r, we propose computing the ratio

$$\beta(t) = v(t)/\theta_{QL_{s,f}}(Q_f(t)) = v(t)/((Q_f(t)+1)/s(t)\mu) = v(t)s(t)\mu/(Q_f(t)+1) \ , \tag{15}$$

and using it to refine the QL_s predictor. That is, the new delay prediction given to a customer arriving to the system at time t, and finding n customers in queue upon arrival, is the following function of n and t:

$$\theta_{QL_r}(t,n) \equiv \beta(t) \times \theta_{QL_s}(t,n) = v(t) \times \frac{n+1}{Q_f(t)+1} \ , \tag{16}$$

for $\beta(t)$ in (15). It is significant that θ_{QL_r} only depends on the number of servers, $s(t)$, through $v(t)$ and $Q_f(t)$. Indeed, the queue length is directly observable in the system, but the potential waiting time requires estimation, which is very difficult in the $M(t)/GI/s(t)+GI$ model. The advantage of using the fluid model is that it provides a way of approximating the potential waiting time.

8.2 The Refined HOL (HOL_r) Delay Predictor

The HOL delay prediction, $\theta_{HOL}(t,w)$, given to a new arrival at time t, is well approximated by the fluid boundary waiting time $w(t)$ in (8). The potential waiting time of that same arrival is approximately equal to $v(t)$ (which is the fluid approximation of the potential waiting time at t). Thus, we propose computing the ratio $v(t)/w(t)$ (after solving numerically for $v(t)$ and $w(t)$), and using it to refine the HOL predictor. Let HOL_r denote this refined HOL delay predictor. The delay prediction, as a function of w and the time of arrival t, is defined as

$$\theta_{HOL_r}(t,w) \equiv \frac{v(t)}{w(t)} \times \theta_{HOL}(t,w) = \frac{v(t)}{w(t)} \times w \ . \tag{17}$$

9 SIMULATION RESULTS WITH A TIME-VARYING NUMBER OF SERVERS

In this section, we describe a sample of our simulation results quantifying the performance of all candidate delay predictors in the $M(t)/M/s(t)+GI$ queueing model. We consider exponential abandonment times (i.e., the $M(t)/M/s(t)+M$ model) and vary the frequency of the arrival process (from slow variation to fast) while holding all other system parameters fixed. We consider alternative abandonment-time distributions in on-going work, Ibrahim and Whitt (2010b).

9.1 Description of the Experiments

We consider $\lambda(t)$ in (7). We consider a sinusoidal number of servers, $s(t)$. Specifically, we assume that

$$s(t) = \bar{s} + \bar{s}\alpha_s \sin(\gamma_s t) \ , \tag{18}$$

where $\bar{s}$ is the average number of servers. As in (7), γ_s is the frequency and α_s is the amplitude. We assume that $\gamma_a = \gamma_s$, and let $\alpha_a = 0.5$ and α_s = 0.3. That is, we assume that $\lambda(t)$ fluctuates more extremely than $s(t)$.

We study the performance of the candidate delay predictors in the $M(t)/M/s(t)+M$ model for alternative values of the arrival-process frequency, γ_a. In particular, we consider values of $\gamma_a = \gamma_s$ ranging from 0.022 ($E[S] = 5$ minutes with a 24 hour cycle) to 1.57 ($E[S] = 6$ hours with a 24 hour cycle); see Table 1. It is important to consider alternative values of $E[S]$ to show that our delay predictors are accurate in different practical settings. We let $\bar{s} = 100$ and $\rho = \bar{\lambda}/\bar{s}\mu = 1.2$.

Table 2 shows that QL_a and HOL_a are the least accurate predictors in this model, for all values of $E[S]$. In contrast, QL_r and HOL_r are much more accurate, especially for large $E[S]$. In general, however, all predictors are less accurate for large $E[S]$. The QL_r predictor is the most accurate predictor in this model, slightly outperforming HOL_r. Indeed, RRASE(QL_r) ranges from about 20% for $E[S] = 5$ minutes to about 25% for $E[S] = 6$ hours. The HOL_r predictor is the second most accurate predictor, and the difference in performance between HOL_r and QL_r depends on $E[S]$. Indeed, Table 2 shows that ASE(HOL_r)/ASE(QL_r) ranges from about 1.6 for $E[S] = 5$ minutes to about 1.3 for $E[S] = 6$ hours. That is, the difference in performance between HOL_r and QL_r decreases as $E[S]$ increases.

Table 2 shows that QL_a and HOL_a are not effective in this model, particularly for large $E[S]$. For example, RRASE(QL_a) ranges from about 26% for $E[S] = 5$ minutes to about 51% for $E[S] = 6$ hours. As expected, the HOL_a predictor performs slightly worse than QL_a: RRASE(HOL_a) ranges from about 30% for $E[S] = 5$ minutes to about 53% for $E[S] = 6$ hours. The ratio ASE(QL_a)/ASE(QL_r) ranges from about 1.8 for $E[S] = 5$ minutes to about 4 for $E[S] = 6$ hours. We show in Ibrahim and Whitt (2010b) that the difference in performance between QL_a and QL_r is even greater with a larger number of servers. That is not surprising since the fluid model is a remarkably accurate approximation of large systems.

Table 2: Performance of the alternative predictors, as a function of $E[S]$, in the $M(t)/M/s(t)+M$ model with $\lambda(t)$ in (7), $s(t)$ in (18), and $\bar{s} = 100$. Estimates of the ASE are shown together with the half width of the 95% confidence interval.

ASE of the predictors in the $M(t)/M/s(t)+M$ model as a function of $E[S]$				
$E[S]$	QL_r	HOL_r	QL_a	HOL_a
5 min.	2.82×10^{-3}	4.49×10^{-3}	5.05×10^{-3}	6.38×10^{-3}
	$\pm 2.5 \times 10^{-4}$	$\pm 4.4 \times 10^{-4}$	$\pm 2.1 \times 10^{-4}$	$\pm 2.1 \times 10^{-4}$
30 min.	2.71×10^{-3}	4.14×10^{-3}	4.54×10^{-3}	6.04×10^{-3}
	$\pm 8.1 \times 10^{-5}$	$\pm 1.2 \times 10^{-4}$	$\pm 3.5 \times 10^{-5}$	$\pm 6.6 \times 10^{-5}$
1 hr.	2.82×10^{-3}	4.44×10^{-3}	4.79×10^{-3}	6.33×10^{-3}
	$\pm 5.2 \times 10^{-5}$	$\pm 8.1 \times 10^{-5}$	$\pm 8.1 \times 10^{-5}$	$\pm 9.5 \times 10^{-5}$
2 hrs.	3.49×10^{-3}	5.38×10^{-3}	6.32×10^{-3}	8.04×10^{-3}
	$\pm 8.0 \times 10^{-5}$	1.2×10^{-4}	$\pm 1.6 \times 10^{-4}$	$\pm 2.0 \times 10^{-4}$
6 hrs.	7.25×10^{-3}	9.40×10^{-3}	2.99×10^{-2}	3.21×10^{-2}
	$\pm 2.2 \times 10^{-4}$	$\pm 2.1 \times 10^{-4}$	$\pm 4.6 \times 10^{-4}$	$\pm 5.6 \times 10^{-4}$

10 CONCLUSIONS

In this paper, we proposed alternative real-time delay predictors for nonstationary many-server queueing systems and showed that they are effective in the $M(t)/M/s+GI$ and $M(t)/M/s(t)+GI$ queueing models. Throughout, we used simulation to study the performance of the candidate delay predictors in several practical settings.

Figure 1 showed that existing delay-history-based-predictors, such as HOL, may be systematically biased in the $M(t)/M/s+GI$ model. Therefore, in Section 4, we proposed a new HOL-based predictor, HOL_a, which is similar to a previous predictor proposed in Ibrahim and Whitt (2009b), QL_a; see (6). In Section 5, we showed that HOL_a (and QL_a) is asymptotically correct in the $M(t)/M/s+GI$ model. Figures 2 and 3 showed that the difference in

performance between HOL and HOL_a can be remarkable, particularly when the number of servers is large. Since both QL_a and HOL_a assume that the number of servers in the system is constant, they may perform poorly with a time-varying number of servers; e.g., see Figure 4. Therefore, in Section 8, we exploited a fluid approximation for the $M(t)/M/s(t)+GI$ model developed in Liu and Whitt (2010) to obtain the new fluid-based delay predictors, QL_r and HOL_r; see (16) and (17). In Section 9, Table 2 showed that both QL_r and HOL_r are asymptotically correct in the $M(t)/M/s(t)+GI$ model, unlike QL_a and HOL_a. In particular, fluid-based predictors have superior performance in large systems especially when $E[S]$ is large.

ACKNOWLEDGMENTS

This research was supported by NSF Grant 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.

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.

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. Real-Time Delay Estimation in Call Centers. *Proceedings of the 2008 Winter Simulation Conference*. 1: 2876-2883.

Ibrahim, R. and W. Whitt. 2010a. Real-Time Delay Estimation Based on Delay History in Many-Server Service Systems with Time-Varying Arrivals. *Working Paper.* IEOR Department, Columbia University, New York. Available at: http://columbia.edu/~rei2101.

Ibrahim, R. and W. Whitt. 2010b. Wait-Time Predictors for Customer Service Systems with Time-Varying Demand and Capacity. *Working Paper.* IEOR Department, Columbia University, New York. Available at: http://columbia.edu/~rei2101.

Liu, Y. and W. Whitt. 2010. A Fluid Approximation for the $G_t/GI/s_t+GI$ Queue. *Working Paper.* IEOR Department, Columbia University, New York. Available at: http://columbia.edu/~ww2040.

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.

AUTHOR BIOGRAPHIES

ROUBA IBRAHIM is a doctoral student in the Department of Industrial Engineering and Operations Research at Columbia University. Her research interests lie in queueing theory, simulation modeling, and healthcare operations.

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