Operations Research
Vol. 59, No. 5, September–October 2011, pp. 1106–1118
ISSN 0030-364X | EISSN 1526-5463 | 11 | 5905 | 1106
http://dx.doi.org/10.1287/opre.1110.0974
© 2011 INFORMS

Wait-Time Predictors for Customer Service Systems with Time-Varying Demand and Capacity

Rouba Ibrahim
Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada,
rouba.ibrahim@mail.mcgill.ca

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027,
ww2040@columbia.edu

We develop new, improved real-time delay predictors for many-server service systems with a time-varying arrival rate, a time-varying number of servers, and customer abandonment. We develop four new predictors, two of which exploit an established deterministic fluid approximation for a many-server queueing model with those features. These delay predictors can be used to make delay announcements. We use computer simulation to show that the proposed predictors outperform previous predictors.

Subject classifications: delay prediction; delay announcements; simulation; time-varying arrival rates; time-varying number of servers; nonstationary queues.
Area of review: Stochastic Models.
History: Received March 2010; revisions received June 2010, September 2010; accepted November 2010.

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 customer contact 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. Our main contribution is to propose new real-time delay predictors that effectively cope with the time variation and abandonment, which are often observed in practice; e.g., see Brown et al. (2005).

Motivating Application. We envision our delay predictions being used to make delay announcements to arriving customers. Delay announcements can be especially helpful with emergency services, such as in a hospital ED. A recent study by Press Ganey (2009), an Indiana-based consulting company specializing in health-care services, found that the average patient waiting time in hospital EDs in the United States is about four hours. Because of those lengthy waits, some patients might opt to leave without being seen by a doctor. Press Ganey (2009) presents statistical evidence suggesting that updating patients on their status (e.g., via delay announcements) makes their waits in the ED more bearable and deters patients from abandoning the ED before treatment.

Delay announcements can also be helpful with other, less critical services. For example, they can be especially helpful when queues are invisible to customers, such as in call centers; see Aksin et al. (2007) for background on call centers. A recent study by Vocalabs (2010), a Minnesota-based consulting company specializing in customer-service surveys, found that customer dissatisfaction with lengthy waits in customer call centers remains a major concern for leading companies such as Apple, Dell, and HP. There is empirical evidence suggesting that making real-time delay announcements is an inexpensive way of increasing customer satisfaction with the service provided; e.g., see Taylor (1994), Hui and Tse (1996), and Munichor and Rafaeli (2007).

Customers typically respond to delay announcements, and their response alters system performance. As discussed by Armony et al. (2009), studying customer responses to delay announcements requires an equilibrium analysis. However, it is not clear whether an equilibrium exists or how to fully characterize it. There might even be multiple equilibria. Here, we do not directly consider customer response. We think of our delay predictions being based on model information obtained after equilibrium has been reached (with the announcements being used). We leave the important extension of directly treating customer response to future research.

Our main purpose in this research is to contribute to the development of a service science. In doing so, we have applications in mind (such as that of making delay announcements), but the present paper is not intended to be about a specific application. Instead, we are aiming for applicable research of broad value. We mention specific application settings as illustrative only. Our main contributions (detailed below) are to (i) propose new and effective predictors, and (ii) to systematically study the performance of those predictors.

Alternative Delay Predictors. Alternative delay predictors differ in the type and amount of information that their implementation requires. (Delay *predictors* might also be called delay *estimators*, as we have done in previous papers, but predictors seems more appropriate, because the predictor is trying to predict a future delay, not to estimate a model parameter.) 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 waiting 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. Delay-history-based predictors directly account for customer response because they depend on the history of delays in the system, which in turn is affected by customer response. 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, $\theta_{\mathrm{HOL}}(t, w) \equiv w$, where w is the elapsed delay of the HOL customer at the time of a new arrival, t.

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.

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; see (2). The mean delay, conditional on some state information, minimizes the MSE. Thus, the most accurate predictor, under the MSE criterion, is the unbiased predictor announcing the conditional mean. Unfortunately, it is usually difficult to determine the conditional mean exactly. We, therefore, rely on approximations. Here, we exploit deterministic fluid approximations for many-server queues with time-varying arrivals and a time-varying number of servers, drawing upon recent work by Liu and Whitt (2010). It is also difficult to determine the MSE of a delay predictor. Therefore, we rely throughout on computer simulation to quantify the accuracy of the alternative delay predictors.

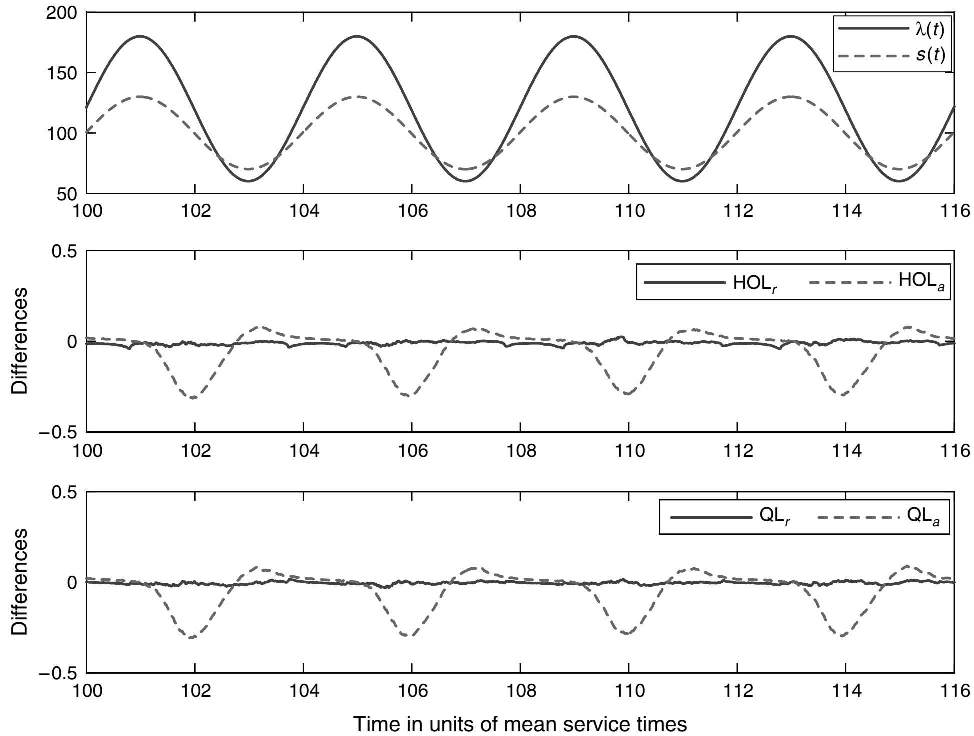
Previous Research. We begin by summarizing the main results of our previous related work. In Ibrahim and Whitt (2009a, b; 2011), we systematically studied the accuracy of various delay predictors in several many-server queueing models. The queueing models considered are controlled environments that mimic real-life customer-service systems. We started with the $GI/M/s$ model and extended to $GI/GI/s$ (nonexponential service times) and $GI/GI/s+GI$ (abandonment with nonexponential patience distributions). We showed that standard queue-length-based predictors, which are commonly used in practice, might perform poorly. We proposed new, more accurate queue-length-based predictors that effectively cope with nonexponential service and abandonment-time distributions, which are often observed in practice; see Brown et al. (2005).

Our most promising predictor, QL_a, draws on the approximations in Whitt (2005): it approximates the $GI/GI/s+GI$ model by the corresponding $GI/M/s+M(n)$ model, with state-dependent Markovian abandonment rates; see §3. Because QL_a assumes a stationary arrival process and a constant number of servers, it might perform poorly with time-varying arrivals and a time-varying number of servers, as we will show. Therefore, there is a need to go beyond QL_a.

We then considered the $M(t)/GI/s+GI$ model with time-varying arrival rates and a constant number of servers. We focused on the HOL delay predictor. We showed that HOL might perform poorly with time-varying arrival rates. When arrival rates vary significantly over time, customer delays may vary systematically as well, which leads to a systematically biased HOL predictor. We proposed refined delay-history-based predictors by analyzing the distribution of customer delay in the system, and we showed that those new predictors perform far better than HOL. Our most promising predictor is another approximation-based predictor, HOL_a. The HOL_a predictor is similar to QL_a; see §3. However, unlike QL_a, HOL_a exploits the HOL delay and does not assume knowledge of the queue length seen upon arrival. The HOL_a predictor has superior performance with a constant number of servers, but we will show that it, too, might perform poorly when the number of servers varies significantly over time. Therefore, there is a need to go beyond HOL_a.

Main Contributions. In this paper, we consider the $M(t)/M/s(t)+GI$ model, which we describe in §2. Because direct analysis of customer delay is complicated in this model, we propose two different approaches: (i) in §3, we propose modified versions of QL_a and HOL_a to account for a time-varying number of servers, and (ii) in §5, we exploit deterministic fluid approximations for many-server queues with time-varying arrivals and a time-varying number of servers, drawing upon recent work by Liu and Whitt (2010). (The fluid model has also been extended to general service and abandonment-time distributions with time-dependent parameters and to networks of queues. We leave such substantially more complicated scenarios to future work.) We propose new queue-length-based and delay-history-based predictors. Extensive

Figure 1. Bias of standard and refined delay predictors in the $M(t)/M/s(t)+M$ model with sinusoidal arrival rates (for model in §6.1).



Note. The differences between delay predictions and actual (potential) delays observed are based on averaging 100 independent simulation replications.

simulation results, of which we show a sample in §6 and in the electronic companion to this paper, available as part of the online version that can be found at http://or.journal.informs.org/, show that those new predictors have a superior performance in the $M(t)/M/s(t)+GI$ model.

In Figure 1, we demonstrate potential problems with HOL_a and QL_a. 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 that lead to significant delays. We assume that $\lambda(t)$ and $s(t)$ have a period equal to 4 times the mean service time; see §6.1. (Without loss of generality, we measure time in units of 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. (The ratio of the peak arrival rate to the average arrival rate is $1+\alpha_a$.) We let the relative amplitude, α_s, for $s(t)$ be equal to 0.3; see Figure 1.

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 is equal to the average number of servers in the system. (In practice, one might use an estimate of, say, the daily average number of servers.) In the second (third) subplot of Figure 1, 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.

Here, we propose a refined HOL-based predictor, HOL_r, and a refined queue-length-based predictor, QL_r. Figure 1 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).

Literature Review. The literature on delay announcements is large and growing. For a review of the growing literature on delay estimation and delay announcements, see Jouini et al. (2011, §2). In broad terms, there are three 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. (2009), Allon et al. (2011a, b), Jouini et al. (2011), and references therein. The second area studies alternative ways of estimating customer delay in service systems; e.g., see Whitt (1999a), Nakibly (2002), and Ibrahim and Whitt (2009a, b; 2011). The third area studies customer psychology in waiting situations; e.g., see Taylor (1994), Hui and Tse (1996), and Munichor and Rafaeli (2007). This paper falls in the second main area of research.

Organization of the Paper. The rest of this paper is organized as follows. In §2, we describe our general framework. In §3, we briefly describe the QL_a and HOL_a predictors, considered in §1, and propose modified predictors, QL_a^m and HOL_a^m, that cope with a time-varying number of servers. In §4, we review a deterministic fluid model, developed in Liu and Whitt (2010), for multi-server queues with time-varying arrival rates and customer abandonment. In §5, we use these fluid approximations to develop new, refined delay predictors. In §6, we present simulation results showing that these new predictors are effective in the $M(t)/M/s(t)+GI$ model. We make concluding remarks in §7. We present more simulation results (including general service-time distributions) in the electronic companion.

2. The Framework

In this section, we describe the $M(t)/M/s(t)+GI$ queueing model and then the performance measures that we use to quantify the performance of the alternative delay predictors.

2.1. The Queueing Model

We consider the $M(t)/M/s(t)+GI$ queueing model, which has a nonhomogeneous Poisson arrival process with an arrival-rate function $\lambda \equiv \{\lambda(u)\colon -\infty < u < \infty\}$. 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. The number of servers varies over time according to the staffing function: $s \equiv \{s(u)\colon -\infty < u < \infty\}$.

As is customarily done in the limited literature that considers the realistic feature of time-varying arrival rates, we assume that the arrival rate and the number of servers are deterministic functions of time, thus leaving out some form of randomness (because the arrival rate and the number of servers are usually not known with certainty in practice). In doing so, we capture what we believe to be the dominant effect of variability, which is the deterministic variation of the arrival rate and number of servers over time.

2.2. Performance Measures

Average Squared Error (*ASE*). In our simulation experiments, we quantify the accuracy of a delay predictor by computing the *average squared error* (ASE), defined by:

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

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.

In our simulation experiments, we measure a_i for both served and abandoning customers. For abandoning customers, we compute the delay experienced, had the customer not abandoned, by keeping him "virtually" in queue until he would have begun service. Such a customer does not affect the waiting time of any other customer in queue. As discussed in Ibrahim and Whitt (2009a, b; 2011), the ASE should approximate the expected MSE for a stationary system in steady state with a constant arrival rate, but the situation is more complicated with time-varying arrivals. We regard ASE as directly meaningful, but now we indicate how it relates to the MSE. Although the MSE (or the ASE) is not the only criterion that might be of interest, we consider it here to gain general insight into the performance of the predictors.

Weighted Mean-Squared Error (*WMSE*). Let $W_{\text{QL}}(t,n)$ 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 $\theta_{\text{QL}}(t,n)$ be some given single-number delay estimate that is based on n and t. Then, the MSE of the corresponding delay predictor is given by

$$\text{MSE}(\theta_{\text{QL}}(t,n)) \equiv E\big[(W_{\text{QL}}(t,n) - \theta_{\text{QL}}(t,n))^2\big], \tag{2}$$

which is a function of t and n. 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.,

$$\text{MSE}(t) \equiv E\big[\text{MSE}\big(\theta_{\text{QL}}(t, Q(t))\big)\big]. \tag{3}$$

Finally, to obtain an average "per-customer" perspective, we consider a weighted MSE (WMSE), defined by

$$\text{WMSE} \equiv \frac{\int_0^T \lambda(t)\text{MSE}(t)\,dt}{\int_0^T \lambda(t)\,dt}. \tag{4}$$

Our ASE is an estimate of the WMSE; for supporting theory, see Massey and Whitt (1994).

3. Modified Delay Predictors: QL_a^m and HOL_a^m

Figure 1 shows that QL_a and HOL_a might be systematically biased when the number of servers, $s(t)$, varies significantly over time. In this section, we propose modified predictors, QL_a^m and HOL_a^m, which account for a time-varying number of servers. For completeness, we begin by reviewing QL_a and HOL_a. Simulation results, described in §6, show that QL_a^m and HOL_a^m are more accurate than QL_a and HOL_a, particularly when the mean service time, $E[S]$, is small.

3.1. The QL_a and HOL_a Predictors

Let $W_{QL}(t, n)$ denote the potential waiting time of a new arrival at time t, such that the queue length at t, excluding the new arrival, is equal to n. We have the representation

$$W_{QL}(t, n) \equiv \sum_{i=0}^{n} Y_i, \tag{5}$$

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 Whitt (2005, §3). 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

$$\psi_j \equiv h(j/\lambda), \quad 1 \leqslant j \leqslant k; \tag{6}$$

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 $t \geqslant 0$, where f is the corresponding density function (assumed to exist).

Here is how (6) is derived. If we knew that a given customer had been waiting for time t, then the rate of abandonment for that customer, at that time, would be $h(t)$. We, therefore, need to estimate the elapsed waiting time of that customer, given the available state information. Assuming that abandonments are relatively rare compared to service completions, it is reasonable to act as if there have been j arrival events since our customer arrived. 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 (6).

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., $\hat{\lambda} \equiv (1/w)\int_{t-w}^{t} \lambda(s)\, ds$. Alternatively, if we do not have information about the recent history of delays in the system and know only the queue length n, then we could, for example, replace w by $\hat{w} \equiv (n+1)/s\mu$ and compute $\hat{\lambda} \equiv (1/\hat{w})\int_{t-\hat{w}}^{t} \lambda(s)\, ds$.

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 $\delta_k = \sum_{j=1}^{k} \psi_j = \sum_{j=1}^{k} h(j/\hat{\lambda})$, $k \geqslant 1$, 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, $n-i+1 \leqslant l \leqslant n$ (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

$$\theta_{QL_a}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}}; \tag{7}$$

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 (7) by $\bar{s}$, defined as the average number of servers in the system. In practice, we would use the daily average number of servers in the system instead of $\bar{s}$.

Unlike QL_a, HOL_a 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 (7).

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

$$m(t, w) \equiv E[N_w(t)] = \int_{t-w}^{t} \lambda(s)(1-F(t-s))\, ds, \tag{8}$$

where F is the abandonment-time cumulative distribution function (cdf).

For step (ii), we use $m(t, w)+1$ as an estimate of the queue length seen upon arrival, at time t. Paralleling (7), the HOL_a 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

$$\theta_{HOL_a}(t, w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s\mu + \delta_n - \delta_{n-i}}, \tag{9}$$

for $m(t, w)$ in (8). 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 (9) by $\bar{s}$.

3.2. Modified Predictors: QL_a^m and HOL_a^m

Now we propose modified predictors, QL_a^m and HOL_a^m, that effectively cope with a time-varying number of servers. In particular, we propose adjusting (7) as follows: we replace s by $s(t_i)$ where t_i denotes the estimated next departure epoch when there are i remaining customers in line ahead of the new arrival, and $t_{n+1} \equiv t$. Here is how we define the QL_a^m delay prediction:

$$\theta_{\mathrm{QL}_a^m}(t, n) = \sum_{i=0}^{n} \frac{1}{s(t_{i+1})\mu + \delta_n - \delta_{n-i}}, \tag{10}$$

where

$$t_i = t_{i+1} + \frac{1}{s(t_{i+1})\mu + \delta_n - \delta_{n-i}} \quad \text{for } 0 \leqslant i \leqslant n, \tag{11}$$

and $t_{n+1} = t$. For HOL_a^m, we proceed similarly. In particular, we use

$$\theta_{\mathrm{HOL}_a^m}(t, w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s(t_{i+1})\mu + \delta_n - \delta_{n-i}}, \tag{12}$$

where t_i is given by (11) and $t_{n+1} = t$.

It is important that QL_a^m and HOL_a^m reduce to QL_a and HOL_a, respectively, with a constant number of servers. Hence, the new predictors are consistent with prior ones, which were shown to be remarkably accurate in simpler models. In §5, we take a different approach and propose new delay predictors based on fluid approximations, which we now review.

4. The Fluid Model with Time-Varying Arrivals

In this section, we review fluid approximations for the $M(t)/M/s(t)+GI$ queueing model, developed by Liu and Whitt (2010). It is convenient to approximate queueing models with fluid models, because performance measures in fluid models are deterministic and mostly continuous in time, which greatly simplifies the analysis.

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 $x > w(t)$:

$$w(t) = \inf\{x > 0\colon q(t, y) = 0 \text{ for all } y > x\}. \tag{13}$$

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. We also assume that there is no fluid in queue at the beginning of every overload phase. For the more general case, accounting for nonzero initial queue content, see Liu and Whitt (2010, §5).

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:

$$q(t+u, x+u) = q(t, x)\frac{\bar{F}(x+u)}{\bar{F}(x)}, \quad 0 \leqslant x \leqslant w(t), \quad \text{and}, \tag{14}$$

$$b(t+u, x+u) = b(t, x)\frac{\bar{G}(x+u)}{\bar{G}(x)}. \tag{15}$$

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

$$Q_f(t) = \int_0^{w(t)} q(t, y)\, dy = \int_0^{w(t)} \lambda(t-x)\bar{F}(x)\, dx, \tag{16}$$

where we use (14) 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

$$v(t - w(t)) = w(t). \tag{17}$$

Therefore, for a given feasible boundary waiting time process $\{w(t)\colon t \geqslant 0\}$, we can determine the associated potential waiting time process $\{v(t)\colon t \geqslant 0\}$, using (17).

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):

$$w'(t) = 1 - \frac{b(t, 0)}{q(t, w(t))} \tag{18}$$

for some initial boundary waiting time; see Liu and Whitt (2010, Theorem 5.3). 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):

$$s(t)\mu + s'(t) \geqslant 0 \quad \text{for all } t. \tag{19}$$

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. To that end, Liu and Whitt (2010, §6.2) develop an algorithm to detect the time of first violation of this condition and construct the minimal feasible staffing function greater than the initial infeasible staffing function.

Using (14), we can write that $q(t, w(t)) = \lambda(t - w(t))\bar{F}(w(t))$. As a result, with exponential service times,

$$w'(t) = 1 - \frac{s(t)\mu + s'(t)}{\lambda(t - w(t))\bar{F}(w(t))}. \tag{20}$$

Note that (20) is valid only 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 (14) and (15) together with (20), 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 is, 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

$$B_f(t) = B_f(t_0)e^{-\mu(t-t_0)} + \int_{t_0}^{t} \lambda(x)e^{-\mu x}\,dx. \tag{21}$$

The first term in (21) 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: $t_1 = \inf\{t > 0\colon B_f(t) \geqslant s(t)\}$, for $B_f(t)$ in (21). 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 (20). We define t_2 as follows: $t_2 = \inf\{t > t_1\colon w(0) = 0\}$. At time t_2 a new underload period begins, and we use (20) to calculate $w(t)$. As such, we obtain $w(t)$ for all values of t. Using $w(t)$, we obtain $v(t)$ via (17) and $Q_f(t)$ via (16) for all t.

Liu and Whitt (2010) also treat the case of nonexponential service times. The analysis is much more complicated in that case, however. The main difficulty lies in determining the service content density, $b(t,x)$, which no longer depends solely on the number of servers, $s(t)$. Indeed, $b(t,x)$ is obtained, with general service times, by solving a complicated fixed point equation; see Liu and Whitt (2010; Theorem 5.1 and Equation (22)) in that paper.

Next, 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.

5. New Fluid-Based Delay Predictors for the $M(t)/M/s(t) + GI$ Model

In this section, we propose new delay predictors for the $M(t)/M/s(t) + GI$ model by making use of the approximating fluid model described in the previous section.

The No-Information-Fluid-Based (*NIF*) *Delay Predictor.* We first propose a simple delay predictor that does not require any information about the system beyond the model. A natural candidate no-information (NI) delay predictor is the mean potential waiting time in the system at time t. Because we do not have a convenient form for the mean, we use the fluid model of §4 to develop a simple approximation. Let the no-information-fluid-based (NIF) delay prediction given to a delayed customer joining the queue at time t_0 be

$$\theta_{\text{NIF}}(t_0) \equiv v(t_0), \tag{22}$$

where $v(t_0)$ is the fluid approximation for the potential waiting time, at t_0. To compute $v(t_0)$, we use (17) and proceed as described in §4. The NIF predictor is appealing because of its simplicity and its ease of implementation. It serves as a useful reference point, because any predictor exploiting additional real-time information about the system should do at least as well as NIF.

The Refined-Queue-Length-Based (QL_r) *Delay Predictor.* We now propose a predictor based on the queue length seen upon arrival to the system. Let QL_r refer to this refined-queue-length-based predictor. The derivation of QL_r is based on that of the simple queue-length-based predictor, QL_s, which was considered in Ibrahim and Whitt (2009b). For a system having $s(t)$ agents at time t, each of whom on average completes one service request in μ^{-1} time units, we may predict that a customer who finds n customers in queue upon arrival will be able to begin service in $(n+1)/s(t)\mu$ minutes. Let QL_s refer to this simple queue-length-based predictor, commonly used in practice. Let the predictor, as a function of n, be

$$\theta_{\text{QL}_s}(t, n) = \frac{n+1}{s(t)\mu}. \tag{23}$$

In Ibrahim and Whitt (2009b), we show that QL_s is the most effective predictor, under the MSE criterion, in the $GI/M/s$ model but that it is not an effective predictor when there is customer abandonment in the system.

Recognizing the simple form of the QL_s predictor in (23) and its lack of predictive power with customer abandonment, we propose a simple refinement of QL_s, QL_r, which makes use of the fluid model in §4. 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 (16). As a result, $QL_{s,f}$ predicts the delay of a customer arriving to the system at time t, in the fluid model, as the deterministic quantity

$$\theta_{QL_{s,f}}(Q_f(t)) = \frac{Q_f(t)+1}{s(t)\mu}.$$

The fluid approximation for the potential waiting time, $v(t)$, is given by (17). For QL_r, we propose computing the ratio

$$\begin{aligned}\beta(t) &= v(t)/((Q_f(t)+1)/s(t)\mu)\\ &= v(t)s(t)\mu/(Q_f(t)+1),\end{aligned} \tag{24}$$

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:

$$\theta_{QL_r}(t,n) \equiv \beta(t) \times \theta_{QL_s}(t,n) = v(t) \times \frac{n+1}{Q_f(t)+1} \tag{25}$$

for $\beta(t)$ in (24). It is significant that θ_{QL_r} depends only 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.

The Refined HOL (HOL_r) *Delay Predictor.* We now propose a refinement of the HOL delay predictor. The HOL delay estimate, $\theta_{HOL}(t,w)$, given to a new arrival at time t, such that the elapsed waiting time of the customer at the head-of-the-line is equal to w, is well approximated by the fluid boundary waiting time $w(t)$ in (13). 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

$$\theta_{HOL_r}(t,w) \equiv \frac{v(t)}{w(t)} \times \theta_{HOL}(t,w) = \frac{v(t)}{w(t)} \times w. \tag{26}$$

The QL_r and HOL_r predictors reduce to the $GI/GI/s+GI$ model, considered in Ibrahim and Whitt (2009b), so that we have "version consistency," as with QL_a^m and HOL_a^m.

6. Simulation Experiments for the $M(t)/M/s(t)+GI$ Model

In this section, we describe simulation results quantifying the performance of all candidate delay predictors in the $M(t)/M/s(t)+GI$ queueing model. Our methods apply to general time-varying functions. To illustrate, we consider sinusoidal functions that are similar to what is observed with daily cycles.

In this section, we consider exponential service and abandonment times (i.e., the $M(t)/M/s(t)+M$ model). We consider nonexponential service and abandonment-time distributions in the electronic companion. We first vary the number of servers (from tens to hundreds) while holding all other system parameters fixed; see Figures 2 and 3. We then vary the frequency of the arrival process (from slow variation to fast) while holding all other system parameters fixed; see Table 2.

6.1. Description of the Experiments

We consider a sinusoidal arrival-rate intensity function given by

$$\lambda(u) \equiv \bar\lambda + \bar\lambda\alpha_a \sin(\gamma_a u),\ -\infty < u < \infty, \tag{27}$$

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 $\lambda(u)$ in (27) should be interpreted relative to the mean service time, $E[S]$. Without loss of generality, we measure time in units of mean service time. 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. We could also choose shorter cycles. For example, assuming an 8-hour cycle (typical number of hours in a workday), $E[S]$ in Table 1 should be divided by 3 (e.g., for $\gamma_a = 0.131$, $E[S] = 10$ minutes).

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

$$s(t) = \bar s + \bar s\alpha_s \sin(\gamma_s t), \tag{28}$$

where $\bar s$ is the average number of servers. As in (27), γ_s is the frequency, and α_s is the amplitude.

Table 1. The relative frequency, γ, as a function of the mean service time, $E[S]$, for a daily (24-hour) cycle.

Relative frequency γ_a	Mean service time $E[S]$
0.0220	5 minutes
0.0436	10 minutes
0.131	30 minutes
0.262	1 hour
1.57	6 hours
3.14	12 hours

In this section, we let $\alpha_a = 0.5$ and $\alpha_s = 0.3$. That is, we assume that $\lambda(t)$ fluctuates more extremely than $s(t)$. (We do so because this case corresponds to longer waiting times in the system, which is when making delay predictions is especially important.) We let the abandonment rate, ν, be equal to 1. That is, the mean time to abandon is assumed to be equal to $E[S]$, which seems reasonable. We define the traffic intensity $\rho \equiv \bar{\lambda}/\bar{s}\mu$ and let $\rho = 1.2$.

We assume that $\gamma_a = \gamma_s$. It is important to emphasize that we do not seek, in this paper, to determine appropriate staffing levels in response to time-varying arrival rates. Indeed, the problem of setting appropriate staffing levels to achieve a time-stable performance (i.e., to stabilize the system's performance measures) is reasonably well understood; e.g., see Eick et al. (1993), Feldman et al. (2008), and references therein. In particular, proper staffing, when it can be done, will make $s(t)$ "out-of-phase" with $\lambda(t)$, i.e., $\gamma_a \neq \gamma_s$. We deliberately violate this restriction because we are interested here in a situation in which the waiting times tend to be large, which is when delay announcements are important. In that setting, (i) customers hmight experience significant delays that motivate the need for making delay announcements, and (ii) we can study the time-varying performance of the system (as opposed to a time-stable performance with appropriate staffing).

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

$$\text{RRASE} \equiv \frac{\sqrt{\text{ASE}}}{(1/k)\sum_{i=1}^{k} p_i}, \tag{29}$$

using the same notation as in (1). The denominator in (29) is the average potential waiting time of customers who must wait. The RRASE is useful because it measures the effectiveness of an predictor relative to the average potential waiting time, given that the customer must wait. Simulation results, which we discuss next, are based on 10 independent replications of length of a few months each (depending on the model), assuming a 24-hour cycle; for a more detailed description of our simulation experiments see §EC.2.

6.2. Simulation Results

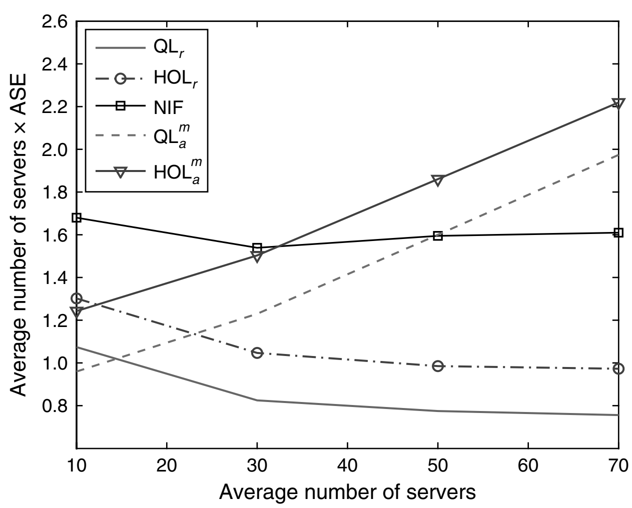
6.2.1. From Small to Large Systems. We study the performance of the candidate delay predictors in the $M(t)/M/s(t)+M$ model with $\gamma_a = \gamma_s = 1.57$. This relative frequency corresponds to $E[S] = 6$ hours with a 24-hour cycle and to $E[S] = 2$ hours with an 8-hour cycle; see Table 1. We consider this relatively large value of $E[S]$ to describe the experience of waiting patients in hospital EDs, where treatment times are typically long (hours or even days in some cases). We study the impact of changing $E[S]$ in §6.2.2. We study the performance of our predictors as a function of $\bar{s}$. In particular, we let $\bar{s}$ range from 10 to 1,000. Hence, our results are applicable to a wide range of real-life systems, ranging from small to very large. The difference between the upper and lower bounds of $s(t)$ in (28) is equal to $2\alpha_s\bar{s}$. Therefore, with $\alpha_s = 0.3$ (fixed), a large value of $\bar{s}$ corresponds to more extreme fluctuations in $s(t)$. For example, with $\bar{s} = 10$, $s(t)$ fluctuates between 7 and 13, whereas with $\bar{s} = 1{,}000$, $s(t)$ fluctuates between 700 and 1,300.

In this section, we present plots of $\bar{s} \times \text{ASE}$ (the average number of servers times the ASE) of the candidate predictors as a function of $\bar{s}$; see Figures 2 and 3. We do not show here separate results for QL_a and HOL_a. Indeed, those two delay predictors perform nearly the same as QL_a^m and HOL_a^m in this case (but not in all cases; see §6.2.2). We present corresponding tables with estimates (for all predictors) of the 95% confidence intervals in the electronic companion; see Table EC.3.

Overview of performance as a function of $\bar{s}$. From Ibrahim and Whitt (2009a, §4) and Ibrahim and Whitt (2009b, §5), we have theoretical results that provide useful perspective for the more complicated models we consider here. For example, we anticipate that the ASE should be inversely proportional to the number of servers and 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.) Similar relations are shown to hold here, too, provided that we use the refined, fluid-based predictors.

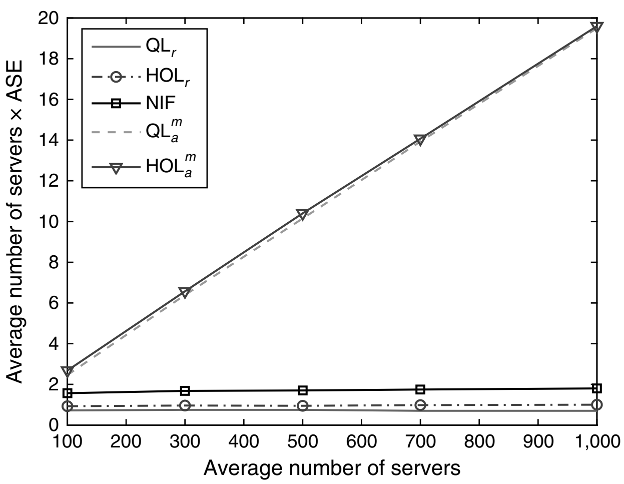
Figures 2 and 3 show that for fluid-based predictors, $\bar{s} \times \text{ASE}$ is roughly constant, particularly for large $\bar{s}$. This means that the ASE of fluid-based predictors is inversely

Figure 2. $\bar{s} \times \text{ASE}$ of the alternative predictors in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (27) and $s(t)$ in (28), and a small average number of servers, $\bar{s}$.



Note. We let $\gamma_a = \gamma_s = 1.57$, which corresponds to $E[S] = 6$ hours with a 24-hour arrival-rate cycle.

Figure 3. $\bar{s} \times$ ASE of the alternative predictors in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (27) and $s(t)$ in (28), and a large average number of servers, $\bar{s}$.



Note. We let $\gamma_a = \gamma_s = 1.57$, which corresponds to $E[S]$ = 6 hours with a 24-hour arrival-rate cycle.

proportional to $\bar{s}$ and thus converges to 0 in large systems. For example, ASE(QL_r) ranges from about 0.1 for $\bar{s} = 10$ to about 7×10^{-4} for $\bar{s} = 1{,}000$. That is, fluid-based predictors are *asymptotically correct*. Additionally, the ratio ASE(HOL_r)/ASE(QL_r) is roughly equal to a constant (equal to 1.3), particularly for large $\bar{s}$. Figures 2 and 3 also show that the ASE of other predictors (i.e., QL_a^m and HOL_a^m) are independent of $\bar{s}$. In particular, $\bar{s} \times$ ASE for those predictors is roughly linear as a function of $\bar{s}$. (That is especially true for large $\bar{s}$.) Consequently, the ASE of those predictors should roughly equal a (nonzero) constant for large systems. For example, Table EC.3 shows that ASE(QL_a^m) and ASE(HOL_a^m) are both roughly constant (equal to 0.02) for large $\bar{s}$.

Additionally, Figures 2 and 3 show that the ASEs of all delay predictors decrease as $\bar{s}$ increases. For example, the ASE of QL_r decreases by a factor of 150 in going from $\bar{s} = 10$ to $\bar{s} = 1{,}000$. (That is not surprising because the fluid model is a remarkably accurate approximation of large systems.) Moreover, the RRASEs of all predictors decrease as well. That is, all predictors are relatively more accurate in large systems. For example, the RRASE of QL_a^m decreases from roughly 64% for $\bar{s} = 10$ to roughly 46% for $\bar{s} = 1{,}000$. (Note that QL_a^m is not a very accurate predictor in this model, even when the number of servers is large.) Although all predictors perform better in large systems, the corresponding ASEs decrease at different rates. Indeed, Figures 2 and 3 clearly show the superiority of fluid-based predictors (i.e., QL_r, HOL_r, and NIF) for moderate to large values of $\bar{s}$, although all predictors perform nearly the same for very small $\bar{s}$ (e.g., $\bar{s} = 10$).

A closer look at the ASEs. For small values of $\bar{s}$, Figure 2 shows that there is no advantage in using fluid-based predictors over QL_a^m and HOL_a^m. Indeed, QL_a^m is the most accurate predictor for $\bar{s} < 15$. However, although QL_a^m is more accurate than fluid-based predictors for small systems, the difference in performance is not great. For one example, ASE(QL_a^m)/ASE(QL_r) is roughly equal to 0.9 for $\bar{s} = 10$. For another example, ASE(QL_a^m)/ASE(NIF) is roughly equal to 0.6 for $\bar{s} = 10$. Simulation experiments with an even smaller number of servers suggest that all predictors perform poorly when the number of servers is too small. For example, with $\bar{s} = 5$ (and all other parameters unchanged), the most accurate delay predictor is QL_a^m, but RRASE(QL_a^m) is roughly equal to 87%.

Figures 2 and 3 show that QL_r and HOL_r are more accurate than the rest of the predictors for $\bar{s} > 30$ (with QL_r being the most accurate predictor). For example, the RRASE of QL_r decreases from roughly 67% for $\bar{s} = 10$ to roughly 8% for $\bar{s} = 1{,}000$. The NIF predictor is competitive for $\bar{s} \geqslant 50$. Indeed, the RRASE of NIF ranges from about 84% for $\bar{s} = 10$ to about 12% for $\bar{s} = 1{,}000$. For large $\bar{s}$, QL_a^m and HOL_a^m perform nearly the same. For example, ASE(HOL_a^m)/ASE(QL_a^m) is roughly equal to 1 for $\bar{s} = 1{,}000$. They are both significantly outperformed by fluid-based predictors. Indeed, ASE(QL_a^m)/ASE(QL_r) ranges from about 0.9 for $\bar{s} = 10$ to about 27 for $\bar{s} = 1{,}000$. Also, ASE(QL_a^m)/ASE(NIF) ranges from about 0.6 for $\bar{s} = 10$ to about 11 for $\bar{s} = 1{,}000$.

Although NIF performs remarkably well in this model, other fluid-based predictors, which exploit some information about current system state, perform better particularly for large $\bar{s}$. For example, ASE(HOL_r)/ASE(NIF) ranges from about 1.5 for $\bar{s} = 10$ to about 2.5 for $\bar{s} = 1{,}000$. Also, ASE(QL_r)/ASE(NIF) ranges from about 1.3 for $\bar{s} = 10$ to about 1.8 for $\bar{s} = 1{,}000$. These ratios are even greater for smaller values of $E[S]$; see §6.2.2.

6.2.2. From Small to Large Frequencies. We now 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. In the following, we will measure $E[S]$ with respect to a 24-hour cycle. It is important to consider alternative values of $E[S]$ to show that our delay predictors are accurate in different practical settings. We let $\lambda(t)$ and $s(t)$ be as in (27) and (28), respectively, and let $\bar{s} = 100$. We leave all other parameters unchanged.

Overview of performance as a function of $E[S]$. With small $E[S]$, the system behaves at every time t like a stationary system with arrival rate $\lambda(t)$. Intuitively, for small $E[S]$, the number of both arrivals and departures during any given interval of time becomes so large that the system approaches steady-state behavior during that interval. Therefore, we expect that delay predictors that use $\lambda(t)$ and $s(t)$ corresponding to each point in time, such as QL_a^m and HOL_a^m (see (10) and (12)), will be accurate for small $E[S]$.

Table 2. Performance (ASE) of the alternative predictors as a function of $E[S]$ in the $M(t)/M/s(t)+M$ model with $\lambda(t)$ in (27), $s(t)$ in (28), and $\bar{s}=100$.

$E[S]$	QL_r	HOL_r	NIF	Ql_a^m	HOL_a^m	QL_a	HOL_a
5 min.	2.82×10^{-3} $\pm2.5\times10^{-4}$	4.49×10^{-3} $\pm4.4\times10^{-4}$	8.89×10^{-3} $\pm2.7\times10^{-4}$	2.20×10^{-3} $\pm1.9\times10^{-4}$	3.56×10^{-3} $\pm1.7\times10^{-4}$	5.05×10^{-3} $\pm2.1\times10^{-4}$	6.38×10^{-3} $\pm2.1\times10^{-4}$
30 min.	2.71×10^{-3} $\pm8.1\times10^{-5}$	4.14×10^{-3} $\pm1.2\times10^{-4}$	9.03×10^{-3} $\pm3.3\times10^{-4}$	2.06×10^{-3} $\pm4.2\times10^{-5}$	3.53×10^{-3} $\pm7.4\times10^{-5}$	4.54×10^{-3} $\pm3.5\times10^{-5}$	6.04×10^{-3} $\pm6.6\times10^{-5}$
1 hr.	2.82×10^{-3} $\pm5.2\times10^{-5}$	4.44×10^{-3} $\pm8.1\times10^{-5}$	9.49×10^{-3} $\pm3.0\times10^{-4}$	2.42×10^{-3} $\pm6.0\times10^{-5}$	4.00×10^{-3} $\pm8.6\times10^{-5}$	4.79×10^{-3} $\pm8.1\times10^{-5}$	6.33×10^{-3} $\pm9.5\times10^{-5}$
2 hrs.	3.49×10^{-3} $\pm8.0\times10^{-5}$	5.38×10^{-3} 1.2×10^{-4}	1.04×10^{-2} 3.4×10^{-4}	4.06×10^{-3} $\pm1.3\times10^{-4}$	5.85×10^{-3} $\pm2.0\times10^{-4}$	6.32×10^{-3} $\pm1.6\times10^{-4}$	8.04×10^{-3} $\pm2.0\times10^{-4}$
6 hrs.	7.25×10^{-3} $\pm2.2\times10^{-4}$	9.40×10^{-3} $\pm2.1\times10^{-4}$	1.57×10^{-2} $\pm5.6\times10^{-4}$	2.44×10^{-2} $\pm4.4\times10^{-4}$	2.66×10^{-2} $\pm5.5\times10^{-4}$	2.99×10^{-2} $\pm4.6\times10^{-4}$	3.21×10^{-2} $\pm5.6\times10^{-4}$

Note. Estimates of the ASE are shown together with the half width of the 95% confidence interval.

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, their modified versions, QL_a^m and HOL_a^m, are much more accurate, especially for small $E[S]$ as expected. For example, ASE(QL_a)/ASE(QL_a^m) is roughly equal to 2.3 for $E[S]=5$ minutes. Also, ASE(HOL_a)/ASE(HOL_a^m) is roughly equal to 1.8 for $E[S]=5$ minutes. This shows the need to go beyond existing delay predictors, such as QL_a and HOL_a. The difference in performance decreases as $E[S]$ increases, however. For example, ASE(QL_a)/ASE(QL_a^m) is roughly equal to 1.2, and ASE(HOL_a)/ASE(HOL_a^m) is roughly equal to 1.1, for $E[S]=6$ hours.

In general, all predictors are more accurate for small $E[S]$. For example, RRASE(HOL_r) ranges from about 25% for $E[S]=5$ minutes to about 29% for $E[S]=6$ hours. Also, RRASE(HOL_a^m) ranges from about 22% for $E[S]=5$ minutes to about 49% for $E[S]=6$ hours. Table 2 shows that although fluid-based predictors perform nearly the same as the remaining predictors for small $E[S]$ (e.g., 5 minutes), they perform much better for large $E[S]$ (e.g., 6 hours).

A closer look at the ASEs. The QL_a^m predictor is the most accurate predictor for small $E[S]$, slightly outperforming QL_r (which is the second most accurate predictor in that case). Indeed, Table 2 shows that ASE(QL_r)/ASE(QL_a^m) is roughly equal to 1.3 for $E[S]=5$ minutes. The HOL_a^m predictor is less accurate than QL_a^m, particularly for small $E[S]$. Indeed, ASE(HOL_a^m)/ASE(QL_a^m) ranges from about 1.6 for $E[S]=5$ minutes to about 1.1 for $E[S]=6$ hours. That is to be expected because QL_a^m exploits additional information about the queue length seen upon arrival, unlike HOL_a^m.

For $E[S]\geqslant 2$ hours, however, QL_r is more accurate than QL_a^m (and all remaining predictors); e.g., ASE(QL_r)/ASE(QL_a^m) is roughly equal to 0.85 for $E[S]=6$ hours. In larger systems, QL_r is more accurate than QL_a^m for even smaller $E[S]$. For example, with $\bar{s}=1{,}000$, ASE(QL_a^m) is slightly larger than ASE(QL_r) for $E[S]=30$ minutes.

The QL_a^m and HOL_a^m predictors both make systematic errors that cause their ASEs to increase dramatically with $E[S]$. They are, therefore, significantly less accurate than fluid-based predictors for large $E[S]$. For example, RRASE(QL_a) ranges from about 27% for $E[S]=5$ minutes to about 52% for $E[S]=6$ hours, whereas RRASE(QL_r) ranges from about 20% for $E[S]=5$ minutes to about 25% for $E[S]=6$ hours. Also, RRASE(HOL_a^m) ranges from about 22% for $E[S]=5$ minutes to about 49% for $E[S]=6$ hours, whereas RRASE(HOL_r) ranges from about 25% for $E[S]=5$ minutes to about 29% for $E[S]=6$ hours. Additionally, Table 2 shows that ASE(QL_a^m)/ASE(QL_r) ranges from roughly 0.8 for $E[S]=5$ minutes to roughly 3.4 for $E[S]=6$ hours, and ASE(HOL_a^m)/ASE(HOL_r) ranges from about 0.8 for $E[S]=5$ minutes to about 2.9 for $E[S]=6$ hours. Fluid-based perform even better with a larger number of servers; e.g., see §6.2.1.

Finally, we now compare the performance of NIF to that of other fluid-based predictors. Table 2 shows that NIF remains less accurate than QL_r and HOL_r. For example, ASE(NIF)/ASE(QL_r) ranges from about 3.1 for $E[S]=5$ minutes to about 2.1 for $E[S]=6$ hours. Also, ASE(HOL_r)/ASE(NIF) ranges from about 2 for $E[S]=5$ minutes to about 1.7 for $E[S]=6$ hours. The NIF predictor is the least accurate predictor for $E[S]\leqslant 2$ hours, yet it performs better as $E[S]$ increases. Indeed, it is more accurate than QL_a^m for large enough $E[S]$. For example, ASE(QL_a^m)/ASE(NIF) ranges from about 0.25 ($E[S]=5$ minutes) to about 1.6 ($E[S]=6$ hours).

6.2.3. Results for Nonexponential Distributions. In the electronic companion, we consider the $M(t)/M/s(t)+GI$ model with H_2 (hyperexponential with balanced means and SCV equal to 4) and E_{10} (Erlang, sum of 10 exponentials) abandonment-time distributions. Simulation results for those models are consistent with those described in this section. In particular, fluid-based predictors are more accurate than other predictors for long enough $E[S]$ and large enough $\bar{s}$, and the difference in performance can be remarkable. For example, in the $M(t)/M/s(t)+E_{10}$ model with $E[S]=6$ hours and $\bar{s}=1{,}000$, ASE(QL_a^m/ASE(QL_r) is roughly equal to 18.

We also study the performance of all delay predictors with both nonexponential service and abandonment-time distributions, i.e., we consider the $M(t)/GI/s(t) + GI$ model (we implement the alternative predictors by approximating the service-time distribution by an exponential with the same mean); see §EC.4. We consider H_2, E_{10}, and D (deterministic) service-time distributions. We find that the performance of the alternative predictors depends largely on the service-time distribution beyond its mean. With H_2 service times, fluid-based-predictors remain more accurate than QL_a^m and HOL_a^m. In Ibrahim and Whitt (2009b), we treated the case of deterministic service times and found that QL_a is not reliable in the $GI/D/s+GI$ model; e.g., see §6.4 of that paper. Nevertheless, QL_a remained effective with minimal variability in the service-time distribution, e.g., with E_{10} service times. Here, we find that fluid-based predictors are ineffective with both D and E_{10} service times. In contrast, we find that QL_a^m and HOL_a^m remain effective with deterministic (or nearly deterministic) service times and that they are considerably more accurate than fluid-based predictors in that case.

7. 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 $M(t)/M/s(t)+GI$ queueing model with time-varying arrival rates and a time-varying number of servers. In Table 3, we summarize the information needed for the implementation of each predictor and the characteristics of systems in which they might be preferred.

Figure 1 showed that existing delay predictors that do not take account of time-varying arrival rate and staffing, such as QL_a and HOL_a, can be systematically biased in the $M(t)/M/s(t)+GI$ model. Therefore, in §3 we proposed the modified predictors, QL_a^m and HOL_a^m. Then in §5 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, HOL_r, and NIF. All new delay predictors proposed in this paper reduce to prior ones that were shown to be remarkably accurate in simpler models. Throughout, we used simulation to study the performance of the candidate delay predictors in several practical settings. We considered alternative values of (i) the number of servers in the system, and (ii) the mean service time, $E[S]$.

Table 3. Summary of the information required for the implementation of each delay predictor.

Predictor	Information about the model	Preferred in systems with
QL_s	$Q(t), s(t), \mu$	Low abandonment, (nearly) exponential service times
QL_a, QL_a^m	$Q(t), s(t), \mu$, $F(x), \lambda(t)$	High abandonment, slow time-variability, small systems
HOL_a, HOL_a^m	$w, s(t), \mu$, $F(x), \lambda(t)$	High abandonment, slow time-variability, small systems
HOL_r	$w, s(t), \mu$, $F(x), \lambda(t)$	High abandonment, fast time-variability, large systems
QL_r	$Q(t), s(t), \mu$, $F(x), \lambda(t)$	High abandonment, fast time-variability, large systems

QL_r is consistently more accurate than both HOL_r and NIF. In terms of efficiency (low ASE), fluid-based predictors are ordered by $\mathrm{QL}_r < \mathrm{HOL}_r <$ NIF. Consistent with prior theoretical results in Ibrahim and Whitt (2009a, b), simulation showed that $\mathrm{ASE}(\mathrm{HOL}_r)/\mathrm{ASE}(\mathrm{QL}_r)$ is roughly equal to a constant between 1 and 2; e.g., see Figures 2 and 3. Although NIF is relatively accurate, particularly in large systems, it performs worse than both QL_r and HOLr because it does not exploit any information about the current system state at the time of prediction.

Fluid-based predictors outperform QL_a^m and HOL_a^m in large systems with large $E[S]$. Figure 3 showed that QL_r, HOL_r, and NIF are asymptotically correct in the $M(t)/M/s(t)+M$ model, with a large $E[S]$, unlike QL_a^m and HOL_a^m; i.e., the ASE of fluid-based predictors is inversely proportional to the number of servers. Moreover, Figure 2 showed that fluid-based predictors remain more accurate than QL_a^m and HOL_a^m even when the number of servers is not too large, provided that $E[S]$ is large enough (e.g., $\bar{s}=30$ and $E[S]=6$ hours).

QL_a^m and HOL_a^m outperform fluid-based predictors in small systems with small $E[S]$. Simulation showed that QL_a^m is the most accurate predictor for small $E[S]$, particularly when the number of servers is small (e.g., $E[S]=5$ minutes and $\bar{s}=10$). Table 2 showed that QL_a^m remains the most accurate predictor even when the system is relatively large (e.g., $E[S]=5$ minutes and $\bar{s}=100$). However, Table 2 also showed that the accuracy of QL_a^m and HOL_a^m decreases steadily as $E[S]$ increases. Indeed, both $\mathrm{RRASE}(\mathrm{QL}_a^m)$ and $\mathrm{RRASE}(\mathrm{HOL}_a^m)$ increase with increasing $E[S]$. Although fluid-based predictors perform worse for large $E[S]$ as well, their RRASEs increase much slower than $\mathrm{RRASE}(\mathrm{QL}_a^m)$ and $\mathrm{RRASE}(\mathrm{HOL}_a^m)$.

In some cases, there is not too much difference in performance between the delay predictors. Figure 2 showed that QL_a^m is only slightly more accurate than QL_r in small systems with large $E[S]$; e.g., $\bar{s}=10$ and $E[S]=6$ hours. The same conclusion also holds in large systems with small $E[S]$. For example, QL_a^m is also only slightly more accurate than QL_r for $\bar{s}=1{,}000$ and $E[S]=5$ minutes. In those cases, all delay predictors proposed are relatively accurate.

8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Acknowledgment

This research was supported by NSF grant CMMI 0948190.

References

Aksin, O. Z., M. Armony, V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Allon, G., A. Bassambo, I. Gurvich. 2011a. "We will be right with you": Managing customer expectations with vague promise and cheap talk. *Oper. Res.* Forthcoming.

Allon, G., A. Bassambo, I. Gurvich. 2011b. Delaying the delay announcements. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonments. *Oper. Res.* **57**(1) 66–81.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100**(469) 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.

Feldman, Z., A. Mandelbaum, W. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Sci.* **53**(6) 962–970.

Hui, M., D. Tse. 1996. What to tell customers in waits of different lengths: An integrative model of service evaluation. *J. Marketing* **60**(2) 81–90.

Ibrahim, R., W. Whitt. 2009a. Real-time delay estimation based on delay history. *Manufacturing Service Oper. Management* **11**(3) 397–415.

Ibrahim, R., W. Whitt. 2009b. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Sci.* **55**(10) 1729–1742.

Ibrahim, R., W. Whitt. 2011. Real-time delay estimation based on delay history in many-server queue with time-varying arrivals. *Production Oper. Management* **20**(5) 654–667.

Jouini, O., Z. Aksin, Y. Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing Service Oper. Management* **13**(4) 535–548.

Liu, Y., W. Whitt. 2010. The $G_t/GI/s_t + GI$ many-server fluid queue. Working paper, IEOR Department, Columbia University, New York. http://columbia.edu/~ww2040.

Massey, W., W. Whitt. 1994. A stochastic model to capture space and time dynamics in wireless communication systems. *Probab. Engrg. Informational Sci.* **8** 541–569.

Munichor, N., A. Rafaeli. 2007. Number of apologies? Customer reactions to tele-waiting time fillers. *J. Appl. Psych.* **92**(2) 511–518.

Nakibly, E. 2002. Predicting waiting times in telephone service systems. M.S. thesis, The Technion, Haifa, Israel.

Press Ganey Pulse Report. 2009. Emergency departments: Patient perspectives on American healthcare. Accessed February 1, 2010, http://www.pressganey.com/.

Taylor, S. 1994. Waiting for service: The relationship between delays and evaluations of service. *J. Marketing* **58**(2) 56–69.

Vocalabs National Customer Service Survey for Computer Tech Support. 2010. Accessed February 1, 2010, http://www.vocalabs.com.

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

Whitt, W. 1999b. Improving service by informing customers about anticipated delays. *Management Sci.* **45**(12) 192–207.

Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Sci.* **51**(2) 221–235.