Real-Time Delay Prediction in Customer Service Systems

Rouba Ibrahim

Submitted in partial fulfillment of the

Requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2010

ABSTRACT

Real-Time Delay Prediction in Customer Service Systems

Rouba Ibrahim

It is common practice in service systems to have customers who cannot be served immediately upon arrival wait in queue until system resources become available to the customer. A customer's waiting experience typically affects his evaluation of the service provided. For service providers, making delay announcements is a relatively inexpensive way of reducing customer uncertainty about delays, thereby improving customer satisfaction with the service provided. Our work focuses on applying queueing theory and computer simulation to develop effective ways to accurately predict customer delay in customer service systems, in real time. Primarily, these real-time delay predictions are intended to help service providers make delay announcements. But, they may also be used by service providers to better manage their systems. For instance, recognizing that customer delay is longer than planned at a service facility, the service provider may elect to provide additional service at that facility in order to reduce customer delay. Our general approach is to consider large heavily-loaded queueing models that mimic the operations of a real-life service system such as call center or a hospital emergency department. We are particularly concerned with the

practical appeal of our delay prediction procedures. That is why we incorporate important real-life phenomena such as customer abandonment, time-varying arrival rates, and a time-varying number of servers. We also consider general arrival, service, and abandonment-time distributions (exponential and non-exponential), which are commonly observed in practice. We use heavy traffic limits and computer simulation to quantify the accuracy of the alternative delay predictors proposed, and to compare them with delay predictors commonly used in practice.

Keywords: Real-time delay prediction; delay announcements, many-server queues; simulation; heavy-traffic; call centers; customer abandonment; time-varying arrival rates; nonstationary queues; time-varying number of servers;

Contents

List of Tables

List of Figures

Acknowledgments

It is a pleasure to thank those who made this thesis possible. First and foremost, I would like to thank my advisor, Prof. Ward Whitt, for his help and support throughout the years. In addition to being an absolutely brilliant scholar, Prof. Whitt is one of the most supportive and encouraging people that I know. I could not have hoped for a better advisor. I would like to thank the staff of the IEOR department: Donella, Jaya, Jenny, Maria, Risa, and Ufei who are always so helpful and friendly. I am grateful to my friends at the IEOR department who have truly made New York a home away from home. I love you all and will forever cherish the great times that we spent together. I look forward to many more great times with you in the future. In particular, Ceci, thanks for bringing so much happiness to me and the department in general.

I owe my deepest gratitude to my parents who tirelessly listened to my constant whining over the years. I would not be where I am now if it was not for them.

Last but not least, I thank my habibi, Yori, for bringing sunshine to my life and inspiring me to be a better person in so many ways. I dedicate this thesis to him.

1

Introduction

The service sector currently dominates the economic landscape of both emerging and developed economies. For example, the "US economy in brief" report (2008) shows that, as of 2008, the service sector in the United States contributes to nearly 70% of the Gross Domestic Product (GDP) and employs roughly 80% of the American workforce. In broad terms, the service sector comprises businesses (systems) that produce a service instead of just an end product; e.g., hospitals are service systems that provide health care services to patients.

Unlike tangible products, services are experienced and not consumed. To increase customer satisfaction, service systems compete in improving the quality of service provided. In general, quality of service is difficult to measure because it involves the perception of the customer being served and his mental state (as well as the provider's) during the service delivery. Nevertheless, some performance measures are commonly used to quantify the service level. One such measure is the delay experienced by customers in the system. Indeed, the time spent waiting for service is one of the most critical attributes of quality of service; e.g., see Maister (1985),

Larson (1987), Taylor (1994), and references therein.

Service providers are often faced with both time-varying and uncertain demand; e.g., see Mandelbaum et al. (2000), Jongbloed and Koole (2001), Avramidis et al. (2004), and Brown et al. (2005). That is especially challenging because service capacity (e.g., number of agents) cannot be inventoried. In an ideal service scenario, service providers have the ability to respond quickly to changing demand by adjusting staffing levels to meet unexpected demand in the short run; see Green et al. (2007), Feldman et al. (2008), and references therein. With appropriate staffing, customer wait times are short and quality of service can be high. We focus here on the less ideal case where service providers lack the resources or the flexibility to meet unexpected surges in demand, leading to long customer wait times. For example, the latter case is common in service-oriented (as opposed to revenue-generating) call centers; e.g., see Aksin et al. (2007).

There is empirical evidence suggesting that long waits lead to poor service evaluation, especially when coupled with feelings of uncertainty about the length of the wait; e.g., see Maister (1985) and Taylor (1994). Hiring and training new agents to alleviate the wait may be too costly. For example, Gans et al. (2003) indicate that "in most call centers capacity costs in general, and human resource costs in particular, account for 60% - 70% of operating expenses" (p.80). Studies show that improving customers' perceptions of the waiting experience can be as effective as reducing the actual length of the wait; see Katz et al. (1991). In particular, making delay announcements is a relatively inexpensive way of reducing customer uncertainty about delays, thereby improving customer satisfaction with the service provided; e.g., see Taylor (1994), Hui and Tse (1996), Munichor and Rafaeli (2007), and references therein. One important issue for service providers is the accuracy of those delay announcements.

1.1 Overview

In this work, we use queueing theory and computer simulation to develop accurate ways to predict customer delay in service systems, in real time. Primarily, these real-time delay predictions are intended to help service providers make delay announcements. But, they may also be used by service providers to better manage their systems. For instance, recognizing that customer delay is longer than planned at a service facility, the service provider may elect to provide additional service at that facility in order to reduce customer delay.

To fully understand a complex service system, we need to study it in detail. However, to help develop a service science, we systematically study various delay predictors in controlled environments, i.e., in structured models. Our general approach is to use queueing models that mimic the operations of real-life service systems. Indeed, we are particularly concerned with the practical appeal of our delay prediction procedures. That is why we incorporate important real-life phenomena such as customer abandonment, time-varying arrival rates, and a time-varying number of servers. We also consider non-exponential arrival, service, and abandonment-time distributions, which are commonly observed in practice; e.g., see Brown et al. (2005).

1.2 Motivating Application: Delay Announcements

We envision our delay predictions being used to make delay announcements to arriving customers. Since it helps to have a definite context in mind, we primarily focus here on two types of service systems: (i) hospital emergency departments (ED) and (ii) telephone call centers.

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 healthcare services, found that the average patient waiting time in hospital ED's in the United States is about four hours. Making real-time delay announcements is important with such long waits. Lengthy waits in hospital ED's are common, due to different factors including: (i) a lack of capacity, which translates into patients having to wait until hospital beds become available, and (ii) unpredictable surges in demand, such as those that emerge from disasters or local epidemics. Due to those lengthy waits, some patients may opt to "leave without being seen" (LWBS) by a doctor. Updating patients on their status (e.g., via delay announcements), would make their long waits in the ED more bearable, and could deter them 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 Gans et al. (2003) and Aksin et al. (2007) for background on call centers. Call center operations are typically regulated by service-level agreements (SLA) which specify target performance levels (such as wait-time level and proportion of abandoning customers). Nevertheless, in service-oriented call centers, such as those providing technical support services to incoming callers, customer wait times can sometimes be long, even when SLA performance levels are met on average. Indeed, 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. Making real-time delay announcements is one way of increasing customer satisfaction.

1.3 Complications

1.3.1 Delay Announcement Framework

We envision the following design: Each delayed customer, upon arrival, is given a single-number delay prediction of that customer's delay until he can start service.

A natural alternative design is to make several (updated) delay announcements, throughout the wait time of a delayed customer, thus creating a sense of progress. Munichor and Rafaeli (2007) show that delay announcements that "create a stronger sense of progress will produce more positive (customer) reactions" (p.512). However, making updated delay announcements is complicated because it involves keeping track of the state of the system at each delay announcement epoch. Additionally, there is a question of when to make the announcements. For example, Allon et al. (2010b) show that, under some conditions, it may be profitable for firms to postpone the delay announcement, i.e., to abstain from communicating information about anticipated delays immediately upon customer arrival.

Finally, there also remains to investigate other types of delay announcements, besides single-number predictions. For example, a service provider may choose to communicate a prediction interval or some upper bound of the wait time to delayed customers. Indeed, Guo and Zipkin (2007) show that accurate delay information may sometimes hurt both the provider and the customers. It may be even be more profitable for service providers to be intentionally vague about anticipated delays; e.g., see Allon et al. (2010a).

To gain insight into more complicated scenarios, it is natural to begin an investigation in a relatively tractable setting, for which we are able to obtain analytical results. Therefore, we leave the important extensions highlighted above to future

research.

1.3.2 Alternative Delay Predictors

Alternative delay predictors differ in the type and amount of information that their implementation requires. For example, this information may involve the model, the system state upon arrival, or the history of delays in the system. Analyzing the performance of the candidate predictors is especially difficult because it involves conditioning on this information. (The analysis becomes even more difficult if customer reaction to delay announcements is taken into account; see §1.3.3.) Computer simulation is thus essential to gain insight into the performance of alternative delay predictors.

Mathematical analysis and extensive simulation experiments show that there is not a single best delay predictor for all circumstances. Therefore, it is necessary to study the performance of multiple candidate predictors. Naturally, a good delay predictor is a delay predictor that is accurate. We use several performance measures to quantify the accuracy of a delay predictor. For example, we use the mean squared error (MSE) which we estimate via simulation by the average squared error (ASE). The MSE is defined as the expected value of the square of the difference between delay prediction and actual delay. Since the predictor typically depends on state information, we use the expected MSE, considering the steady-state distribution of the state information. The analysis becomes even more complicated with time-varying arrivals where the MSE is also a function of time; e.g., see §4.3. The mean delay, conditional on some state information, minimizes the expected 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. Therefore, we rely on approximations.

It is also important that the delay predictor be easy to implement in a real-life system, i.e., that it uses information that is readily available. An important insight, which applies broadly, is that simplicity and ease of implementation are often obtained at the expense of statistical accuracy.

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. For example, as in Armony et al. (2009), a standard delay-history-based predictor is the waiting time of the last customer to have entered service (LES) at the new arrival epoch. That is, $\theta_{LES}(t, w_L) \equiv w_L$, where w_L is the delay of the LES 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. For previous work on delay predictors exploiting system state information, see Whitt (1999a).

1.3.3 Customer Reactions

Customers typically respond to delay announcements, and their response alters system performance. For example, some customers may elect to balk, upon arrival, in response to a delay announcement. As a result, the arrival rate to the system would become state dependent. Moreover, customers who decide to stay may have different abandonment behavior in response to the announcement. They may become increasingly impatient if they have to wait more than their announced delay. As a result, the abandonment distribution of customers in queue would depend on their elapsed waiting time. Changes in system performance alter, in turn, the delay predictions given. 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 may 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). More generally, we regard our work as an essential first step toward studying the performance impact of delay announcements in the queueing models considered. It is not hard to see how the delay prediction methods of this work can be applied to the more complicated setting involving customer response. Indeed, 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. Delay-history-based predictors are appealing precisely because they directly apply to models involving customer response.

The queue-length-based predictors can also be extended to account for changes in customer behavior. For example, we could use an iterative simulation-based

algorithm to develop approximations of the equilibrium performance of the queueing model with delay announcements. During each iteration, we would give real-time delay predictions to arriving customers, and model their response. We would then re-estimate model parameters that are affected by customer response, and feed these new estimates into the subsequent iteration. The algorithm would continue until the observed difference between successive estimates of model parameters is negligible. It is significant that our proposed queue-length-based predictors apply directly to the successive iterations of this algorithm, using the different set of model parameters from each iteration. However, there remains to determine appropriate regularity conditions under which this algorithm terminates, i.e., under which there exists a unique equilibrium in the system. We leave such important extensions to future research.

1.4 Queueing Models

In this work, we study ways of predicting customer delay in several queueing models. Here, we briefly describe those models. A more detailed description is relegated to subsequent chapters.

In chapter 2, we consider the simple idealized setting of the $GI/M/s$ queueing model. This model has independent and identically distributed (i.i.d) exponential service times and s homogeneous servers. The interarrival times are i.i.d with a general distribution. Customers are served in order of arrival, i.e., we consider the first-come-first-served (FCFS) service discipline. The $GI/M/s$ model has the advantage of mathematical tractability. In the $GI/M/s$ model, we are able to obtain analytical results for several delay predictors.

As in Garnett et al. (2002), customer abandonment is an important phenomenon in real-life service systems. For example, the Help Desk Institute (2009) indicates, in its annual report, that about 40% of call centers observe an abandonment rate of over 10%. Moreover, non-exponential abandonment-time distributions are often observed in practice; e.g., see Brown et al. (2005). In chapter 3, we consider the $GI/GI/s + GI$ model which includes independent sequences of i.i.d. service and abandonment times with general distributions. The $GI/GI/s + GI$ model is difficult to analyze directly, so we rely on approximations in Whitt (2005b, 2006) to develop new delay predictors which effectively cope with customer abandonment.

The $GI/M/s$ and $GI/GI/s + GI$ models both assume a stationary arrival process. However, arrival processes to service systems, in real life, typically vary significantly over time; e.g., see Avramidis et al. (2004), Brown et al. (2005), and Shen and Huang (2008a, b). Therefore, in chapter 4, we consider the $M(t)/GI/s$ and $M(t)/GI/s + GI$ queueing models with a nonhomogeneous Poisson arrival process. Finally, since service providers typically adjust their staffing level in response to time-varying demand we consider, in chapter 5, the $M(t)/GI/s(t) + GI$ model with time-varying arrivals and a time-varying number of servers. To develop new delay predictors that effectively cope with time-varying demand and capacity, we rely on deterministic fluid approximations in Liu and Whitt (2010).

Real-life service systems are often much more complicated than structured queueing models. For one example, there may be multiple customer classes and multiple service pools with some form of skill-based routing; see Gans et al. (2003). For a second example, as with Web chat, servers may serve several customers simultaneously, different servers may participate in a single service, and there may be interruptions in the service times. For a third example, the arrival rate in a real-life system is often not known with certainty (as is assumed with a nonhomogeneous

Poisson arrival process). Therefore, it could be assumed to be a random variable; e.g., see Jongbloed and Koole (2001). However, such generalizations greatly complicate the analysis and are left to future research. The results of this work provide useful background for similar studies in even more complicated settings.

1.5 Literature Review

In broad terms, there are two main areas of research intimately related to our work: The first area studies the effect of delay announcements on system dynamics, and the second area studies alternative ways of predicting customer delay in service systems.

1.5.1 Effect of Delay Announcements

We begin by reviewing related literature from the first area. In general, this body of literature considers the issue of "optimal" wait-time quotes (by assuming appropriate cost structures) and studies the benefits of both overestimating and underestimating anticipated delays. That is different from our work where we focus solely on *accurately* predicting anticipated delays. Additionally, this stream of papers explicitly models customer reactions to the delay announcements made. In contrast, we do not incorporate customer reactions into our models, and leave this question to future research.

One of the earliest papers about customers influenced by delay information is Naor (1969). In that paper, identical risk-neutral customers decide, based on the observed queue length, whether or not to balk by comparing the expected cost of waiting with the reward from being served. In that context, Naor showed that imposing

appropriate tolls (upon entry) to arriving customers may regulate the system and lead to social optimality. Hassin (1986) adopted the same model as Naor (1969), and determined conditions under which a revenue-maximizing server should reveal (induce balking) or suppress (prevent balking) information about the queue length in the system to arriving customers.

Some researchers studied the problem of quoting manufacturing lead times in systems. For example, Duenyas and Hopp (1995) investigated the problem of quoting optimal lead times from the point of view of a revenue-maximizing manufacturer. First, they considered an infinite-capacity system and calculated the profit-maximizing lead time quote. (On one hand, a high lead time quote induces many customers to balk, and on the other hand, the manufacturer incurs a penalty if the order is not filled on time.) Then, they considered the case where capacity is finite and derived profit-maximizing lead time quotes under two scenarios: (i) where the lead time is dictated by the market, and (ii) where firms are able to compete on the basis of lead time. They found that the optimal lead time quote policy is state-dependent and increasing in the state.

Spearman and Zhang (1999) considered two different problems. The first problem seeks to minimize the average lead time quote of jobs subject to a constraint on the fraction of tardy jobs. The second problem uses the same objective subject to a constraint on average tardiness. The optimal lead time quote for the first problem is inconsistent with "ethical practice". In particular, the authors showed that it may sometimes be optimal for a firm to quote lead times that it has no hope of achieving. They attributed this conclusion to the inadequacy of using the fraction of tardy jobs as a service measure. On the other hand, the lead time quote policy which solves the second problem is more reasonable: Lead-time quotes are monotonically increasing with the level of congestion in the system.

Dobson and Pinker (2005) compared two scenarios where the firm shares either state-dependent or steady-state lead time information. Dobson and Pinker assumed that the firm cannot choose a lead time quote arbitrarily and must provide accurate quotes that are achieved (in a probabilistic sense) at least a given fraction of the time. Their focus on accuracy is similar to ours in this thesis. Dobson and Pinker showed that, in many cases, providing state-dependent lead time information is better than information based on the long-run lead time distribution, for both the firm and the customers. (State-dependent information increased throughput for the firm and decreased expected waiting time for the customers.) One main conclusion reached is that the possible benefits of sharing more information with customers are highly sensitive to modeling assumption (in their case, to the nature of a customer's sensitivity to waiting).

Guo and Zipkin (2007) considered a single server Markovian queue with balking and compared the effects of three different levels of delay information on system dynamics: (i) no information (steady-state distribution of wait-time), (ii) partial information (number of customers in the system), and (iii) full information (exact delay information). For (iii), Guo and Zipkin assumed that customers bring the realization of their service times upon arrival. They derived sufficient conditions under which more information helps the service provider (by increasing throughput) or the customers (by increasing the average utility). They found that, in some cases, more information can actually hurt one or the other.

Allon et al. (2010a) studied a retail operations problem where customers are strategic in both their actions and the way that they interpreted information communicated to them by the firm. In a dynamic game framework, they discussed the equilibrium language emerging between the retailer and its customers. When there is a single retailer, they find that equilibrium language only emerges when no information is

revealed. When there is more than one retailer, they found that firms are capable of credibly sharing unverifiable information. Interestingly, they found that firms are better off under an equilibrium in which intentional vagueness is used. Allon et al. (2010b) studied a related problem. They addressed the issue of delaying the announcements and showed that doing so may actually create credibility for the firm and augment the equilibrium language. They also showed that credibility, whenever created, improves not only the profit for the firm but also customer utility overall.

Other research has focused on studying delay announcements in call centers, as we do in this thesis. Whitt (1999b) studied the effect of communicating information about anticipated delays on system dynamics. In particular, he compared two Markovian models. In Model 1, no delay announcements are made, and customers may either balk upon arrival or join the queue and abandon if their wait exceeds their patience. In Model 2, information about current system state is communicated to arriving customers who, as a result, either balk upon arrival or remain in queue (customer abandonment is entirely replaced by balking). Whitt showed that the number of customers in Model 1 is stochastically larger than in Model 2. Intuitively, more customers balk immediately upon arrival in Model 2, thus alleviating the congestion in the system. Whitt's paper provides theoretical support to using delay announcements as a control mechanism by the service provider.

Armony and Maglaras (2004) considered a Markovian model with two different modes of service: real-time and postponed with a delay guarantee. They, too, focused specifically on call centers. In their model, customers are informed about their anticipated wait upon arrival and are offered a call-back option whereby the system will call them back within a specified amount of time. The authors proposed a delay prediction scheme and showed that it is asymptotically correct in the Quality and Efficiency Driven (Halfin-Whitt) regime. They also proposed a routing policy that

asymptotically minimizes real-time delay subject to the deadline of the postponed service mode. Consistent with our approach in this thesis, Armony and Maglaras focused on quoting wait times as accurately as possible.

Our work is partly motivated by Armony et al. (2009). The authors of that paper studied delay announcements in many-server queues with customer abandonment, focusing on customer response to the announcements, leading to balking and new abandonment behavior. They developed ways to approximately describe the equilibrium system performance using LES delay announcements (considered in this work as well). More specifically, they use deterministic fluid approximations in Whitt (2006) to derive conditions under which a unique equilibrium exists for the system with announcements. Armony et al. (2009) discussed the motivation for the LES delay predictor and other delays predictors based on recent delay history. Here, we consider a wider range of delay-history-based predictors.

Jouini et al. (2010) extended the model in Whitt (1999b). In particular, they assumed that customers who are given a delay announcement either balk immediately upon arrival or abandon, after joining the queue, if they end up waiting more than the delay announcement. Jouini et al. characterized, analytically, the performance measures for this model and used a numerical study to explore when informing customers about delays is beneficial, and what the optimal precision announcement is.

1.5.2 Predicting Customer Delay

The second body of literature focuses on accurately predicting customer waiting times in service systems, in real time, where the predictions could be used to make delay announcements. We note in passing that there is a stream of papers concen-

trated on correctly predicting lead times in the manufacturing setting, e.g., Morton and Vepsalainen (1987), Ornek and Collier (1988), and Shanthikumar and Sumita (1988), but their context is different from ours.

In addition to Armony et al. (2009), our work is also partly motivated by Whitt (1999a). In that paper, Whitt investigated delay predictions based on the state of the system in multiserver queues without customer abandonment. He showed how additional information about system state (e.g., number of customers in the system and elapsed service times) leads to better predictions. We reach similar conclusions in this work as well. We also use one predictor, QL, which was proposed in Whitt (1999a); see chapter 2. However, we consider here more complicated queueing models incorporating additional features such as time-varying arrival rates.

Nakibly (2002) focuses mostly on queueing models with priority and studies ways to predict waiting times based on information about system state upon arrival. She uses difference equations to predict waiting times in a model with two servers and two service types. She then uses matrix geometric methods to predict waiting times in more general systems with priorities (but without customer abandonment).

This thesis mainly contributes to the second area of research. It comprises edited versions of four papers, Ibrahim and Whitt (2009a,b, 2010a,b), where we study the accuracy of alternative delay predictors in queueing models with several realistic features.

1.6 Main Contributions

Here are our general contributions. We describe our specific contributions, in more detail, at the beginning of each subsequent chapter.

First and foremost, we develop several new real-time delay predictors applying to a broad range of models. The delay predictors that we propose are appealing because (i) they are easy to implement in practice, and (ii) they effectively cope with multiple real-life features such as non-exponential service and abandonment-time distributions, and time-varying demand and capacity. The main novelty of this work lies in systematically considering all those features which are relevant in practice. Direct mathematical analysis is often complicated in our models. That is why we resort to approximations which we also show are effective.

Second, we establish heavy-traffic limits that generate approximations for the expected MSE of some delay predictors in the $GI/M/s$ and $GI/GI/s+GI$ models. We verify the effectiveness of our approximations through computer simulation. Indeed, our approach to the delay prediction problem, throughout this thesis, combines both theoretical analysis and numerical support.

Third, we describe results of a wide range of simulation experiments, in a variety of settings, evaluating all alternative delay predictors proposed. As in some previous research (see §1.5), we also compare the accuracy of predictors exploiting system-state information to others relying only on model parameters. Our simulation study is exhaustive and provides ample support to our general conclusions.

Finally, we focus especially on delay predictors which are commonly used in practice, such as the QL predictor of chapter 2 and the (no-information) NI predictor (announcing the expected waiting time) of chapters 2 and 3, and alternative delay-history-based predictors. We study the performance of those predictors, show how and when they may not be effective, and propose new and more accurate predictors.

1.7 Organization

The rest of this thesis is organized as follows. In chapter 2, we consider the $GI/M/s$ model and study the performance of alternative delay-history-based predictors in that context. We compare the performance of those predictors to the standard QL predictor, commonly used in practice. In chapter 3, we consider the $GI/GI/s + GI$ model. We exploit established approximations for performance measures with a non-exponential abandonment-time distribution to obtain new delay predictors that effectively cope with non-exponential abandonment-time distributions. In chapter 4, we focus especially 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 refined delay-history-based predictors that effectively cope with time-varying arrival rates together with non-exponential service-time and abandonment-time distributions. In chapter 5, 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. In chapter 6, we draw conclusions. We present additional simulation results in the appendix. Even more simulation results are presented in online supplements to Ibrahim and Whitt (2009a,b, 2010a,b), available on the authors' webpages.

2

Delay Prediction in the $GI/M/s$ Model

2.1 Introduction

In this chapter, we study the performance of alternative real-time delay predictors based on recent customer delay experience in the standard $GI/M/s$ queueing model, emphasizing the case of large s. The main predictors considered are: (i) the delay of the last customer to enter service (LES), (ii) the delay experienced so far by the customer at the head of the line (HOL), and (iii) the delay experienced by the customer to have arrived most recently among those who have already completed service (RCS). We compare these delay-history-based predictors to the standard predictor based on the queue length (QL), commonly used in practice, which requires knowledge of the mean interval between successive service completions in addition to the queue length. We characterize performance by the mean squared error (MSE).

Our main contributions are to: (i) obtain analytical results for the conditional distribution of the delay given the observed HOL delay and propose its mean value as a refined predictor, (ii) establish heavy-traffic limits quantifying the difference in per-

formance between QL and HOL (LES) in the many-server and classical heavy-traffic limiting regimes, (iii) show that delay-history-based predictors are all asymptotically relatively efficient (ratio of MSE to the square of the mean converges to 0) in the many-server and classical heavy-traffic limiting regimes, and (iv) describe results of a wide range of simulation experiments evaluating the alternative delay predictors. This chapter is an edited version of Ibrahim and Whitt (2009a).

2.1.1 The $GI/M/s$ Model

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

$$A(t) \equiv \max\{n \geq 0 : U_1 + \cdots + U_n \leq t\}, \quad t \geq 0 \ . \tag{2.1}$$

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

2.1.2 The Standard Queue-Length (QL) Delay Predictor

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

$$\theta_{QL}(t) \equiv \frac{Q(t)+1}{r(t)}, \tag{2.2}$$

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

2.1.3 Predictors Based on Delay History

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

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

The HOL delay predictor was used as an announcement in an Israeli bank studied by Mandelbaum et al. (2000) and is mentioned as a candidate delay announcement by Nakibly (2002) in her study of delay predictions. In this study, we are motivated in part by recent work by Armony et al. (2009), who studied delay announcements in many-server queues with customer abandonment, focusing on customer response to the announcements, leading to balking and new abandonment behavior. They developed ways to approximately describe the equilibrium system performance using LES delay announcements. Armony et al. (2009) discuss the motivation for the LES delay predictor and other delays predictors based on recent delay history.

2.1.4 Quantifying the Effectiveness

We quantify the effectiveness of the delay predictors through the mean squared error (MSE), which we approximate analytically and estimate via simulation. To illustrate, let $W_{LES}(w)$ denote the random delay of a new arrival, conditional on that customer having to wait and an observed LES delay of w (under specified conditions, e.g., in steady state). Let $\theta_{LES}(w)$ be a candidate predictor based on this information. We will primarily be concerned with the direct predictor $\theta^d_{LES}(w) \equiv w$, the refined

predictor $\theta^r_{LES}(w) \equiv E[W_{LES}(w)]$ and approximations of the refined predictor, since the refined predictor is difficult to determine. The MSE of such an predictor is

$$MSE \equiv MSE(\theta_{LES}(w)) \equiv E\left[(W_{LES}(w) - \theta_{LES}(w))^2\right] \ . \tag{2.3}$$

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

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

$$ASE \equiv \frac{1}{n}\sum_{j=1}^{n}(a_j - p_j)^2 \ , \tag{2.4}$$

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

2.1.5 Study in an Idealized Setting

In this chapter, we study the performance of the delay-history-based predictors and compare them to the standard QL delay predictor in the relatively simple idealized setting of the $GI/M/s$ queueing model. Let m denote the mean service time. For this $GI/M/s$ model, the QL predictor $\theta_{QL}(t) \equiv m(Q(t) + 1)/s$ is an ideal delay predictor. Indeed, there are no serious competitors, as far as statistical precision is concerned (provided that we have no information about remaining service times). Given the queue length, the future evolution of the system is independent of the

past. (This even remain true for more general arrival processes.) Consequently, $\theta_{QL}(t)$ is the conditional mean delay given all information available at time t, so that it minimizes the MSE.

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

2.1.6 Motivation for Considering Alternative Delay Predictors

Whenever the actual service system is well modeled by a $GI/M/s$ queueing model and the system state is known accurately at each time, then there is little motivation for considering other delay predictors besides the standard QL predictor. However, real service systems rarely are as simple as the $GI/M/s$ model. First, the service-time distribution might well be non-exponential, as shown for call centers by Brown et al. (2005). Second, the number of servers and mean service times often change over time, in part because the servers are humans who serve in different shifts and may well have different service-time distributions. Third, the queue length may not be directly observable. That is nicely illustrated by the ticket queues studied by Xu et al. (2007). Upon arriving at a ticket queue, each customer is issued a numbered ticket. The number currently being served is displayed. The queue length is not

known to ticket-holding customers or even to system managers, because they do not observe customer abandonments.

Finally, the system is often much more complicated: For one example, there may be multiple customer classes and multiple service pools with some form of skill-based routing (SBR); see Gans et al. (2003). For a second example, with web chat, servers may serve several customers simultaneously, different servers may participate in a single service, and there may be interruptions in the service times, as the customers explore material on the web in between conversations with agents. For a third example, when delays are large – which is when we most want to make delay announcements – customers often abandon from queue; see chapter 3. In these more complicated settings, the queue length is typically known, but the rate customers enter service if often not known and/or difficult to estimate reliably. That causes problems for the QL predictor.

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

2.1.7 Example (non-exponential service times)

To dramatically illustrate the possible difficulties with the QL delay predictor in the presence of a non-exponential service-time distribution (without trying to be realistic), we consider a limiting hyperexponential (H_2) distribution, in which each service time is either an exponential with mean 10, with probability $1/10$, or the deterministic value 0, with probability $9/10$. Thus the service time has mean 1, but busy servers will only be serving customers with the exponential distribution. Let $s = 100$ and suppose that an arrival finds the queue empty but all the servers busy. Then the QL delay prediction for this new arrival is $1/s = 1/100$, but the actual delay is

exponentially distributed with mean 1/10 (the minimum of 100 exponential random variables, each with mean 10). Hence, the actual mean delay is ten times greater than predicted by the QL delay predictor. Consistent with this extreme example, we have found that our alternative delay predictors actually outperform the QL delay predictor in the $D/H_2/100$ model with moderately variable H_2 distributions. ■

Similarly, when there is a large amount of customer abandonment, the QL predictor will tend to overestimate the potential delay (the delay assuming that the customer has infinite patience), because many customers in queue may abandon before entering service, and the standard QL predictor fails to take that into account. As discussed in Whitt (1999a), the QL predictor can be revised to provide an accurate prediction of delays with abandonments when the time-to-abandon distribution is exponential. However, as discussed in Whitt (2006), the performance measures in the overloaded $M/M/s+GI$ model, with non-exponential time-to-abandon distribution, depend strongly on the time-to-abandon distribution beyond its mean. Since the time-to-abandon distribution has been found to be non-exponential in practice, see Brown et al. (2005), there also are potential difficulties with the generalized QL predictor based on the $M/M/s+M$ model. We investigate alternative delay predictors in the presence of abandonments in chapter 3. There we give examples with non-exponential distributions in which both the standard QL predictor and the refinement for the $M/M/s+M$ model are outperformed by delay predictors based on recent delay history.

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

2.1.8 This Study

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

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

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

2.1.9 Organization of the Chapter

We start in §2.2 by defining alternative delay predictors based on recent delay history and giving some expressions for them for the $GI/M/s$ model. We present results

of initial simulation experiments in §2.3. We establish properties of two basic delay predictors – LES and the Head-of-the-Line (HOL) predictor – in §2.4. We present confirming simulations related to those analytical results in §2.5. We establish heavy-traffic limits for some predictors in §2.6. We make concluding remarks in §2.7. We present additional simulation results in the appendix.

2.2 Alternative Predictors

2.2.1 The No-Information (NI) Steady-State Predictor

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

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

2.2.2 The Full-Information Queue-Length (QL) Delay Predictor

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

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

$$W_Q(n) \equiv \sum_{i=1}^{n+1} (S_i/s) \ , \tag{2.5}$$

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

We have the division by s in (2.5) because the times between successive service completions when all servers are busy are i.i.d. random variables distributed as

the minimum of s exponential random variables, each with mean 1, which makes the minimum exponential with mean $1/s$. It is significant that this predictor is independent of the arrival process and thus also of the traffic intensity. It applies equally well to steady-state and transient settings.

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

$$E[W_Q(n)] = \frac{n+1}{s}, \quad Var[W_Q(n)] = \frac{n+1}{s^2}$$
$$\text{and} \quad c^2_{W_Q(n)} \equiv \frac{Var[W_Q(n)]}{(E[W_Q(n)])^2} = \frac{1}{n+1}\ , \tag{2.6}$$

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

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

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

e.g., see Proposition 5.6.3 of Ross (1996). Hence,

$$\begin{aligned} E[MSE(W_Q(Q_\infty^w))] &\equiv \sum_{n=0}^{\infty} MSE(W_Q(n))P(Q_\infty^w = n) \\ &= E[Var(W_Q(Q_\infty^w))] \\ &= \frac{1}{s^2(1-\rho)} \ , \end{aligned} \tag{2.7}$$

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

$$\frac{MSE(\theta_{NI})}{MSE(\theta_{QL}(Q_\infty^w))} = \frac{Var(W_\infty | W_\infty > 0)}{E[Var(W_Q(Q_\infty^w))]} = \frac{1/s^2(1-\rho)^2}{1/s^2(1-\rho)} = \frac{1}{1-\rho} \ . \tag{2.8}$$

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

2.2.3 The Last Customer to Enter Service (LES)

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

To a large extent, the alternative refined LES delay predictors (and others as well) are obtained by replacing the known queue length n in (2.5) by random variables that estimate the queue length, based on the available delay history. Let $W_{LES}(w, d)$ be the delay of a new arrival, given that the new arrival must wait before starting service

and given that the last customer to enter service experienced delay w before entering service and there was elapsed time d since that customer entered service. Let t_a be the arrival epoch of the new customer and t_e be the time the last customer entered service prior to t_a. (Throughout this paper we use the fact that, almost surely, no two events – arrivals or service completions – will occur simultaneously.) Necessarily, $d = t_a - t_e$ and $t_e - w$ is the arrival epoch of the customer entering service at t_e. With Poisson arrivals, a key observation is that the queue length at time t_e must be distributed as $A(w)$, because customers enter service from the queue in order of arrival. However, $W_{LES}(w, d)$ has a relatively complicated exact distribution, even with Poisson arrivals, because we do not know precisely what happens in the interval $[t_e, t_a]$.

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

$$W_{LES}(w, d) \equiv \sum_{i=1}^{A(w+d)+1} (S_i/s) \ , \tag{2.9}$$

where the summands are i.i.d. and independent of $A(w + d)$, because the queue length seen by the new arrival at time t_a will be $A(w+d)$, the number of arrivals in the interval of length $w+d$ preceding the arrival epoch t_a. With a general (not Poisson) renewal arrival process, the analysis is complicated by conditioning on both arrival epochs, t_a and $t_e - w$, which may affect the distribution of the number of arrivals

between those two epochs; e.g., see §2.2.4.1 for an example. In general, we think of the arrival process as if it were Poisson. Formula (2.9) allows us then to characterize the distribution of $W_{LES}(w, d)$. Just like (2.5), (2.9) requires knowledge of s and the mean service time as well as w. Here we also require knowledge of the renewal arrival process or, equivalently, the interarrival-time distribution.

An important reference point for the refined LES predictor in (2.9) is the $D/M/s$ model, with a deterministic arrival process, having constant interarrival times, because under the extra condition leading to (2.9), we then have $W_{LES}(w, d) = W_Q(Q(t_a))$, since $A(w + d) = Q(t_a)$, making (2.5) coincide with (2.9). Thus we see that the loss of efficiency in going from QL to LES (direct or refined) is primarily due to the variability in the arrival process.

We assume that the experienced LES waiting time w is always available, but we might not know d, so that we might want to consider as an alternative refined predictor the mean of the random variable $W_{LES}(w)$, which assumes d is unavailable, but dropping d makes the distribution even more complicated. If we can assume that $w >> d$, then there should be negligible difference. In general, we have the natural approximations based on (2.9):

$$W_{LES}(w) \approx \sum_{i=1}^{A(w+(S_0/s))+1} (S_i/s) \approx \sum_{i=1}^{A(w+(1/s))+1} (S_i/s) \ , \tag{2.10}$$

where S_0 is an exponential random variable with mean 1 independent of S_i for $i \geq 1$, because the time between successive service completions when all servers are busy is distributed as S_0/s. (Assuming that the queue is nonempty at time t_e, that time is a service completion epoch. Then d is the age of the Poisson all-servers-busy departure process with rate s under Poisson inspection by the arrival process.) The second approximation is obtained by inserting the expected value. It is also based

on the extra condition, which will hold approximately for large w.

2.2.4 The Head-Of-The-Line (HOL) Predictor

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

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

the previous formulas, with a Poisson arrival process, we have

$$W_{HOL}(w) \equiv \sum_{i=1}^{A(w)+2} (S_i/s) \ . \tag{2.11}$$

2.2.4.1 $W_{HOL}(w)$ with a renewal (non-Poisson) arrival process

Formula (2.11) is exact with a Poisson arrival process and is an approximation more generally. Indeed, if we condition on the current time being an arrival epoch, then given the observed HOL delay w, we know that the current HOL customer and the new arriving customer are exactly w time units apart. For a general interarrival-time distribution, this adds extra information and considerably complicates the analysis. To illustrate, suppose that the interarrival-time distribution is a mixture of two distributions. With probability 0.999, the interarrival time is exponential with mean 1, and with probability 0.001 it is the deterministic value $M >> 1$. If the HOL delay w at an arrival epoch is equal to M, then we know with probability 1 that there must have been 0 arrivals between the HOL arrival and the new arrival (and that the new arrival must wait for exactly 2 service completions to begin service). That is, conditioning on both arrival epochs changes the distribution of the number of arrivals between those epochs, and formula (2.11) no longer applies.

In this work, we approximate the renewal arrival process by a Poisson process, with the same interarrival-time mean, in which case formula (2.11) is exact. We can also take an alternative view for which formula (2.11) is exact even with a general interarrival-time distribution. Indeed, we can think of predicting delays in continuous time, thus making predictions at time t for a hypothetical arrival at time t. With a continuous-time view, we avoid conditioning on t being an arrival epoch, thus simplifying the analysis.

2.2.5 The Delay of the Last Customer to Complete Service (LCS)

A third candidate direct delay predictor is the delay of the last customer to *complete* service (LCS). We naturally would want to consider this alternative predictor if we only learn customer delay experience after they complete service. That might be the case for customers and outside observers.

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

2.2.6 The Delay of the Most Recent Arrival to Complete Service (RCS)

Under some circumstances, the LCS and LES direct predictors will be similar, but they actually can be very different when s is large, because the last customer to complete service may have experienced his waiting time much before the last customer to enter service. We emphasize that customers need not depart in order of

arrival. Indeed, with exponential service times, when all s servers are busy, each of the s servers is equally likely to generate the next service completion. Thus, for large s the LCS predictor is not really a viable alternative, as we will show. Consequently, we propose other candidate delay predictors based on the delay experience of customers that have already completed service. The first is the delay experienced by the customer that arrived most recently (and thus entered service most recently) among those customers who have already completed service (RCS). We find that RCS is far superior to LCS when s is large.

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

A disadvantage of the RCS predictor is that we must analyze a lot of data, going arbitrarily far back in the past. From heavy-traffic analysis in §2.6, we deduce that the most recent arrival time of a customer that has completed service is very likely to occur among the last $c\sqrt{s}$ customers when s is large (and the system is normally loaded). So we introduce a new predictor, which requires less information processing: Let RCS-$c\sqrt{s}$ be the delay of the customer to have arrived most recently among the last $c\sqrt{s}$ customers who have already completed service. Clearly, these last three predictors LCS, RCS and RCS-$c\sqrt{s}$ are complicated, so that we primarily rely on simulation to evaluate their relative performance. Through extensive simulation experiments, we found that the average squared error of RCS-$c\sqrt{s}$ is essentially identical to that of RCS when $c = 4$, differs by at most 1% when $c = 2$ and differs by at most 10% when $c = 1$; for corresponding simulation results, see the appendix.

2.2.8 Averages

Our main predictors are individual delays experienced by a recent customer, rather than an average over many past delays. Only the no-information steady-state predictor ($W_\infty | W_\infty > 0$) can be said to use averages. We can extend the LES, LCS, RCS and RCS-$c\sqrt{s}$ predictors to get LES-k, LCS-k, RCS-k and RCS-$c\sqrt{s}-k$ by averaging over the last k experienced delays. With the exception of LCS with large s (which does not have desirable properties), we have found that averages do not help, when the delays are relatively large (the case of primary interest to us). There is a simple explanation: When delays are large, the delays change relatively slowly compared to the size of the delays. Theoretically, this can be explained by the heavy-traffic snapshot principle; see Section 2.6. In this setting it is better to use recent information than to eliminate noise by averaging.

2.3 Initial Simulation Experiments: Comparing the Predictors

In this section we present initial simulation experiments, aiming to compare the alternative predictors defined in §2.2. We focus on the *average squared error* (ASE) of the predictor, defined in (2.4). For large samples, the ASE should agree with the MSE in steady state.

2.3.1 Overall Performance of the Predictors

Table 2.1 shows the ASE's for seven different delay predictors in the $GI/M/s$ model with $s = 100$. Tables 2.2 and 2.3 shows the ASE's of the same predictors with

$s = 10$ and $s = 1$, respectively. We consider three categories of predictors: (i) the two reference predictors QL and NI, (ii) the direct delay predictors LES and HOL, and (iii) the three predictors based on delays of customers who have already completed service - LCS, RCS and RCS-$\sqrt{s}$. We consider three interarrival-time distributions - M, D and H_2 - and four values of the traffic intensity ρ - 0.98, 0.95, 0.93 and 0.90. The H_2 distribution has SCV $c_a^2 = 4$ and balanced means (the two component exponential distributions contribute equally to the mean). We performed 10 independent replications of long runs in each case. The half width of the 95% confidence interval is shown below each estimate. Corresponding results for other values of s - 400 and 900 - are contained in the appendix.

2.3.2 The Case with $s = 100$

The predictors appear in Table 2.1 with the better performance toward the left; i.e., in terms of efficiency (low ASE), the predictors are ordered by

$$QL > LES \approx HOL > RCS \approx RCS - \sqrt{s} > LCS > NI \ . \qquad (2.12)$$

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

Estimated ASE in units of 10^{-3}

$M/M/s$ **model with** $s = 100$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	5.03 ±0.02	10.2 ±0.05	10.2 ±0.05	12.5 ±0.05	12.9 ±0.05	26.7 ±0.06	255 ±36
0.95	2.04 ±0.02	4.3 ±0.05	4.3 ±0.05	6.4 ±0.05	6.7 ±0.05	16.5 ±0.06	41.8 ±2.7
0.93	1.44 ±0.002	3.07 ±0.003	3.08 ±0.003	5.06 ±0.003	5.32 ±0.003	13.1 ±0.13	20.8 ±1.2
0.90	0.99 ±0.003	2.2 ±0.006	2.2 ±0.006	3.9 ±0.008	4.2 ±0.009	9.4 ±0.27	9.7 ±0.7

$D/M/s$ **model with** $s = 100$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	2.48 ±0.05	2.62 ±0.05	2.62 ±0.05	3.77 ±0.05	3.94 ±0.05	10.3 ±0.11	61.5 ±3.9
0.95	1.01 ±0.02	1.15 ±0.02	1.15 ±0.02	2.20 ±0.03	2.34 ±0.03	6.38 ±0.12	10.1 ±0.40
0.93	0.73 ±0.02	0.87 ±0.02	0.87 ±0.02	1.85 ±0.03	1.96 ±0.03	4.90 ±0.13	5.20 ±0.32
0.90	0.52 ±0.015	0.67 ±0.016	0.66 ±0.017	1.54 ±0.035	1.63 ±0.037	3.44 ±0.15	2.68 ±0.23

$H_2/M/s$ **model with** $s = 100$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	12.4 ±0.70	60.4 ±3.2	60.4 ±3.2	66.1 ±3.2	67.0 ±3.2	103.4 ±34.0	1505 ±226
0.95	4.82 ±0.095	22.5 ±0.46	22.5 ±0.47	27.7 ±0.45	28.4 ±0.45	56.3 ±0.58	243.3 ±22.7
0.93	3.44 ±0.094	15.5 ±0.44	15.5 ±0.44	20.4 ±0.49	21.1 ±0.50	44.5 ±1.02	121.4 ±10.2
0.90	2.35 ±0.040	10.2 ±0.21	10.2 ±0.21	14.6 ±0.24	15.2 ±0.24	33.1 ±0.53	55.4 ±2.9

Table 2.1: A comparison of the efficiency of different real-time delay predictors for the $GI/M/100$ queue as a function of the traffic intensity ρ and the interarrival-time distribution (M, D and H_2). Only the direct predictors are considered. Estimates of the average squared error ASE are shown together with the half width of the 95% confidence interval. The units are 10^{-3} throughout. The ASE's are measured in units of mean service time squared per customer.

Estimated ASE in units of 10^{-1}

$M/M/s$ **model with** $s = 10$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	4.95 ±0.23	10.1 ±0.42	10.1 ±0.41	10.8 ±0.41	10.9 ±0.42	11.9 ±0.41	257.2 ±48.1
0.95	1.98 ±0.025	4.16 ±0.040	4.17 ±0.042	4.83 ±0.039	4.94 ±0.041	5.87 ±0.041	39.61 ±2.3
0.93	1.42 ±0.013	3.03 ±0.032	3.05 ±0.037	3.67 ±0.036	3.77 ±0.033	4.62 ±0.036	20.01 ±0.66
0.9	1.00 ±0.017	2.19 ±0.033	2.20 ±0.042	2.79 ±0.036	2.88 ±0.035	3.63 ±0.036	10.10 ±0.49

$D/M/s$ **model with** $s = 10$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	2.49 ±0.084	2.63 ±0.083	2.63 ±0.086	2.99 ±0.085	3.05 ±0.086	3.57 ±0.086	59.3 ±10.2
0.95	1.01 ±0.018	1.16 ±0.018	1.16 ±0.020	1.50 ±0.019	1.55 ±0.019	2.00 ±0.019	10.1 ±0.83
0.93	0.730 ±0.010	0.876 ±0.011	0.877 ±0.013	1.21 ±0.012	1.26 ±0.011	1.66 ±0.012	5.24 ±0.29
0.9	0.518 ±0.0058	0.663 ±0.0057	0.663 ±0.0091	0.977 ±0.0077	1.02 ±0.0066	1.37 ±0.0078	2.66 ±0.12

$H_2/M/s$ **model with** $s = 10$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.98	12.8 ±0.69	62.6 ±4.0	62.6 ±4.1	64.4 ±4.1	65.1 ±4.1	67.3 ±5.6	1594 ±258
0.95	4.81 ±0.081	22.3 ±0.47	22.3 ±0.48	23.9 ±0.47	24.6 ±0.47	26.5 ±0.81	229 ±9.1
0.93	3.42 ±0.069	15.4 ±0.35	15.4 ±0.37	17.0 ±0.35	17.5 ±0.35	19.4 ±0.35	115 6.8
0.9	2.34 ±0.036	10.1 ±0.18	10.1 ±0.20	11.6 ±0.19	11.8 ±0.18	13.7 ±0.18	54.4 ±2.9

Table 2.2: A comparison of the efficiency of different real-time delay predictors for the $GI/M/10$ queue as a function of the traffic intensity ρ and the interarrival-time distribution (M, D and H_2). Only the direct predictors are considered. Estimates of the average squared error ASE are shown together with the half width of the 95% confidence interval. The units are 10^{-1} throughout. The ASE's are measured in units of mean service time squared per customer.

Estimated ASE

$M/M/s$ **model with** $s = 1$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.95	20.1 ±0.42	42.2 ±0.77	42.4 ±0.79	44.1 ±0.78	44.1 ±0.78	44.1 ±0.78	405.0 ±23.4
0.93	14.4 ±0.19	30.6 ±0.37	30.7 ±0.39	32.4 ±0.37	32.4 ±0.37	32.4 ±0.37	207.5 ±10.4
0.9	9.99 ±0.084	21.8 ±0.19	22.0 ±0.21	23.5 ±0.19	23.5 ±0.19	23.5 ±0.19	100.6 ±3.4
0.85	6.68 ±0.043	15.1 ±0.093	15.4 ±0.095	16.6 ±0.010	16.6 ±0.010	16.6 ±0.010	44.9 ±0.88

$D/M/s$ **model with** $s = 1$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.95	10.1 ±0.15	11.6 ±0.15	11.6 ±0.16	12.6 ±0.15	12.6 ±0.15	12.6 ±0.15	101.1 ±7.2
0.93	7.32 ±0.081	8.79 ±0.078	8.79 ±0.086	9.73 ±0.080	9.73 ±0.080	9.73 ±0.080	52.7 ±2.4
0.9	5.19 ±0.038	6.64 ±0.037	6.65 ±0.041	7.56 ±0.040	7.56 ±0.040	7.56 ±0.040	26.8 ±0.94
0.85	3.53 ±0.018	4.96 ±0.018	4.95 ±0.020	5.82 ±0.020	5.82 ±0.021	5.82 ±0.020	12.4 ±0.36

$H_2/M/s$ **model with** $s = 1$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.95	48.7 ±1.13	226.4 ±5.14	226.5 ±5.23	231.1 ±5.15	231.1 ±5.15	231.1 ±5.15	2339 ±425
0.93	34.3 ±0.63	154.4 ±2.9	154.4 ±2.9	158.9 ±3.0	158.9 ±3.0	158.9 ±3.0	1151 ±181
0.9	23.48 ±0.37	101.3 ±2.3	101.4 ±2.4	105.5 ±2.4	105.5 ±2.4	105.5 ±2.4	552.9 ±103
0.85	14.95 ±0.104	60.0 ±0.52	60.2 ±0.53	63.9 ±0.51	63.9 ±0.51	63.9 ±0.51	224.4 ±6.2

Table 2.3: A comparison of the efficiency of different real-time delay predictors for the $GI/M/1$ queue as a function of the traffic intensity ρ and the interarrival-time distribution (M, D and H_2). Only the direct predictors are considered. Estimates of the average squared error ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

$s = 10$ and $s = 1$ in Tables 2.2 and 2.3, but do not discuss those results here. In broad terms, the predictor LCS fares better as s decreases. The ASE's of LCS and RCS do not differ greatly for $s = 10$ and are identical for $s = 1$.

It is instructive to look at the relative average squared error (RASE), which is obtained by dividing the ASE by $E[W_\infty|W_\infty > 0]^2$, because the associated steady-state relative mean squared error (RMSE), defined as $MSE/E[W_\infty|W_\infty > 0]^2$, is *linear* as a function of ρ for the QL predictor: $RMSE(QL) = (1 - \rho)$. (The RMSE is identically 1 for the NI predictor.) We show the RASE plots for the $D/M/100$ model in Figure 2.1. The LES and HOL predictors are virtually identical (with the plots lying on top of each other), so we only show LES. Both LES and HOL are nearly as good as QL and much better than RCS; LCS is so bad that it is not even shown. We display the RASE's for the $M/M/100$ and $H_2/M/100$ models in Figures 2.2 and 2.3. Again we see linear or near-linear performance as a function of ρ. The advantage of QL over LES increases as c_a^2 increases.

2.3.3 Performance Conditional on the Level of Delay

Since delay predictions are more relevant when the observed delays in the system are longer, it is natural to consider the behavior of the predictors for larger delays. We have complemented the experiments described above by considering how the delay predictors perform when we only consider actual delays that fall in one of the intervals: $(E[W|W > 0], 2E[W|W > 0])$, $(2E[W|W > 0], 4E[W|W > 0])$, $(4E[W|W > 0], 6E[W|W > 0])$ and $(6E[W|W > 0], \infty)$. Table 2.4 illustrates the results for the $M/M/100$ model when the observed delays fall in the interval $(4E[W|W > 0], 6E[W|W > 0])$. The performance of the predictors for these larger delays is approximately as in Table 2.1. Other cases appear in the appendix.

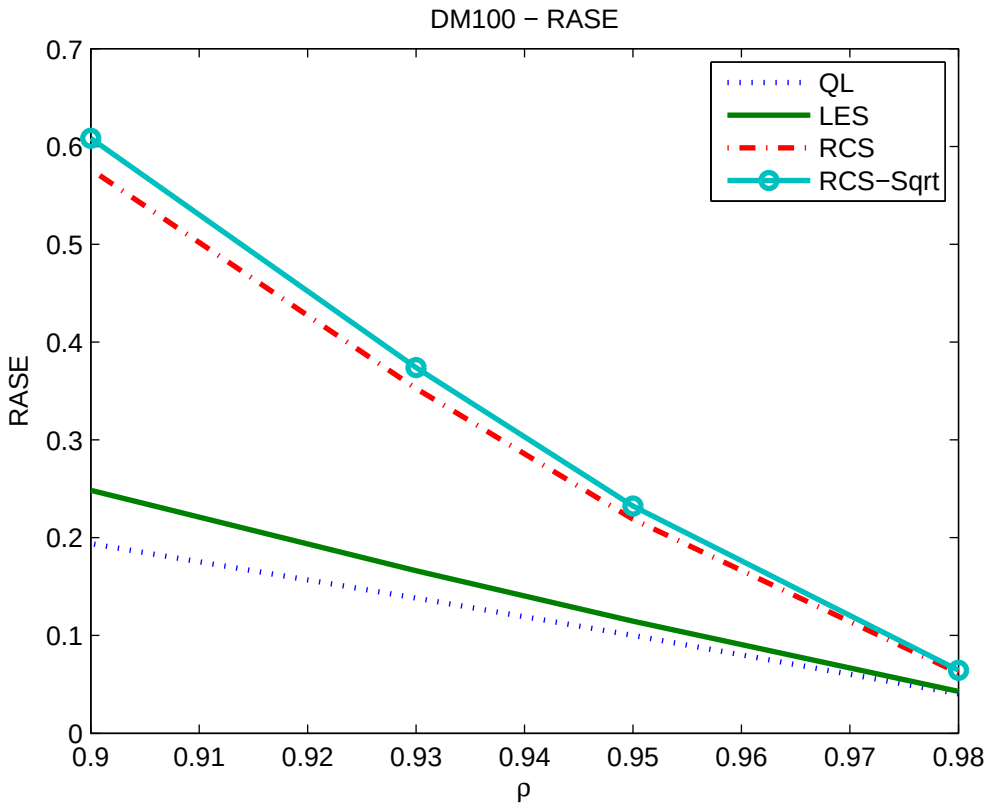

Figure 2.1: The relative average squared error ($RASE$) for the $D/M/100$ model.

Experience shows that the NI predictor performs especially poorly in very heavy traffic, while LCS performs especially poorly with large s in light traffic. For large s and small ρ, LCS even performs worse than the NI predictor. There is only one case in Table 2.4; more cases can be seen when $s = 400$ and $s = 900$ in the appendix.

2.4 Analysis of the HOL and LES Predictors

The representation (2.11) allows us to characterize the probability distribution of the random variable $W_{HOL}(w)$, which we do both for its own sake and as an approximation for the random variables $W_{LES}(w)$ and $W_{RCS}(w)$. When we use the HOL

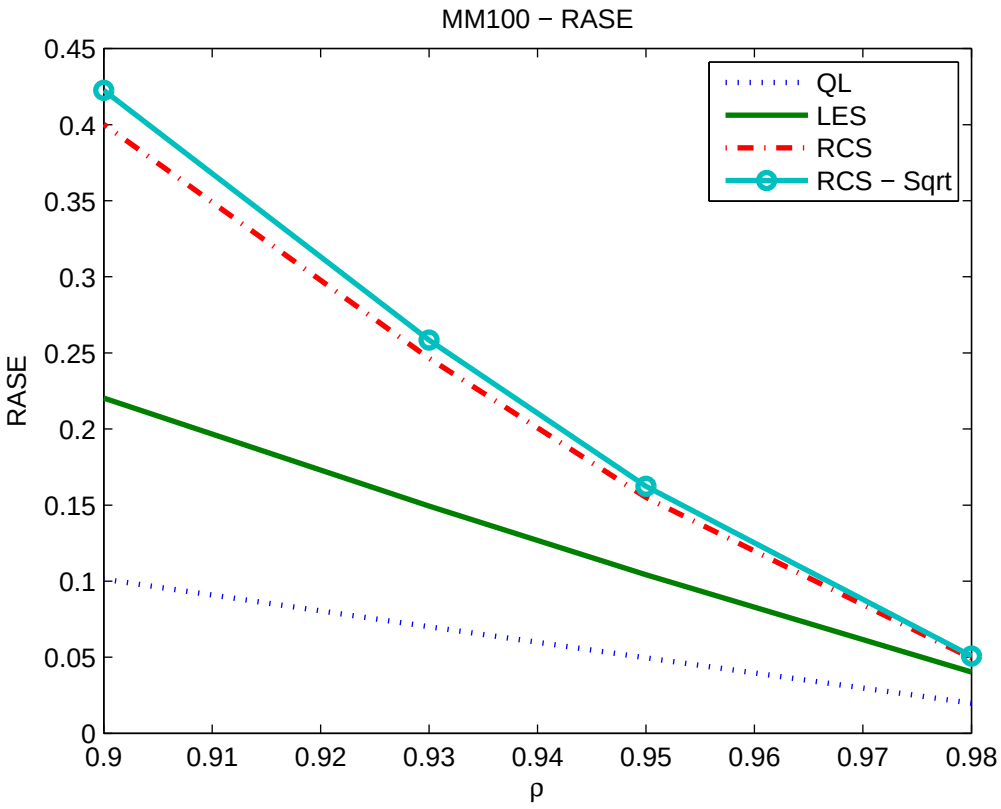

Figure 2.2: The relative average squared error ($RASE$) for the $M/M/100$ model.

predictor, we assume that there is at least one customer in queue at the new arrival epoch t_a. Very similar formulas hold for the LES predictor based on formula (2.9), under the extra assumption given there. Since the formulas are virtually identical, we do not display separate results for LES.

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

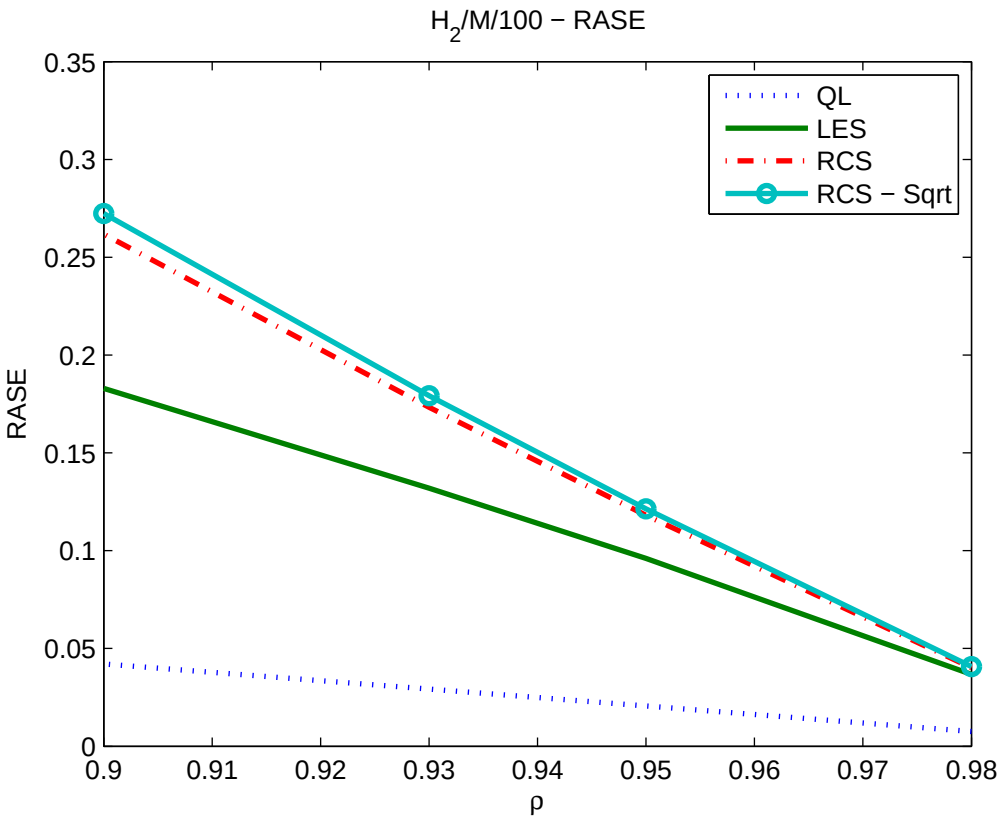

Figure 2.3: The relative average squared error ($RASE$) for the $H_2/M/100$ model.

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

Theorem 2.4.1 (*dependence upon s*) *For the $GI/M/s$ model,*

$$W_{HOL,s}(w) \stackrel{\mathrm{d}}{=} \frac{W_{HOL,1}(sw)}{s} \tag{2.13}$$

for all ρ, w and s.

Proof. We show the equality in distribution by establishing equality w.p.1 for a special construction. We construct a convenient family of systems indexed by s.

Conditional ASE for the $M/M/100$ model in units of 10^{-3}
Observed delays in between $4E[W|W>0]$ and $6E[W|W>0]$

ρ	QL	LES	HOL	RCS	RCS-$\sqrt{s}$	LCS	NI
0.99	49.4 ±7.0	86.6 ±6.9	86.3 ±6.9	89.4 ±7.2	90.1 ±7.2	108.8 ±10.2	11,586 ±1250
0.98	24.8 ±1.8	47.5 ±3.1	47.3 ±3.0	50.1 ±3.1	50.6 ±3.1	69.6 ±3.7	3,542 ±431
0.95	10.5 ±0.23	20.4 ±0.63	20.1 ±0.62	23.5 ±0.82	24.0 ±0.80	50.4 ±3.3	564 ±27
0.93	7.54 ±0.20	15.2 ±0.31	14.9 ±0.29	18.7 ±0.43	19.3 ±0.45	52.0 ±3.2	286 ±8.0
0.90	5.62 ±0.21	11.1 ±0.38	10.7 ±0.38	15.3 ±0.61	16.1 ±0.66	50.9 ±25.2	137.4 ±6.7

Table 2.4: A comparison of the efficiency of different real-time delay predictors conditional on the level of delay experienced for the $M/M/100$ model as a function of the traffic intensity ρ. Actual delays are considered that fall in the interval $(4E[W|W>0], 6E[W|W>0])$. Estimates of the conditional average squared error ASE are shown together with the half width of the 95% confidence interval. The units are 10^{-3} throughout. The ASE's are measured in units of mean service time squared per customer.

For each s, let the service times be exponential random variables S_n with mean 1 as before. Start by defining interarrival times U_n with mean $1/\rho$ to use for the case of $s = 1$. Then in the system with $s > 1$, let the n^{th} interarrival time be U_n/s. Let $\{A_s(t) : t \geq 0\}$ be the renewal counting process in system s, having interarrival times U_n/s. Then $A_s(w/s) = A_1(w)$ for all s and w; since we have re-scaled the interarrival times, we just re-scale time in the associated renewal counting process. This construction yields equality for the random variables in (2.13) and all $w \geq 0$. Since the distribution is independent of the construction, that implies the claimed relation (2.13). ∎

We now show that we get relatively simple asymptotic expressions characterizing the distribution of $W_{HOL,s}(w)$ when $sw \to \infty$. That applies when $w \to \infty$ for fixed s, but it also can apply when $s \uparrow \infty$ and $w \downarrow 0$, as occurs in the QED many-server

heavy-traffic limiting regime, to be discussed in §2.6; then $w = O(1/\sqrt{s})$ so that $sw \to \infty$ while $w \to 0$.

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

Theorem 2.4.2 (*distribution of* $W_{HOL,s}(w)$) *Consider the* $GI/M/s$ *queue with traffic intensity* ρ *operating in the time interval* $[t_a - w, t_a]$. (*a*) *For any* $\rho > 0$, $s \geq 1$ *and* $w > 0$,

$$E[W_{HOL,s}(w)] = \frac{E[A(w)] + 2}{s} \tag{2.14}$$

and

$$Var[W_{HOL,s}(w)] = E[A(w) + 2]Var(S/s) + Var(A(w) + 2)(E[S/s])^2. \tag{2.15}$$

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

$$E[W_{HOL,s}(w)] = \rho w + \frac{2}{s} \tag{2.16}$$

and

$$Var[W_{HOL,s}(w)] = \frac{2\rho w}{s} + \frac{2}{s^2} \ , \tag{2.17}$$

so that

$$c^2_{W_{HOL,s}(w)} = \frac{2}{\rho s w} - \frac{6}{(\rho s w)^2} + O\left(\frac{1}{(\rho s w)^3}\right) \quad as \quad sw \to \infty \ . \tag{2.18}$$

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

$$sE[W_{HOL,s}(w)] - \rho s w \to \frac{(c_a^2 + 3)}{2} \ , \tag{2.19}$$

$$\frac{W_{HOL,s}(w)}{w} \to \rho \quad w.\ p.\ 1 \quad and \quad \frac{E[W_{HOL,s}(w)]}{w} \to \rho \ , \tag{2.20}$$

$$s^2 Var(W_{HOL,s}(w)) - \rho s w(c_a^2+1) \to \left(\frac{5(c_a^2+1)^2}{4} - \frac{2\nu_a^3}{3} + 1\right) \ , \tag{2.21}$$

$$s^2 E[(W_{HOL,s}(w) - \rho w)^2] - \rho s w(c_a^2+1) \to K \ , \tag{2.22}$$

$$s^2 E[(W_{HOL,s}(w) - w)^2] - (sw)^2(1-\rho)^2 - sw\left[(2\rho-1)c_a^2 + 4\rho - 3\right] \to K \ , \tag{2.23}$$

where

$$K \equiv K(c_a^2, \nu_a^3) \equiv \left(\frac{3c_a^4}{2} + 4c_a^2 + \frac{9}{2} - \frac{2\nu_a^3}{3}\right) \ , \tag{2.24}$$

$$sw c^2_{W_{HOL,s}(w)} \to \frac{c_a^2+1}{\rho} \quad and \quad \frac{W_{HOL,s}(w) - \rho w}{\sqrt{\rho w(c_a^2+1)/s}} \Rightarrow N(0,1) \ . \tag{2.25}$$

Proof. Since $W_{HOL}(w)$ in (2.11) is a random sum of i.i.d. random variables, where $A(w)$ is independent of the summands S_i/s, we have (2.14). Formula (2.15) follows from the conditional variance formula, e.g., p. 51 of Ross (1996). For (2.18), we use elementary operations on series, as in 3.6.22 in Abramowitz and Stegun (1972). When we let sw increase, we first apply Theorem 2.4.1 to reduce the analysis to the case $s = 1$. Henceforth assume that $s = 1$. When we restrict attention to $s = 1$, it suffices to simply let $w \to \infty$. When we let w increase,

$$E[A(w) + 2] - \rho w \to \frac{(c_a^2+1)}{2} + 1 \quad \text{as} \quad w \to \infty \ , \tag{2.26}$$

see Corollary 3.4.7 of Ross (1996) or (2.7) and (2.8) of Whitt (1982a), which review a classic result. Combining (2.26) and (2.14) gives (2.19), which immediately implies the second limit in (2.20). For the w.p.1 limit in (2.20), we apply the strong law of large numbers for the partial sums of S_n and the renewal arrival process $A(w)$:

With probability one,

$$\frac{\sum_{i=1}^{n} S_i}{n} \to E[S] = 1 \quad \text{and} \quad \frac{A(w)+2}{w} \to \frac{1}{E[U]} = \rho \ , \tag{2.27}$$

so that

$$\frac{\sum_{i=1}^{A(w)+2} S_i}{w} = \frac{A(w)+2}{w} \times \frac{\sum_{i=1}^{A(w)+2} S_i}{A(w)+2} \to \rho \quad \text{w. p. 1} \ . \tag{2.28}$$

The asymptotic variance formula (2.21) follows from (2.15) and the asymptotic form of the variance for a renewal process, e.g., as in (2.7) and (2.8) of Whitt (1982a):

$$\begin{aligned} Var(A(w)+2) \ &= Var(A(w)) = \rho w c_a^2 + \tfrac{5(c_a^2+1)^2}{4} - \tfrac{2\nu_a^3}{3} - \tfrac{(c_a^2+1)}{2} + o(1) \quad (2.29) \\ &\text{as} \quad w \to \infty \ . \end{aligned}$$

The associated limits (2.22) and (2.23) follow from (2.21). For (2.22), we use

$$\begin{aligned} E[(W_{HOL,s}(w) - \rho w)^2] \ &= \ var(W_{HOL,s}(w) - \rho w)) + (E[W_{HOL,s}(w) - \rho w])^2 \\ &= \ var(W_{HOL,s}(w)) + (E[W_{HOL,s}(w) - \rho w])^2 \ . \quad (2.30) \end{aligned}$$

The calculation for (2.23) is similar. The first limit in (2.25) follows immediately from (2.19) and (2.21). The central limit theorem in (2.25) follows from the central limit theorem for renewal-reward processes, e.g., Theorem 7.4.1 of Whitt (2002). We use the convergence-together theorem, Theorem 11.4.7 of Whitt (2002), to justify neglecting the asymptotically negligible terms. ∎

Remark 2.4.1 (exact values by numerical inversion) It is possible to exploit (2.14) and (2.15) in order to compute the exact means and variances. To do so, we can exploit numerical transform inversion of Laplace transforms, as discussed in

§13 of Abate and Whitt (1992). The Laplace transform of $E[A(t)]$ is $\hat{m}_1(s) \equiv \hat{f}(s)/[s(1-\hat{f}(s)]$, where $\hat{f}(s)$ is the Laplace transform of the density function of the interarrival-time cdf F (here assumed to exist). The associated Laplace transform of $E[A(t)^2]$ is $2\hat{m}_1(s)^2 - \hat{m}_1(s)$, as can be seen from exercise XI.13 on p. 386 of Feller (1971). Since we are interested in prediction for relatively large delays, we will rely on the asymptotic approximations. ■

Remark 2.4.2 (nonhomogeneous Poisson arrival process) We can also analyze the random variable $W_{HOL,s}(w)$ in the case of a nonhomogeneous Poisson arrival process with intensity function $\{\lambda(t) : t \geq 0\}$. The exact relations (2.16) and (2.17) have natural extensions to that case. We again have representation (2.11), but now with $A(w)$ being a Poisson random variable having mean

$$m_a(w) \equiv \int_{t_a - w}^{t_a} \lambda(t)\, dt \ , \tag{2.31}$$

which depends on the arrival time t_a and the intensity function as well as the experienced waiting time w. Unless we specify how the intensity function behaves, we have no simple asymptotic story as w increases, though. For more on the analysis of the HOL predictor with time-varying arrivals, see chapter 4 ■

Theorem 2.4.2 shows that the first-order asymptotic behavior of the random variable $W_{HOL,s}(w)$ as sw increases depends on the general interarrival-time distribution F only through its first two moments or, equivalently, through the mean $E[U] = 1/\rho s$ and the SCV c_a^2. Equations (2.21) and (2.25) show that both the variance $Var(W_{HOL,s}(w))$ and the SCV $c^2_{W_{HOL,s}(w)}$ are approximately proportional to $c_a^2 + 1$ for large sw.

Theorem 2.4.2 shows that it may be useful to consider various refined predictors instead of the direct predictor $\theta^d_{HOL} \equiv w$. We would want to use the refined predictor

$\theta^r_{HOL} \equiv E[W_{HOL,s}(w)]$, because the mean necessarily minimizes the MSE, but we do not have a convenient formula for the mean. Theorem 2.4.2 leads us to consider two other refined predictors: the *simple refined predictor* $\theta^{sr}_{HOL} \equiv \rho w$ and the *asymptotic refined predictor* $\theta^{ar}_{HOL} \equiv \rho w + (c_a^2 + 3)/(2s)$, based on the the limit (2.19) as $sw \to \infty$. Note that the formulas for the mean and variance for Poisson arrivals in (2.16) and (2.17) are exact, whereas the formulas for non-Poisson formulas are only approximations.

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

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

Theorem 2.4.3 (*comparison of alternative HOL predictors*) *Consider the $GI/M/s$ queue with traffic intensity $\rho < 1$ in steady state. If the arrival process is Poisson or if we take the limit in* (2.19) *as the exact mean, then the steady-state MSE's are*

ordered by

$$MSE(\theta^{ar}_{HOL}(W_\infty)) < MSE(\theta^{d}_{HOL}(W_\infty)) < MSE(\theta^{sr}_{HOL}(W_\infty)). \tag{2.32}$$

Moreover,

$$\begin{aligned} & MSE(\theta^{d}_{HOL}(W_\infty)) - MSE(\theta^{ar}_{HOL}(W_\infty)) \\ = \quad & E\left[\left((1-\rho)(W_\infty|W_\infty>0) - \tfrac{(c_a^2+3)}{2s}\right)^2\right] \\ < \quad & \tfrac{(c_a^2+3)^2}{4s^2} = MSE(\theta^{sr}_{HOL}(W_\infty)) - MSE(\theta^{ar}_{HOL}(W_\infty)) \ . \end{aligned} \tag{2.33}$$

Proof. The MSE formulas in (2.33) are obtained by directly adding and subtracting the mean inside the MSE formula, with the mean here regarded as being given by (2.19). The key inequality in (2.33) follows from a bound on the mean steady-state waiting time in the $GI/M/1$ queue. The conditional delay $(W_\infty|W_\infty > 0)$ in the $GI/M/s$ model has the same exponential distribution as in the $GI/M/1$ model; e.g., see p. 398 of Wolff (1989). Its mean is $(1-\omega)^{-1}$, where ω is the root of the transform equation $\hat{f}(1-\omega) = \omega$, where $\hat{f}(s)$ is the Laplace-Stieltjes transform of the interarrival-time cdf. However, it is known that $1-\omega > 2(1-\rho)/(c_a^2+1)$; e.g., apply Theorem 2 of Whitt (1984), noting that in the $D/M/1$ queue $1-\omega > 2(1-\rho)$, which follows from elementary inequalities for the exponential function: $e^{-2(1-\rho)} \geq 1-2(1-\rho)$. From (2.33), we see that $MSE(\theta^{d}_{HOL}(W_\infty)) < MSE(\theta^{sr}_{HOL}(W_\infty))$ if and only if

$$E\left[\left((1-\rho)(W_\infty|W_\infty>0) - \frac{(c_a^2+3)}{2s}\right)^2\right] < \frac{(c_a^2+3)^2}{4s^2} \ , \tag{2.34}$$

which, upon expanding the quadratic and using the fact that the second moment is

twice the square of the first moment, holds if and only if

$$E[W_\infty | W_\infty > 0] < \frac{c_a^2 + 3}{s(1-\rho)} , \tag{2.35}$$

which is implied by the delay bound. ∎

To illustrate, we show numerical results in Tables 2.5 and 2.6 for the candidate delay predictors θ^d_{HOL}, θ^{sr}_{HOL} and θ^{ar}_{HOL} in the $H_2/M/s$ and $M/M/s$ models, respectively, with $s = 100$ and $s = 1$. We display the values of their approximate MSE's in steady state predicted by formulas (2.23), (2.22) and (2.21), and we show the contributing terms, displayed in the order given in Theorem 2.4.2. In each case, one term grows without bound as ρ increases while the other terms remains constant or nearly constant. We take the expected value of each MSE formula, where w is distributed randomly as the steady-state conditional delay $(W_\infty | W_\infty > 0)$. We use the simulation estimates of the first two moments of the conditional delay. Table 2.5 is consistent with Theorem 2.4.3. As a consequence of Theorem 2.4.3, we suggest using the asymptotic refined predictor θ^{ar}_{HOL}.

Paralleling Table 2.5, we show corresponding results for the $M/M/s$ model with $s = 100$ and $s = 1$ in Table 2.6. We have used simulation to estimate all quantities here, even though we could compute them analytically. This case thus provides a crosscheck on both our analytic formulas and the simulations.

We remark that the limit in (2.25) implies that $W_{HOL,s}(w)$ should be approximately normally distributed when sw is not too small. Our simulation experiments show that all the random variables $W_{HOL,s}(w)$, $W_{LES,s}(w)$ and $W_{RCS,s}(w)$ tend to be normally distributed when sw is not too small.

We can combine (2.25) and (2.6) to compare the efficiency of the QL and refined HOL predictors under high congestion. Let $W(t)$ be the virtual waiting time at time

Evaluating the alternative HOL predictors

Approximations in the $H_2/M/100$ model

ρ	0.88	0.92	0.96	0.98
$E[W\|W>0]$	0.1902	0.2964	0.6114	1.307
conf. int.	$\pm$0.0030	$\pm$0.0067	$\pm$0.029	$\pm$0.17
$E[W^2\|W>0]$	0.07205	0.1761	0.7446	3.436
conf. int.	$\pm$0.0022	$\pm$0.0095	$\pm$0.060	$\pm$0.67
$MSE(\theta^d)$	0.00826	0.0135	0.0293	0.0640
term 1	0.00103	0.00113	0.00119	0.00137
term 2	0.00677	0.0120	0.0276	0.0622
term 3	0.00045	0.00045	0.00045	0.00045
$MSE(\theta^{sr})$	0.00882	0.00141	0.00298	0.0645
term 1	0.00837	0.0136	0.0293	0.0640
term 2	0.00045	0.00045	0.00045	0.00045
$MSE(\theta^{ar})$	0.00759	0.0129	0.0286	0.0632
term 1	0.00837	0.0136	0.0293	0.0640
term 2	-0.000775	-0.000775	-0.000775	-0.000775

Approximations in the $H_2/M/1$ model

ρ	0.85	0.90	0.95	0.98
$E[W\|W>0]$	15.01	23.50	48.64	115.7
conf. int.	$\pm$0.18	$\pm$0.42	$\pm$1.6	$\pm$8.80
$E[W^2\|W>0]$	446.2	1105.7	4707.1	25650.5
conf. int.	$\pm$8.03	$\pm$39.2	$\pm$263.2	$\pm$3280
$MSE(\theta^d)$	62.59	104.9	230.3	565.7
term 1	10.04	11.06	11.76	10.26
term 2	48.04	89.3	214.0	550.9
term 3	4.5	4.5	4.5	4.5
$MSE(\theta^{sr})$	68.31	110.3	235.5	571.6
term 1	63.81	105.8	231.0	567.1
term 2	4.5	4.5	4.5	4.5
$MSE(\theta^{ar})$	56.06	98.02	223.3	559.3
term 1	63.81	105.8	231.0	567.1
term 2	-7.75	-7.75	-7.75	-7.75

Table 2.5: Evaluation of the MSE approximations for the predictors θ^d_{HOL}, θ^{sr}_{HOL} and θ^{ar}_{HOL} in steady-state using (2.23), (2.21) and (2.22) together with simulation estimates of the first two moments of the conditional delay $E[W_\infty|W_\infty>0]$. The $H_2/M/s$ model is considered as a function of the traffic intensity ρ for $s=100$ and $s=1$. The ASE's are measured in units of mean service time squared per customer.

Evaluating the alternative HOL predictors

Approximations in the $M/M/s$ model for $s = 100$ and $s = 1$

ρ	0.85	0.90	0.93	0.95	0.98	0.99
$E[W\|W>0]$	0.0666	0.0993	0.1435	0.2012	0.500	0.901
conf. int.	±0.0018	±0.0027	±0.0018	±0.0019	±0.037	±0.059
$E[W^2\|W>0]$	0.0089	0.0196	0.0414	0.0811	0.500	1.53
conf. int.	±0.0006	±0.0012	±0.0016	±0.0026	±0.097	±0.24
$MSE(\theta^d)$	0.00153	0.00219	0.00307	0.00422	0.01020	0.01823
term 1	0.00020	0.00020	0.00020	0.00020	0.00020	0.00015
term 2	0.00073	0.00139	0.00227	0.00342	0.00940	0.01748
term 3	0.00060	0.00060	0.00060	0.00060	0.00060	0.00060
$MSE(\theta^{sr})$	0.00173	0.00239	0.00327	0.00442	0.01040	0.01844
term 1	0.00113	0.00179	0.00267	0.00382	0.00980	0.01784
term 2	0.00060	0.00060	0.00060	0.00060	0.00060	0.00060
$MSE(\theta^{ar})$	0.00133	0.00199	0.00287	0.00402	0.01000	0.01804
term 1	0.00113	0.00179	0.00267	0.00382	0.00980	0.01784
term 2	0.00020	0.00020	0.00020	0.00020	0.00020	0.00020

Approximations in the $M/M/1$ model

ρ	0.80	0.85	0.90	0.95	0.96	0.98
$E[W\|W>0]$	5.01	6.68	9.98	20.04	24.80	50.70
conf. int.	±0.03	±0.04	±0.08	±0.36	±0.33	±2.4
$E[W^2\|W>0]$	50.3	89.6	200.3	806.6	1211	5290
conf. int.	±0.69	±1.36	±5.1	±37.4	±45	640
$MSE(\theta^d)$	12.02	15.36	21.98	42.08	51.58	103.4
term 1	2.01	2.01	2.00	2.02	1.94	2.11
term 2	4.01	7.35	13.98	34.07	43.64	95.25
term 3	6.00	6.00	6.00	6.00	6.00	6.00
$MSE(\theta^{sr})$	14.02	17.35	23.97	44.07	53.61	105.31
term 1	8.02	11.35	17.97	38.07	47.61	99.31
term 2	6.00	6.00	6.00	6.00	6.00	6.00
$MSE(\theta^{ar})$	10.02	13.35	19.97	40.07	49.61	101.31
term 1	8.02	11.35	19.97	38.07	47.61	99.31
term 2	2.00	2.00	2.00	2.00	2.00	2.00

Table 2.6: Evaluation of the MSE approximations for the predictors θ^d_{HOL}, θ^{sr}_{HOL}, and θ^{ar}_{HOL} in steady-state using (2.23), (2.21) and (2.22) together with simulation estimates of the first two moments of the conditional delay $E[W_\infty|W_\infty > 0]$. The $M/M/s$ model is considered as a function of the traffic intensity ρ for $s = 100$ and $s = 1$. The ASE's are measured in units of mean service time squared per customer.

t, the time an arrival at time t would have to wait before beginning service. Since

$$W(t) = \sum_{i=1}^{Q(t)+1} (S_i/s) \ , \tag{2.36}$$

the law of large numbers implies that $W(t)/Q(t) \to 1/s$ as $Q(t) \to \infty$. Thus, when $Q(t)$ is large, we have $W(t) \approx Q(t)/s$ (even if $W(t)$ itself is not large). Assuming that n is large with $w \approx n/s$ in (2.25) and (2.6), we have both sw and n large and

$$\frac{c^2_{W_{HOL,s}(w)}}{c^2_{W_{Q,s}(n)}} \approx \frac{(c_a^2+1)/\rho s w}{1/(n+1)} \approx \frac{c_a^2+1}{\rho} \ . \tag{2.37}$$

Since we have introduced HOL partly as an approximation for LES, it is interesting to consider the difference between the HOL and LES observed delays and the difference between the random variables $W_{HOL,s}(w)$ and $W_{LES,s}(w, d/s)$. (We let $t_a - t_e = d/s$ because it should be proportional to $1/s$ with s servers.) First note that if at least one customer remains in queue after the last customer to enter service at time t_e, then the HOL customer at time t_e (after the customer entered service) will remain the HOL customer at time t_a. As a consequence, the HOL customer arrived immediately after the LES customer. Thus the HOL customer waits more than the LES customer by the time $t_a - t_e$ but less by the single interarrival time between them. Clearly these differences should become asymptotically negligible in the appropriate scaling.

We now compare the random variables $W_{HOL,s}(w)$ and $W_{LES,s}(w, d)$. We establish a stochastic bound between these random variables. Let $\leq_{st}$ denote ordinary stochastic order; see §9.1 of Ross (1996). The following bound shows that the difference between $W_{HOL,s}(w)$ and $W_{LES,s}(w, d)$ is stochastically bounded and thus asymptotically negligible compared to w and these individual random variables as $sw \to \infty$.

We say that a family of random variables $\{X(w) : w > 0\}$ is *stochastically bounded* if for any $\epsilon > 0$ there exists a positive constant $K(\epsilon)$ such that $P(|X(w)| > K(\epsilon)) < \epsilon$. By Markov's inequality, for nonnegative random variables it suffices to have the means $E[X(w)]$ uniformly bounded: $P(|X(w)| > K(\epsilon)) \leq E[X(w)]/K(\epsilon)$.

Theorem 2.4.4 (*bound on the difference between $W_{HOL,s}(w)$ and $W_{LES,s}(w, d/s)$*) *Consider the $GI/M/s$ model. Assume that there is at least one customer in queue at the new arrival epoch, so that (2.11) is valid for HOL and (2.9) is valid for LES. Then*

$$W_{LES,s}(w, d/s) - X(s, w, d) \leq_{st} W_{HOL,s}(w) \leq_{st} W_{LES,s}(w, d/s) + X(s, w, d) \ , \tag{2.38}$$

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

$$X(s, w, d) \equiv \sum_{i=1}^{A(w+(d/s))-A(w)+1} (S_i/s) \ . \tag{2.39}$$

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

$$\frac{|W_{HOL}(w) - W_{LES}(w, d)|}{w} \to 0 \quad as \quad sw \to \infty \ . \tag{2.40}$$

For the $M/M/s$ model,

$$X(s, w, d) = \sum_{i=1}^{A(d/s)+1} (S_i/s) \ , \tag{2.41}$$

so that

$$E[X(s, w, d)] = (\rho d + 1)/s \quad and \quad Var(X(s, w, d)) = (2\rho d + 1)/s^2 \ . \tag{2.42}$$

Proof. Without altering the individual distributions of $W_{HOL,s}(w)$ and $W_{LES,s}(w, d/s)$, we can make a special construction in which we use exactly the same exponential random variables S_i/s for the two predictors. The random numbers of summands differ by $A(w + (d/s)) - A(w) - 1$, which is bounded above by $A(w + (d/s)) - A(w) + 1$, which we use in (2.39). Since the renewal process A has rate ρs, we can then apply Blackwell's renewal theorem, p. 155 of Asmussen (2003), to get $E[A(w + d/s) - A(w)] \to \rho d$ as $sw \to \infty$. Recall that we have assumed that the interarrival time cdf F is non-lattice. Hence we get $E[X(s, w, d)]/w \to 0$ as $sw \to \infty$, which implies (2.40). ∎

2.5 Simulations Related to Theorem 2.4.2

Based on (2.23) in Theorem 2.4.2, we approximate the MSE of the direct HOL, LES and RCS predictors by

$$MSE(\theta^d_{HOL}(w)) \approx (1-\rho)^2 w^2 + \frac{((2\rho-1)c_a^2 + 4\rho - 3)w}{s} + \frac{K}{s^2}\ , \tag{2.43}$$

for K in (2.24). As above, let $MSE(\theta^d_{HOL}(W_\infty))$ denote the MSE in steady state, i.e., when we replace w in (2.43) by $(W_\infty | W_\infty > 0)$. We obtain

$$\begin{aligned} MSE(\theta^d_{HOL}(W_\infty)) &\approx (1-\rho)^2 E[W_\infty^2 | W_\infty > 0] \\ &+ \frac{((2\rho-1)c_a^2 + 4\rho - 3)E[W_\infty | W_\infty > 0]}{s} + \frac{K}{s^2}\ , \end{aligned} \tag{2.44}$$

where W_∞ is the steady-state delay.

We have compared the ASE for HOL, LES and RCS to $MSE(\theta^d_{HOL}(W_\infty))$ and found close agreement, with the agreement being slightly better for HOL and LES than for

RCS. In making this comparison, we substitute the simulation estimates of the two moments $E[W_\infty|W_\infty > 0]$ and $E[W_\infty^2|W_\infty > 0]$ into (2.44). We must calculate or approximate these conditional moments in order to have a full approximation, but we do not consider that step here. We obtain good results comparing approximation (2.44) to the ASE for the cases of exponential (M), hyperexponential (H_2 with $c_a^2 = 4$) and Erlang (E_2) interrenewal-time distributions. We did experiments for $s = 1, 10, 100, 400, 900$, each for four values of ρ, increasing with s in order to represent typical cases. The errors were consistently less than 5% for HOL and LES in these experiments, as illustrated by the results for LES with M and H_2 interarrival-time distributions in Table 2.7.

Testing the MSE(HOL$_\infty$) approximations in the $GI/M/100$ model

ρ	M	% diff.	D	% diff.	H_2	% diff.
0.98	10.20	−0.3%	2.67	−1.9%	62.8	−3.9%
0.95	4.20	1.4%	1.20	−4.1%	22.9	−1.9%
0.93	3.06	0.4%	0.92	−5.8%	15.9	−2.1%
0.90	2.20	−1.5%	0.72	−7.5%	10.5	−3.2%

Table 2.7: Evaluation of the approximations for the steady-state MSE of HOL in (2.44) and (2.46) by comparing to simulation estimates of the ASE for LES in the $GI/M/100$ model as a function of the interarrival-time distribution and the traffic intensity ρ. The simulation estimates appear in Table 2.1. The approximations in units of 10^{-3} and the relative percent differences are shown here. The ASE's are measured in units of mean service time squared per customer.

We found that the approximation in (2.44) does not perform nearly as well for the case of a deterministic (D) arrival process, which should not be surprising, because the deterministic interrenewal-time distribution is a lattice distribution not covered by Theorem 2.4.2. Instead of (2.43), we propose the following approximation for the direct predictor with a D arrival process:

$$MSE(\theta^d_{HOL,D}(w)) \approx (1-\rho)^2 w^2 + \frac{\rho w + (2/s)}{s} , \tag{2.45}$$

which is obtained by making the simple approximation $A(w) \approx \rho s w$. We then obtain the following analog of the steady-state approximation (2.44):

$$MSE(\theta^d_{HOL,D}(W_\infty)) \approx (1-\rho)^2 E[W^2_\infty | W_\infty > 0] + \frac{\rho E[W_\infty | W_\infty > 0] + (2/s)}{s} . \tag{2.46}$$

Approximation (2.46) performs much better than approximation (2.44) with $c_a^2 = 0$, yielding errors of about 5% (ranging up to 11%), instead of about $5-25\%$, as shown in Table 2.7. For the refined predictor, we would also change the mean predictor to (2.16) instead of (2.19).

In order to evaluate the approximations for a specified observed delay w, we consider data from the simulation where the observed HOL delay falls in a small interval about $w \equiv 2E[W_\infty | W_\infty > 0]$. (We choose interval widths to make roughly reasonable, comparable sample sizes.) Table 2.8 shows the results of such an experiment for the $GI/M/100$ model with $\rho = 0.95$. (The width of the sampling interval in each case was chosen to have roughly comparable sample sizes.) Table 2.8 shows that the approximations for the HOL conditional mean and variance are remarkably accurate approximations for all three predictors: HOL, LES and RCS, with the variance being slightly higher for RCS. We found that the estimated distribution of the actual delay is approximately normally distributed in each case, as predicted by the limit in (2.25).

2.6 Heavy-Traffic Limits for Several Predictors

We can gain additional insight about the performance of the different predictors by considering heavy-traffic limits for the $GI/M/s$ model. To do so, we consider a family of models indexed by the parameter ρ, so we introduce a second subscript ρ

Testing the approximations (2.19) and (2.21) with observed w in a small interval about $2E[W_\infty|W_\infty > 0]$

interarrival-time dist.	M	D	H_2
$2E[W_\infty\|W_\infty > 0]$	0.40	0.20	0.96
selected HOL w interval	[0.39, 0.41]	[0.19, 0.21]	[0.94, 0.98]
sample size	128,287	99,747	151,556
sample mean observed	0.3998	0.2000	0.9597
$E[W_{HOL}](w)$ est.	0.4003	0.1996	0.9625
$Var(W_{HOL})(w)$ est.	0.0080	0.0020	0.0448
$E[W_{LES}(w)]$ est.	0.3996	0.1995	0.9617
$Var(W_{LES}(w))$ est.	0.0081	0.0021	0.0450
$E[W_{RCS}(w)]$ est.	0.3938	0.1929	0.9586
$Var(W_{RCS}(w))$ est.	0.0103	0.0029	0.0507
Predicted mean by (2.19)	0.400	0.205	0.947
Pred. variance by (2.21)	0.0076	0.0021	0.0455

Table 2.8: Comparing the approximations for $E[W_{HOL}(w)]$ and $Var(W_{HOL}(w))$ for fixed w following from (2.19) and (2.21) with simulation estimates of the mean and variance of the HOL, LES and RCS predictors in the $GI/M/100$ model with $\rho = 0.95$ as a function of the interarrival-time distribution. Data are collected for observed waiting times contained in a small interval about $2E[W_\infty|W_\infty > 0]$. The resulting sample sizes are shown. The ASE's are measured in units of mean service time squared per customer.

in addition to s. We let the service times remain unchanged. We assume that we start with interarrival times U_n having mean $1/s$. In system (s, ρ), we use interarrival times U_n/ρ, so that they have mean $1/s\rho$. That makes the traffic intensity in model ρ be ρ.

We consider both the classical heavy-traffic (HT) regime in which $\rho \uparrow 1$ for fixed s and the Quality-and-Efficiency-Driven (QED) many-server heavy-traffic (HT) regime in which both $\rho \uparrow 1$ and $s \to \infty$ with $((1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$. For more on the QED regime for $GI/G/s$ queues, see Halfin and Whitt (1981), Puhalskii and Reiman (2000), Jelenkovic et al. (2004) and Whitt (2004b, 2005). The queue length tends to be of order $1/(1-\rho)$ in both limiting regimes, but the delays behave differently. The delay are of order $1/(1-\rho)$ in the classical HT regime, but are of

order $1-\rho$ or $1/\sqrt{s}$ in the QED HT regime.

2.6.1 Insights from the Heavy-Traffic Snapshot Principle

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

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

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

$$(1-\rho)W_{s,\rho}((1-\rho)^{-2}t) \Rightarrow RBM(t) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{2.47}$$

where the limit stochastic process $RBM(t)$ is a reflected Brownian motion, which

has continuous sample paths, and the convergence in distribution is for the entire stochastic process with sample paths in the function space D; see Whitt (2002). The space scaling in (2.47) implies that the waiting times will be of order $O(1/(1-\rho))$, while the time scaling in (2.47) implies that the waiting times will only change significantly over time intervals of length of order $O(1/(1-\rho)^2)$. As a consequence, we conclude that the HOL and LES predictors are relatively consistent in the classical HT regime.

A similar story holds in the QED HT regime. The stochastic-process limit for the virtual waiting time process in the QED regime is obtained by Puhalskii and Reiman (2000). Let $W_{s,\rho}(t)$ be the virtual waiting time at time t in model (s,ρ). Paralleling (2.47), in the QED regime we have the stochastic-process limit

$$\sqrt{s}W_{s,\rho}(t) \Rightarrow Y(t) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{2.48}$$

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

The time and space scaling in (2.48) is drastically different from (2.47), but we nevertheless obtain the same conclusions about our predictors. Now the waiting times are getting small instead of large, being of order $O(1/\sqrt{s})$, but there is no time scaling at all, so that the waiting times will only change significantly over time intervals of length of order $O(1)$. As a consequence, we conclude that the HOL and LES predictors are also relatively consistent in the QED HT regime. Again, we conclude that there will be no advantage to averaging the delays experienced over past customers.

2.6.2 Steady-State Heavy-Traffic Limits

In the following, we establish heavy-traffic limits in both regimes for steady-state random variables. We focus on the HOL predictor; by Theorem 2.4.4, the LES predictor behaves the same. We see what happens "on average" to the random variable $W_{HOL,s,\rho}(w)$ (where the observed delay w has the steady-state distribution). From the steady-state HT limits, we deduce that both the direct QL and HOL predictors are (weakly) relatively consistent: the ratio of the predictor to the random quantity being estimated converges to 1. We also establish limits establishing the asymptotic efficiency of the different predictors (comparing MSE's). In these HT limits the direct and refined predictors have asymptotically the same efficiency, while the QL predictor is asymptotically more efficient than these delay-history predictors by the constant factor $c_a^2 + 1$, consistent with Theorem 2.4.2. Since associated heavy-traffic stochastic-process limits have been established for other models, the predictors should have similar nice properties for other models.

The Classical Heavy-Traffic Regime. We start with the classic heavy-traffic (HT) regime in which $\rho \uparrow 1$ with fixed s. We look at the distribution of $W_{HOL,s}(w)$, assuming that the observed waiting time w experienced by the customer at the head of the line is a random variable $W^h_{\infty,s,\rho}$, assumed to be the steady-state delay in model (s,ρ) experienced by a customer at the head of the line at an arrival epoch, conditional on there being at least one customer in the queue. Thus let $W_{HOL,s,\rho}(W^h_{\infty,s,\rho})$ denote a random variable with the distribution

$$P(W_{HOL,s,\rho}(W^h_{\infty,s,\rho}) \le x) \equiv \int_0^\infty P(W_{HOL,s,\rho}(w) \le x)\, dP(W^h_{\infty,s,\rho} \le w) \ , \quad (2.49)$$

in model (s,ρ), where in this subsection s is held fixed.

This means that $E[W_{HOL,s,\rho}(W^h_{\infty,s,\rho})] \equiv E[E[W_{HOL,s,\rho}(W^h_{\infty,s,\rho})|W^h_{\infty,s,\rho}]]$. The ran-

dom variable $W^h_{\infty,s,\rho}$ is not quite distributed as the steady-state waiting time at the arrival epoch, $W_{\infty,s,\rho}$, or the conditional steady-state waiting time, $(W_{\infty,s,\rho}|W_{\infty,s,\rho} > 0)$, but it is asymptotically equivalent to both of these in the heavy-traffic limit.

In order to relate the HOL and QL predictors, it is important to exploit the joint convergence of the steady-state queue length and waiting time. Such joint convergence is discussed extensively for the single-server queue in Chapter 9 of Whitt (2002); it was also used in Iglehart and Whitt (1970), which treated more general models. Let $(Q_{\infty,s,\rho}, W_{\infty,s,\rho})$ be a random vector with the limiting steady-state distribution of $(Q_{k,s,\rho}, W_{k,s,\rho})$, where $Q_{k,s,\rho}$ is the queue length and $W_{k,s,\rho}$ is the delay just before $A_{k,s,\rho}$, where $A_{k,s,\rho}$ is the k^{th} arrival epoch, all in model (s,ρ).

Here we will use the following established steady-state heavy-traffic limit:

$$(1-\rho)(Q_{\infty,s,\rho}, W_{\infty,s,\rho}) \Rightarrow (L, L/s) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{2.50}$$

where $L \stackrel{\text{d}}{=} Exp(c_a^2+1)/2$ with $Exp(m)$ denoting a random variable having an exponential distribution with mean m. We give a detailed proof in a subsection below starting from the known steady-state distribution for $Q_{\infty,s,\rho}$. The joint convergence follows from the limit for $Q_{\infty,s,\rho}$ and the law of large numbers, using the representation

$$(Q_{\infty,s,\rho}, W_{\infty,s,\rho}) = \left(Q_{\infty,s,\rho}, (Q_{\infty,s,\rho}+1) \left(\left[\sum_{i=1}^{Q_{\infty,s,\rho}+1} (S_i/s) \right] / (Q_{\infty,s,\rho}+1) \right) \right) . \tag{2.51}$$

We can apply (2.50) and previous results to get the following limits for our predictors. Let RMSE $\equiv MSE/Mean^2$ be the relative mean squared error. Let $c^2_{W_{Q,s,\rho}}(Q_{\infty,s,\rho})$ be the random variable assuming the value $c^2_{W_{Q,s,\rho}}(n)$ with probability $P(Q_{\infty,s,\rho}=n)$ for $n \geq 0$. Let other random variables involving c^2 and RMSE be defined analogously.

We prove the following theorem in a subsection below.

Theorem 2.6.1 (*classical heavy-traffic limit*) *If $\rho \uparrow 1$ in the family of $GI/M/s$ models indexed by (s, ρ) with fixed s, then*

$$\frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{E[W_{Q,s,\rho}(Q_{\infty,s,\rho})|Q_{\infty,s,\rho}]} = \frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{(Q_{\infty,s,\rho}+1)/s} \Rightarrow 1 \ , \tag{2.52}$$

$$\frac{W_{\infty,s,\rho}}{W^h_{\infty,s,\rho}} \Rightarrow 1 \quad \textit{and} \quad \frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}} \Rightarrow 1 \ , \tag{2.53}$$

from which we can deduce that

$$(1-\rho)(Q_{\infty,s,\rho}, W_{\infty,s,\rho}, W^h_{\infty,s,\rho}, W_{Q,s,\rho}(Q_{\infty,s,\rho}), W_{HOL,s,\rho}(W^h_{\infty,s,\rho}))$$

$$\Rightarrow (L, L/s, L/s, L/s, L/s) \tag{2.54}$$

and

$$(1-\rho)^{-1}(c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}, c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}, RMSE(W^h_{\infty,s,\rho}))$$

$$\Rightarrow (1/L, (c_a^2+1)/L, (c_a^2+1)/L) \tag{2.55}$$

where $L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2)$ as above, so that

$$\frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W_{Q,s,\rho}(Q_{\infty,s,\rho})} \Rightarrow 1, \quad \frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2+1 \ , \tag{2.56}$$

$$\frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}} \Rightarrow 1 \quad \textit{and} \quad \frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2+1 \ . \tag{2.57}$$

The limits in (2.52) and (2.53) show that the direct QL and HOL predictors are (weakly) relatively consistent in the classical heavy-traffic limit, while the limits in (2.55)–(2.57) compare the asymptotic efficiency of the different predictors. In this heavy traffic limit, the direct and refined HOL predictors have asymptotically

the same efficiency, while the QL predictor is asymptotically more efficient by the constant factor $c_a^2 + 1$.

We conjecture (but have not yet proved) that there is appropriate uniform integrability, so that the moments of these random variables converge as well as distributions, see p. 31 of Billingsley (1999). Then from (2.55) and (2.56) we obtain associated convergence of the moments:

$$E\left[\frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}}\right] \to c_a^2 + 1 \quad \text{and} \quad \frac{E[c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}]}{E[c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}]} \to c_a^2 + 1 \ , \tag{2.58}$$

and similarly for the direct predictor. These limits supplement the previous limits, implying that the QL delay predictor is asymptotically more efficient than the HOL and LES delay predictors by the constant factor $c_a^2 + 1$ in the classical heavy-traffic limit.

The QED Many-Server Heavy-Traffic Regime.

We now consider the QED HT regime, in which both $\rho \uparrow 1$ and $s \uparrow \infty$ with $(1-\rho)\sqrt{s} \to \beta$ for some positive constant β.

This alternative QED regime is appealing because, unlike the classical HT regime, the probability that a customer is delayed approaches a nondegenerate limit, strictly between 0 and 1:

$$P(W_{\infty,s,\rho} > 0) \to \alpha \quad \text{and} \quad P(Q_{\infty,s,\rho} > 0) \to \alpha, \quad 0 < \alpha < 1 \ , \tag{2.59}$$

where $\alpha \equiv \alpha(\beta/\sqrt{c_a^2+1})$ for $\alpha(x) \equiv [1 + x\Phi(x)/\phi(x)]^{-1}$, where ϕ is the cdf and ϕ is the probability density function (pdf) of the standard normal $N(0,1)$; see (1.1) of Whitt (2004b).

With minor modifications, the story is the same as for the classical HT regime, so

we will be brief. A major difference is that the queue length is of order $O(\sqrt{s}) = O(1/(1-\rho))$, while the waiting time is of order $O(1/\sqrt{s}) = O((1-\rho))$. As before, the ratio $W_{\infty,s,\rho}/Q_{\infty,s,\rho}$ is of order $O(1/s)$, but now $s \to \infty$.

Paralleling (2.50), we have the joint limit

$$\left(\frac{Q_{\infty,s,\rho}}{\sqrt{s}}, (1-\rho)Q_{\infty,s,\rho}, \sqrt{s}W_{\infty,s,\rho}, \frac{W_{\infty,s,\rho}}{1-\rho} \right) \Rightarrow (Z, \beta Z, Z, Z/\beta) \ , \tag{2.60}$$

where $P(Z > 0) = \alpha$ for the same $\alpha \equiv \alpha(\beta/\sqrt{c_a^2+1})$ defined above and $(Z|Z > 0) \stackrel{\mathrm{d}}{=} L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2)$. The limit for $Q_{\infty,s,\rho}$ was established by Halfin and Whitt (1981), but Whitt (2004b) corrects an error in the expression for α when the arrival process is non-Poisson. The joint limit with $W_{\infty,s,\rho}$ can be established as in (2.51). Paralleling (2.87), here we have

$$\begin{aligned} & \left((1-\rho)(Q_{\infty,s,\rho}|Q_{\infty,s,\rho} > 0), \frac{W_{\infty,s,\rho}|W_{\infty,s,\rho} > 0}{1-\rho}, \frac{W^h_{\infty,s,\rho}}{1-\rho}, (1-\rho)A(W^h_{\infty,s,\rho}) \right) \\ \Rightarrow \ & (\beta L, L/\beta, L/\beta, \beta L) \ , \end{aligned} \tag{2.61}$$

where again $L \stackrel{\mathrm{d}}{=} (Z|Z > 0) \stackrel{\mathrm{d}}{=} Exp(c_a^2+1)/2$; as before, the important point is that the same random variable L appears in all four components on the right.

We now state the theorem, omitting the proof.

Theorem 2.6.2 (*QED heavy-traffic limit*) *If $\rho \uparrow 1$ and $s \uparrow \infty$ so that $(1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$ in the family of $GI/M/s$ models indexed by ρ and s, then*

$$\frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{(Q_{\infty,s,\rho}+1)/s} \Rightarrow 1 \quad \textit{and} \quad \frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}} \Rightarrow 1 \ . \tag{2.62}$$

$$(1-\rho)^{-1}(W_{Q,s,\rho}(Q_{\infty,s,\rho}), W_{HOL,s,\rho}(W^h_{\infty,s,\rho})) \Rightarrow (L/\beta, L/\beta) \tag{2.63}$$

and

$$(1-\rho)^{-1}(c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}, c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}, RMSE(W^h_{\infty,\rho,s})) \Rightarrow (1/\beta L, (c_a^2+1)/\beta L, (c_a^2+1)/\beta L) \tag{2.64}$$

where $L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2)$ *as above, so that*

$$\frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W_{Q,s,\rho}(Q_{\infty,s,\rho})} \Rightarrow 1, \quad \frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2+1 \ . \tag{2.65}$$

$$\frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}} \Rightarrow 1 \quad \textit{and} \quad \frac{RMSE(W^h_{\infty,s,\rho})}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} \Rightarrow c_a^2+1 \ . \tag{2.66}$$

Just as in the classical HT regime, we conjecture that there is appropriate uniform integrability, so that the moments converge as well as distributions. Then we will obtain associated convergence of the moments, just as in (2.58).

Heavy-Traffic Detail: Proof of (2.50).

In this section we prove the classical heavy-traffic limit for the steady-state joint distribution of the queue length and waiting time at arrival epochs stated in (2.50):

$$(1-\rho)(Q_{\infty,\rho}, W_{\infty,\rho}) \Rightarrow (L, L/s) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{2.67}$$

where $L \stackrel{\mathrm{d}}{=} Exp(c_a^2+1)/2$ with $Exp(m)$ denoting a random variable that is exponentially distributed with mean m. We consider this a known result, but we cannot point to a place where a proof is given.

We draw on well-known properties of the steady-state distribution of the $GI/M/s$ queue. The key initial result is the fact that the conditional distribution of the queue length at an arrival epoch, given that the arrival must wait, is a geometric

distribution, i.e.,

$$P(Q_{\infty,\rho} = j | W_{\infty,\rho} > 0) = (1-\omega)\omega^j, \quad j \geq 0 \ , \tag{2.68}$$

where the single parameter ω in (2.68) is the unique root of the equation

$$\omega = \int_0^\infty e^{-(1-\omega)sx} \, dF(x) \equiv \hat{f}((1-\omega)s) \ , \tag{2.69}$$

where $\hat{f}$ is the Laplace-Stieltjes transform of the cdf F, i.e.,

$$\hat{f}(z) \equiv \int_0^\infty e^{-zx} \, dF(x) \ ; \tag{2.70}$$

see (14.10), (14.11), (14.12) and (14.19) of Cooper (1982). This property was used in the proof of Theorem 2.4.3.

The key then is the way that the root $\omega \equiv \omega(\rho)$ depends on the traffic intensity ρ as $\rho \uparrow 1$. Anticipating that we should have $\omega(\rho) \uparrow 1$ as $\rho \uparrow 1$, we see that the argument of the Laplace-Stieltjes transform should approach 0 in the limit. It should thus come as no surprise that we can rigorously establish the desired result by expanding the Laplace transform $\hat{f}(z)$ in a Taylor series about $z = 0$; see p. 435 of Feller (1971) for supporting theory. As was first observed by Smith (1953, p. 461), it follows that

$$\frac{1-\omega(\rho)}{1-\rho} \to \frac{2}{c_a^2 + 1} \quad \text{as} \quad \rho \uparrow 1 \ . \tag{2.71}$$

The expansion appears in a more general context in formula (17) of Abate and Whitt (1994). In the special case of the $GI/M/s$ queue, equation (7) there reduces to equation (2.69) here. An alternative approach involving upper and lower bounds is given in Whitt (1984); that focuses on the more elementary $GI/M/1$ model, but the

key root has the same structure. The equation differs only by the constant factor s appearing in the equation (2.69). Additional theoretical results about characterizing roots for queues appears in Neuts (1986), Choudhury and Whitt (1994) and Glynn and Whitt (1994).

It is well known – see pages 1-2 of Feller (1971) – that if X_m is a random variable with a geometric distribution having mean m, then

$$\frac{X_m}{cm} \Rightarrow Exp(1/c) \quad \text{as} \quad m \to \infty \ . \tag{2.72}$$

By (2.68), $(Q_{\infty,\rho}|W_{\infty,\rho} > 0)$ has a geometric distribution with mean $1/(1-\omega(\rho))$. Thus we can combine (2.68), (2.71) and (2.72) to obtain

$$(1-\rho)(Q_{\infty,\rho}|W_{\infty,\rho} > 0) \Rightarrow Exp((c_a^2+1)/2) \quad \text{as} \quad \rho \uparrow 1 \ . \tag{2.73}$$

It is also known that

$$P(W_{\infty,\rho} > 0) = \frac{A}{1-\omega} \quad \text{where} \quad A = \left[\frac{1}{1-\omega} + X\right]^{-1} \ , \tag{2.74}$$

with $X \equiv X(\rho) \to X(1)$, $0 < X(1) < \infty$, as $\rho \uparrow 1$; see (14.14)–(14.17) of Cooper (1982). Hence

$$P(W_{\infty,\rho} > 0) = [1 + (1-\omega(\rho))X(\rho)]^{-1} \to 1 \quad \text{as} \quad \rho \uparrow 1 \ . \tag{2.75}$$

Combining (2.73) and (2.75), we obtain the first part of (2.67):

$$(1-\rho)Q_{\infty,\rho} \Rightarrow L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2) \quad \text{as} \quad \rho \uparrow 1 \ . \tag{2.76}$$

Given that

$$W_{\infty,\rho} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{Q_{\infty,\rho}+1} (S_i/s) \ , \tag{2.77}$$

we have

$$\frac{W_{\infty,\rho}}{Q_{\infty,\rho}+1} \Rightarrow \frac{1}{s} \quad \text{as} \quad \rho \uparrow 1 \tag{2.78}$$

by the weak law of large numbers, since $Q_{\infty,\rho} \Rightarrow \infty$ as a consequence of (2.76). We then apply Theorem 11.4.5 of Whitt (2002) to write the joint limit

$$((1-\rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho}+1)) \Rightarrow (L, (1/s)) \ . \tag{2.79}$$

We then can apply the continuous mapping theorem with the function $h : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $h(x,y) = (x, xy)$ to get

$$h(((1-\rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho}+1)) \Rightarrow h(L, (1/s)) = (L, L/s) \ , \tag{2.80}$$

but

$$h(((1-\rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho}+1)) = \left((1-\rho)Q_{\infty,\rho}, (1-\rho)W_{\infty,\rho}\frac{Q_{\infty,\rho}}{Q_{\infty,\rho}+1}\right) \ . \tag{2.81}$$

Since $Q_{\infty,\rho} \Rightarrow \infty$,

$$\frac{Q_{\infty,\rho}}{Q_{\infty,\rho}+1} \Rightarrow 1 \quad \text{as} \quad \rho \uparrow 1 \ . \tag{2.82}$$

Hence,

$$|h(((1-\rho)Q_{\infty,\rho}, W_{\infty,\rho}/(Q_{\infty,\rho}+1)) - (1-\rho)(Q_{\infty,\rho}, W_{\infty,\rho})| \Rightarrow 0 \quad \text{as} \quad \rho \uparrow 1 \ . \tag{2.83}$$

Thus we can combine (2.80), (2.83) and the convergence-together theorem, Theorem 11.4.7 of Whitt (2002), to complete the proof of (2.67).

Proof of Theorem 2.6.1.

First we show that $W^h_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$. As a consequence of the limit in (2.50), we must have $W_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$. Suppose that we do *not* have $W^h_{\infty,s,\rho} \Rightarrow \infty$. Then there must exist a subsequence $\{\rho_k\}$ with $\rho_k \uparrow 1$ as $k \to \infty$, a constant K and a positive constant $\epsilon > 0$ such that $P(W^h_{\infty,s,\rho_k} > K) > \epsilon$ for all k. Since

$$W_{\infty,s,\rho} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{A(W^h_{\infty,s,\rho})+2} (S_i/s) \ , \tag{2.84}$$

conditional on $W_{\infty,s,\rho} > 0$, which holds with probability 1 in the limit, there must exist a new constant K' such that $P(W_{\infty,s,\rho_k} > K') > \epsilon/2$ for all k as well, but that contradicts the established limit $W_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$. Hence we must have $W^h_{\infty,s,\rho} \Rightarrow \infty$ as $\rho \uparrow 1$, as claimed above.

Given that $\rho \uparrow 1$ and $W^h_{\infty,s,\rho} \Rightarrow \infty$, we get $A(W^h_{\infty,s,\rho})/W^h_{\infty,s,\rho} \Rightarrow s$ and

$$\frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}} = \left(\frac{\sum_{i=1}^{A(W^h_{\infty,s,\rho})+2}(S_i/s)}{A(W^h_{\infty,s,\rho})+2}\right)\left(\frac{A(W^h_{\infty,s,\rho})+2}{W^h_{\infty,s,\rho}}\right) \Rightarrow (1/s) \times s = 1 \ , \tag{2.85}$$

by the law of large numbers for partial sums and renewal processes. Similarly, by (2.50), we also have $Q_{\infty,s,\rho} \Rightarrow \infty$, so that

$$\frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{Q_{\infty,s,\rho}+1} = \frac{\sum_{i=1}^{Q_{\infty,s,\rho}+1}(S_i/s)}{Q_{\infty,s,\rho}+1} \Rightarrow 1/s \ . \tag{2.86}$$

The limits (2.85) and (2.86) imply (2.52) and (2.53).

Since the limits in (2.85) and (2.86) are deterministic, we can apply Theorem 11.4.5

of Whitt (2002) to obtain joint convergence of all these with the limits in (2.50):

$$\left((1-\rho)Q_{\infty,s,\rho}, (1-\rho)W_{\infty,s,\rho}, (1-\rho)W^h_{\infty,s,\rho}, \frac{W_{Q,s,\rho}(Q_{\infty,s,\rho})}{Q_{\infty,s,\rho}+1}, \frac{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}{W^h_{\infty,s,\rho}}\right)$$
$$\Rightarrow \left(L, \frac{L}{s}, \frac{L}{s}, \frac{1}{s}, 1\right) \ . \tag{2.87}$$

We next apply the continuous mapping theorem, see Section 3.4 of Whitt (2002), with the function $h : \mathbb{R}^5 \to \mathbb{R}^5$ defined by $h(v, w, x, y, z) = (v, w, x, vy, xz)$ to get (2.54) from (2.87).

To continue, we next consider the random variable $c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}$. Starting from the limit in (2.54), we can apply the Skorohod representation theorem, Theorem 3.2.2 on p. 78 of Whitt (2002), to get random variables $\tilde{W}^h_{\infty,s,\rho}$ with the same probability law as $W^h_{\infty,s,\rho}$ but for which we have the convergence $(1-\rho)\tilde{W}^h_{\infty,s,\rho} \to \tilde{L}/s$ as $\rho \uparrow 1$ w.p.1, where $\tilde{L} \stackrel{\mathrm{d}}{=} L \stackrel{\mathrm{d}}{=} Exp((c_a^2+1)/2)$. Next note that $c^2_{W_{HOL,s,\rho}(w)}/c^2_{W_{HOL,s,1}(w)} \to 1$ w.p.1 as $\rho \uparrow 1$ and $w \to \infty$ in any order. Then, by (2.25),

$$\frac{c^2_{W_{HOL,s,\rho}(\tilde{W}^h_{\infty,s,\rho})}}{1-\rho} = \left(\frac{c^2_{W_{HOL,s,\rho}(\tilde{W}^h_{\infty,s,\rho})}}{c^2_{W_{HOL,s,1}(\tilde{W}^h_{\infty,s,\rho})}}\right)\left(\frac{\tilde{W}^h_{\infty,s,\rho}\, c^2_{W_{HOL,s,1}(\tilde{W}^h_{\infty,s,\rho})}}{(1-\rho)\tilde{W}^h_{\infty,s,\rho}}\right) \to \frac{(c_a^2+1)/s}{\tilde{L}/s} \tag{2.88}$$

as $\rho \uparrow 1$ w.p.1. Essentially the same reasoning applies to the random variable RMSE $(W^h_{\infty,s,\rho})$, giving the same limit. The equality in distribution then implies the associated convergence in distribution for the last two components of the original random vector in (2.55). We now treat the first component. Since $(Q_{\infty,s,\rho} +$

$1)c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})} = 1$, a deterministic quantity, by (2.5), we can apply (2.25) to get

$$\begin{aligned}\frac{c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}} &= \left(\frac{Q_{\infty,s,\rho}+1}{W^h_{\infty,s,\rho}}\right)\left(\frac{W^h_{\infty,s,\rho}c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})}}{(Q_{\infty,s,\rho}+1)c^2_{W_{Q,s,\rho}(Q_{\infty,s,\rho})}}\right) \\ &= \left(\frac{Q_{\infty,s,\rho}+1}{W^h_{\infty,s,\rho}}\right)W^h_{\infty,s,\rho}c^2_{W_{HOL,s,\rho}(W^h_{\infty,s,\rho})} \\ &\Rightarrow s \times \frac{c_a^2+1}{s} = c_a^2+1\ . \end{aligned} \tag{2.89}$$

We then reason as before in establishing (2.87), first to express this limit jointly with the last two components of (2.55) and then to apply the continuous mapping theorem to complete the proof of (2.55) itself. Finally, (2.56) and (2.57) follow from the previous results. ■

2.6.3 Customers Who Have Completed Service

In this final subsection, supplementing the application of the snapshot principle above, we consider the predictors based on the delays experienced by previous customers to *complete* service. Unlike for the LES and HOL predictors, we find that the LCS predictor behaves very differently in the classical and QED HT regimes. The way to see this is to observe that the LCS customer completed service a full service time in the past. That LCS customer arrived a waiting time plus a service time in the past.

In both heavy-traffic regimes, the service time is an exponential random variable with mean 1. In the classical HT regime, the waiting times are exploding in heavy traffic, so that a service time is negligible compared to the waiting time. Thus we see that LCS will be asymptotically equivalent to LES and HOL in the classical HT regime, for any fixed number of servers. The LCS predictor will be consistent as

well in the classical heavy-traffic regime.

However, the story is very different in the QED HT regime. The service times remain unchanged, but now the waiting times become smaller, being of order $O(1/\sqrt{s})$. Now the service time is the same order as the time scaling. The stochastic-process limit in (2.48) describes the waiting time experience of each customer, but for the last customer to complete service at time t, we have a different limit. Let $A^L_{s,\rho}(t)$ denote the arrival time of the last customer to complete service at time t in model (s,ρ). The relevant limit now will be

$$\sqrt{s}W_{s,\rho}(A^L_{s,\rho}(t)) \Rightarrow Y(t-S) \quad \text{as} \quad \rho \uparrow 1 \ , \tag{2.90}$$

where $Y(t)$ is the limit process in (2.48) and S is a service time, an exponential random variable with mean 1. In other words, the waiting time at time t is approximately $Y(t)/\sqrt{s}$, while the waiting time of the last customer to complete service immediately prior to time t is approximately $Y(t-S)/\sqrt{s}$. Thus, in the QED HT limit the LCS predictor is *not* consistent. The effectiveness of the LCS predictor depends on the difference between $Y(t-S)$ and $Y(t)$. However, we do not attempt to do further analysis; here we are content to observe that the LCS predictor has inferior asymptotic performance in the QED HT regime. That is consistent with our simulation results, which show that the LCS predictor performs poorly for large s.

Fortunately, there is better information that we can obtain from customers who have already completed service in the QED HT regime. Other customers who have completed service are very likely to have arrived much more recently than the last customer to complete service. The minimum service time among the last m customers to complete service is $1/m$. Since the waiting times are of order $1/\sqrt{s}$,

it is natural to consider $m = O(\sqrt{s})$; then the minimum service time among these customers also will be of order $O(1/\sqrt{s})$.

As a bound, first consider the customer among the last $c\sqrt{s}$ customers to complete service with the minimum service time. That customer's service time is exponentially distributed with mean $1/c\sqrt{s} = O(1/\sqrt{s})$. By (2.48), the customer's waiting time is also of order $O(1/\sqrt{s})$. Since the times between successive service completions are i.i.d. exponential random variables with mean $1/s$, the last $c\sqrt{s}$ service completions occur over a time interval having mean $c/\sqrt{s} = O(1/\sqrt{s})$. Hence this customer arrived $O(1/\sqrt{s})$ in the past. Hence we deduce that if we consider the customer among the last $c\sqrt{s}$ customers to complete service with the minimum service time, then that delay predictor is consistent in the QED HT regime.

Even better will be the RCS and RCS-$c\sqrt{s}$ predictors, because those customers necessarily arrive at least as recently. We summarize these conclusions in the following theorem. To state the theorem, let $W^{RCS}_{\infty,s,\rho}$ and $W^{RCS-c\sqrt{s}}_{\infty,s,\rho}$ be the steady-state RCS and RCS-$c\sqrt{s}$ delays in model (s,ρ); and let $W_{RCS,s,\rho}(w)$ and $W_{RCS-c\sqrt{s},s,\rho}(w)$ be the associated random variables having the conditional distribution of the delay to be estimated given the observed RCS and RCS-$c\sqrt{s}$ delays.

Theorem 2.6.3 (*performance of LCS, RCS and RCS-$c\sqrt{s}$ in the QED HT regime*) *If $\rho \uparrow 1$ and $s \uparrow \infty$ so that $(1-\rho)\sqrt{s} \to \beta$ for $0 < \beta < \infty$ in the family of $GI/M/s$ models indexed by s and ρ, then the RCS and RCS-$c\sqrt{s}$ predictors are relatively consistent, i.e.,*

$$\frac{W_{RCS,s,\rho}(W^{RCS}_{\infty,s,\rho})}{W^{RCS}_{\infty,s,\rho}} \Rightarrow 1 \quad \textit{and} \quad \frac{W_{RCS-c\sqrt{s},s,\rho}(W^{RCS-c\sqrt{s}}_{\infty,s,\rho})}{W^{RCS-c\sqrt{s}}_{\infty,s,\rho}} \Rightarrow 1 \ , \tag{2.91}$$

but the LCS predictor is not relatively consistent.

In this relatively crude sense, the predictors LES, HOL, RCS and RCS-$c\sqrt{s}$ are all asymptotically equivalent in the QED regime, but LCS is not. However, it remains to describe the asymptotic efficiency of RCS and RCS-$c\sqrt{s}$, paralleling the results for the HOL (and LES) predictor SCV's in (2.64) and (2.65).

2.7 Concluding Remarks

Insights that can be Generalized. Even though we are primarily interested in service systems that are more complex than the $GI/M/s$ queueing model, in this chapter we studied the performance of alternative delay predictors in this relatively simple idealized $GI/M/s$ setting. Our goal has been to gain insight into how the predictors will perform in more complex settings. Our results for the $GI/M/s$ model indicate what to expect more generally.

Performance of the Predictors. An important reference point for the delay predictors based on delay history is the standard QL predictor based on the observed queue length, defined in (2.2). For QL, the only source of uncertainty is the remaining service times of the customers ahead of the arrival. That uncertainty can be reduced if the remaining service times can be reliably estimated, as emphasized by Whitt (1999a).

As can be seen from formulas (2.9)-(2.11), to a large extent, the LES and HOL predictors can be regarded as the QL predictor modified by replacing the known queue length by an estimate of that queue length. Since the queue length is equal (or approximately equal) to the number of arrivals during the observed waiting time, the queue length is estimated by the expected number of arrivals during the observed waiting time. Thus the increase in MSE in going from QL to the LES, HOL and

RCS predictors is primarily due to variability in the arrival process. The MSE tends to be larger for LES and HOL than QL by the constant factor $(c_a^2 + 1)$, where c_a^2 is the SCV of an interarrival time, a common measure of variability for a renewal arrival process; see Whitt (1982a).

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

Overall, we conclude that the greatest source of prediction uncertainty is the remaining service times. After that, it is the arrival-process variability, as partially characterized by the SCV c_a^2. We conclude that the predictors $\theta_{QL}(n)$, $\theta_{LES}^d(w)$, $\theta_{HOL}^d(w)$ and $\theta_{RCS}^d(w)$ can be very useful, but they are not extraordinarily accurate. The refined predictors for HOL, LES and RCS can remove all or nearly all of the bias, but non-negligible variance remains. The greatest hope for more reliable prediction seems to lie in being able to better predict the remaining service times, which is certainly possible if the service times are actively *controlled*, and is possible to some extent if either the service-time distribution is non-exponential or if it is possible to classify the customers, as discussed in Whitt (1999a).

We considered several different delay predictors based on recent delay history, notably LES, HOL and RCS. Through analysis and extensive simulation experiments, we conclude that the LES and HOL delay predictors are very similar, with both being more accurate than the others based on delay history, but less accurate than the

full-information queue-length (QL) predictor. For large s, RCS is far superior to the delay of the last customer to complete service (LCS), because customers need not complete service in the same order they arrive. For low traffic intensities with large s, LCS was even outperformed by the no-information predictor (NI). The reason is that the LCS customer may have arrived too long ago. We conclude that RCS should only be preferred to HOL and LES if delay information is not available until after customers complete service, but the MSE is not much greater for RCS than for LES and HOL.

In §2.6 we established heavy-traffic limits that provide important insight. The heavy-traffic snapshot principle provides strong support for all these delay-history estimation procedures, and shows that there should be little benefit from averaging over past customer delays, under heavy loads. The relative errors of the LES and HOL predictors are asymptotically negligible in both the classical and many-server heavy-traffic regimes. The MSE relative to the mean is asymptotically negligible for all the candidate delay predictors based on delay history The QL predictor is asymptotically more efficient than HOL and LES by the constant factor $c_a^2 + 1$ in both heavy-traffic regimes. Since similar heavy-traffic limits have already been established for much more general models, these heavy-traffic properties can be expected to hold more generally.

3

Delay Prediction in the $GI/GI/s + GI$ Model

3.1 Introduction

In this chapter, we use heavy-traffic limits and computer simulation to study the performance of alternative real-time delay predictors in the overloaded $GI/GI/s + GI$ multiserver queueing model, allowing customer abandonment. Our main contributions are: (i) to propose new, effective, and simple ways to do better delay prediction in overloaded many-server queues with customer abandonment, (ii) to establish heavy-traffic limits that generate approximations for the expected mean squared error (MSE) of some delay predictors, and (iii) to describe results of simulation experiments evaluating alternative delay predictors. We obtain more effective delay predictors by exploiting approximations for performance measures in many-server queues with a non-exponential abandonment-time distribution, from Whitt (2005b, 2006). This chapter is an edited version of Ibrahim and Whitt (2009b).

For completeness, we redefine relevant notation and restate some related results from chapter 2.

3.1.1 The $GI/GI/s + GI$ Model

In this chapter, we consider the steady-state behavior of an overloaded $GI/GI/s+GI$ queueing model. This model has independent and identically distributed (i.i.d.) interarrival times with mean λ^{-1} and a general distribution. We only use the i.i.d. assumption for the interarrival times when simulating the model; it is not required for the implementation of our delay predictors. Service times are i.i.d. with mean μ^{-1} and a general distribution. Each arriving customer will abandon if he is unable to start service before a random time with mean ν^{-1} and a general distribution. Abandonments times are i.i.d.; the arrival, service and abandonment processes are all mutually independent. There is unlimited waiting space and arriving customers are served in order of arrival; i.e., we use the first-come-first-served (FCFS) service discipline. The traffic intensity is $\rho \equiv \lambda/s\mu$.

We focus on overloaded scenarios, in which the arrival rate exceeds the maximum possible total service rate. Customer abandonment makes the system stable in this case. (That can be proved by bounding the model above by the $GI/GI/\infty$ model obtained by removing all servers; then the abandonment times can be thought of as service times. For more on the stability of the $GI/GI/\infty$ model, see p.178 of Whitt (1982b).) We consider overloaded systems because we are primarily interested in predicting delays when they are large. For example, many call centers are overloaded at least some of the time, especially service-oriented ones in which emphasis is placed on efficiency rather than on quality of service.

3.1.2 Potential Waiting Times

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

3.1.3 Quantifying Performance: Average Squared Error (ASE)

As in chapter 2, we rely on simulation to evaluate the alternative delay predictors. In our simulation experiments, we quantify the performance of a delay predictor by computing the average squared error (ASE), defined by:

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

where $w_i > 0$ is the *potential* waiting time of delayed customer i, p_i is the delay prediction corresponding to customer i, and k is the number of customers in our sample. (Here, we define the ASE slightly differently than in the $GI/M/s$ model where there is no distinction between actual and potential waiting times; see (3.1).) In our simulation experiments, we measure w_i for both served and abandoning cus-

tomers. 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. The ASE should approximate the expected MSE in steady state.

3.1.4 Mean Squared Error (MSE)

Paralleling (2.5), let $W_Q(n)$ represent a random variable with the conditional distribution of the *potential* delay of an arriving customer, given that this customer must wait before starting service, and given that the queue length at the time of his arrival, t, not counting himself, is $Q(t) = n$. In this framework, the event "$Q(t) = 0$" corresponds to all servers being busy and our arriving customer being the first in queue. Let $\theta_{QL}(n)$ be some given single-number delay prediction which is based on the queue length, n. Then, the MSE of the corresponding delay predictor is given by:

$$MSE \equiv MSE(\theta_{QL}(n)) \equiv E[(W_Q(n) - \theta_{QL}(n))^2] \ .$$

The MSE of a queue-length-based delay predictor is a function of n, the number of customers seen in queue upon arrival. By looking at the ASE, we are looking at the expected MSE averaging over all n, where the arrival must wait, in steady state. As explained in chapter 2, it is known that the conditional mean, $E[W_Q(n)]$, minimizes the MSE. In the $GI/M/s$ model, we can easily calculate $E[W_Q(n)]$; see (2.6). In contrast, it is difficult to find a closed-form expression for this mean in the $GI/GI/s+GI$ model, so we develop approximations of it.

3.1.5 Root Relative Squared Error

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

$$RRASE \equiv \frac{\sqrt{ASE}}{(1/k)\sum_{i=1}^{k} p_i} , \tag{3.2}$$

using the same notation as in (3.1). The denominator in (3.2) is the average potential waiting time of customers who must wait. For large samples, the RRASE should agree with the expected *root relative mean squared error* (RRMSE), in steady state. The RRASE and RRMSE are useful because they measure the effectiveness of an predictor relative to the mean, so that they are easy to interpret.

3.1.6 Organization

The rest of this chapter is organized as follows: In §3.2, we describe a no-information delay predictor (NI) in the efficiency-driven many-server heavy-traffic limiting regime, which serves as a useful reference point. In §3.3, we define new queue-length-based delay predictors, and discuss relevant results. For completeness, we briefly describe alternative delay-history-based delay predictors in §3.4. (For a more complete description, see chapters 1 and 2.) We establish heavy-traffic limits for several delay predictors in the $G/M/s + M$ model in §3.5, and present simulation results for the $M/M/s + GI$ model in §3.6. In §3.7, we present additional experimental results for non-exponential service-time distributions. In §3.8, we present simulation results substantiating the heavy-traffic limits of §3.5 by considering the $GI/M/s + M$ model with alternative interarrival-time distributions and alternative values of the abandonment rate ν. We make concluding remarks in §3.10.

3.2 A Theoretical Reference Point

An important theoretical reference is the many-server heavy-traffic limit for the number in the system in the Markovian $M/M/s+M$ queue with customer abandonment, in the efficiency-driven (ED) regime, as discussed in Garnett et al. (2002), Whitt (2004a) and Talreja and Whitt (2008). That limit describes how the model behaves as the arrival rate λ and number of servers s increase, while the individual service rate μ and individual abandonment rate ν remain unchanged, with the traffic intensity held fixed at a value $\rho \equiv \lambda/s\mu > 1$. (There are also some results for the more general $G/GI/s+GI$ model in the ED regime in Zeltyn and Mandelbaum (2005) and Whitt (2006).)

Let $W_s(\infty)$ represent the steady-state waiting time as a function of s in the ED regime, and let $\Rightarrow$ denote convergence in distribution. Whitt (2004a) shows that

$$W_s(\infty) \Rightarrow w \equiv \frac{1}{\nu} \ln(\rho) > 0 \quad \text{as} \quad s \to \infty \ , \tag{3.3}$$

while Theorem 6.1 of Zeltyn and Mandelbaum (2005) (Theorem 5 below) and Theorem 6.4 of Talreja and Whitt (2008) show that

$$\sqrt{s}\,(W_s(\infty) - w) \Rightarrow N(0, 1/\nu\mu) \quad \text{as} \quad s \to \infty, \tag{3.4}$$

where $N(m, \sigma^2)$ denotes a normal random variable with mean m and variance σ^2. These limits lead to the deterministic fluid approximation $W_s(\infty) \approx w$ and the stochastic refinement $W_s(\infty) \approx N(w, 1/s\nu\mu)$.

The deterministic fluid approximation w in (3.3) and the steady-state mean $E[W_s(\infty)]$ it approximates, are candidate *no-information* (NI) predictors, θ_{NI}, paralleling the NI predictor for the $GI/M/s$ model considered as a reference point in chapter 2

(e.g., see §2.2.1). In fact, the NI predictor is much more appealing now, because it is much more effective with customer abandonment than without. Based on the limits above (plus appropriate uniform integrability, which can also be established), we have

$$MSE(\theta_{NI}) \approx Var(W_s(\infty)) \approx \frac{1}{s\nu\mu} \to 0 \quad \text{as} \quad s \to \infty. \tag{3.5}$$

Unlike in the $GI/M/s$ model, here the squared coefficient of variation (SCV, variance divided by the square of the mean), c^2_{NI}, is asymptotically negligible as well, because here $E[W_s(\infty)] \to w > 0$ as $s \to \infty$. For the $GI/M/s$ model, $c^2_{NI} \to 1$ as $\rho \uparrow 1$ for all s. The limit in (3.5) implies that any reasonable predictor ought to be effective in the ED regime as s gets larger. We will want to see that our proposed predictors outperform NI as well as become effective as s increases.

3.3 Queue-Length-Based Delay predictors

In this section, we describe alternative predictors based on the queue length seen upon arrival to the system. The information needed for the implementation of each of these queue-length-based predictors is summarized in Table 3.1. For completeness, we begin by reviewing the QL predictor which was extensively studied in chapter 2.

3.3.1 The Simple Queue-Length-Based (QL) Delay Predictor

For a system having s agents, 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 $(n+1)/s\mu$ minutes. Let QL refer to this simple queue-length-based predictor, commonly used in practice. Let

	Information About the Model
QL	Q(t), s, μ
QL_r^m	Q(t), s, μ, ν
QL_r	Q(t), s, μ, $F(x)$, λ
QL_m	Q(t), s, μ, ν
QL_a	Q(t), s, μ, $F(x)$, λ

Table 3.1: Summary of the information required for the implementation of each queue-length delay predictor.

the predictor, as a function of n, be:

$$\theta_{QL}(n) \equiv (n+1)/s\mu \ . \tag{3.6}$$

The QL predictor is appealing due to its simplicity and ease of implementation: It uses information about the system that usually is readily available. In the $GI/M/s$ model, $W_Q(n)$ is the time necessary to have exactly $n+1$ consecutive departures from service (service completions). But, the times between successive service completions, when all servers are busy, are i.i.d. random variables distributed as the minimum of s exponential random variables, each with mean μ^{-1}, which makes them i.i.d. exponential with mean $1/s\mu$; see (2.5). The optimal delay predictor, using the MSE criterion, is the one announcing the conditional mean, $E[W_Q(n)]$. But, following the analysis above, $E[W_Q(n)] = \theta_{QL}(n)$ in (3.6). Hence, QL is optimal for the $GI/M/s$ model, under the MSE criterion. Extensive simulation experiments in chapter 2 show the superiority of QL in that simple idealized setting.

When there is customer abandonment, the QL predictor overestimates the potential delay, because customers in queue may abandon before entering service, and QL fails to take that into account. That is confirmed by our simulation results in §3.6, but we now analytically quantify the effect for the Markovian $M/M/s+M$ model. To

do so, we use the steady-state fluid approximations to the $M/M/s+M$ model in the ED regime discussed in §3.2. In the steady-state fluid limit, all served customers wait the same deterministic amount of time w in (3.3) and they all see the same number of customers, q, in queue upon arrival. From (2.26) of Whitt (2004a),

$$q = \frac{s\mu}{\nu}(\rho - 1). \tag{3.7}$$

In the fluid limit,

$$\theta_{QL}(q) = \frac{q+1}{s\mu} \approx \frac{q}{s\mu} = \frac{1}{\nu}(\rho - 1) > w = \frac{1}{\nu}\ln(\rho) \ .$$

Consistent with intuition, we see that QL overestimates w. Indeed,

$$\frac{\theta_{QL}(q) - w}{w} = \frac{(\rho - 1)/\nu - \ln(\rho)/\nu}{\ln(\rho)/\nu} \ ; \tag{3.8}$$

e.g., there is 10% relative error when $\rho = 1.2$, 19% relative error when $\rho = 1.4$, and much greater error when ρ is larger. (Exploiting the asymptotic expansion of the logarithm: $\ln(1+\delta) \approx \delta - \delta^2/2$ when δ is small, we can obtain the simple rough approximation to (3.8) of $(\rho - 1)/(3 - \rho) \approx (\rho - 1)/2$ when ρ is slightly greater than 1.)

Motivated by the simple form of the QL delay prediction, $\theta_{QL}(n)$ in (3.6), we now propose modified queue-length-based delay predictors that account for customer abandonment, and that are also easy to implement in practice.

3.3.2 The Markovian Queue-Length-Based Delay Predictor (QL_m)

As in Whitt (1999a), this predictor QL_m approximates the $GI/GI/s+GI$ model by the corresponding $GI/M/s+M$ model with the same service-time and abandon-time means. For the $GI/M/s+M$ model, we have the representation:

$$W_Q(n) \equiv \sum_{i=0}^{n} Y_i \ , \tag{3.9}$$

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

$$E[W_Q(n)] = \sum_{i=0}^{n} E[Y_i] = \sum_{i=0}^{n} \frac{1}{s\mu+i\nu} \ . \tag{3.10}$$

The QL_m predictor given to a customer who finds n customers in queue upon arrival is $\theta_{QL_m}(n) \equiv E[W_Q(n)]$. Under the MSE criterion, QL_m is the best possible, in the $GI/M/s+M$ model, but we find that it is not always so good for the more general $GI/GI/s+GI$ model. Non-exponential service-time and abandonment-time distributions are commonly observed in practice; see Brown et al. (2005), and Mandelbaum and Zeltyn (2004, 2007). It is therefore important to propose other queue-length-based delay predictors that effectively cope with non-exponential distributions. Approximations are needed because direct mathematical analysis is difficult. Next, we propose two queue-length-based delay predictors, QL_r and QL_a, exploiting approximations for performance measures in many-server queues with a

non-exponential abandonment-time distribution, developed in Whitt (2005b, 2006).

3.3.3 The Simple-Refined Queue-Length-Based Delay Predictor ($\mathbf{QL}_r$)

We now propose a simple refinement of QL by making use of the steady-state fluid approximations to the general $G/GI/s+GI$ model, in the ED limiting regime, as developed by Whitt (2006). For that purpose, let F be the cumulative distribution function (cdf) of the abandon-time distribution, and let F^c be the complementary cdf (ccdf) associated with F. (That is, $F^c(t) = 1 - F(t)$, for all t.) In this steady-state fluid limit, the deterministic waiting time w and the deterministic queue length q are given by Equations (3.6) and (3.7) of Whitt (2006), which we restate. Since "rate in" $\equiv \lambda F^c(w) = s\mu \equiv$ "rate out", we have:

$$\rho F^c(w) = 1 \ . \tag{3.11}$$

The associated equation for q is

$$q = \lambda \int_0^w F^c(x)dx = s\rho\mu \int_0^w F^c(x)dx \ . \tag{3.12}$$

In the fluid limit, QL estimates a customer's delay as the deterministic quantity:

$$\theta_{QL}(q) = \frac{q+1}{s\mu} \approx \frac{q}{s\mu} = \rho \int_0^w F^c(x)dx \ .$$

For QL_r, we propose computing the ratio $\beta = w/(q/s\mu) = ws\mu/q$ (after solving numerically for w and q), and using it to refine the QL predictor. That is, the new

delay prediction is:

$$\theta_{QL_r}(n) \equiv \beta \times \theta_{QL}(n) = \beta(n+1)/s\mu \ .$$

The QL_r predictor is appealing because it is only a minor modification of the QL predictor, but performs much better in models with customer abandonment, as we show in §3.6. In particular, it is remarkably effective with non-exponential abandonment-time distributions. Note that in addition to s, n and μ, we need to know ρ or, equivalently, λ, and the abandonment-time cdf F in order to implement QL_r.

3.3.4 The Exponential Abandonment Case (QL_r^m)

We now propose a modification of QL_r which does not depend on ρ. It is based on assuming that the abandonment-time cdf F is exponential. Using the corresponding values of w and q for the $GI/M/s+M$ model, given respectively by (3.3) and (3.7), we obtain the ratio $\beta = \ln(\rho)/(\rho-1)$. From (3.7), we get $\rho = 1+\nu q/s\mu$, yielding

$$\beta = \frac{\ln(1+\nu q/s\mu)}{\nu q/s\mu} \ .$$

The corresponding delay prediction, as a function of n, is given by

$$\theta_{QL_r^m}(n) = \beta \times \theta_{QL}(n) = \frac{\ln(1+\nu n/s\mu)}{\nu n/s\mu} \times \frac{n+1}{s\mu} \ .$$

Thus, the implementation of QL_r^m requires knowledge of n, s, μ, and ν, but not of ρ or, equivalently, λ. It approximates the abandonment-time distribution by the exponential distribution. We will see that QL_r^m performs nearly the same as QL_m, which is good when the abandonment is nearly exponential, but not necessarily

otherwise.

3.3.5 The Approximation-Based Queue-Length Delay Predictor (QL_a)

Our most promising predictor QL_a draws on the approximations in Whitt (2005b): It approximates the $GI/GI/s + GI$ model by the corresponding $GI/M/s + M(n)$ model, with state-dependent Markovian abandonment rates.

We begin by describing the Markovian approximation for abandonments, as in §3 of Whitt (2005b). As an approximation, we assume that a customer who is jth from the *end* of the queue has an exponential abandonment time with rate ν_j, where ν_j is given by

$$\nu_j \equiv h(j/\lambda), \quad 1 \leq j \leq k \ ; \tag{3.13}$$

k is the current queue length, and h is the abandonment-time hazard-rate function, defined as $h(t) \equiv f(t)/F^c(t)$, $t \geq 0$, where f is the corresponding density function (assumed to exist). Having ν_j depend on h instead of F is convenient, because it is natural to prediction F via h; e.g., see Brown et al. (2005). From (3.13), we see that the predictor QL_a depends on the abandonment distribution having a relatively smooth density. We assume that is the case.

We now explain the derivation of (3.13). 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 $h(t)$. The goal is to produce, as an approximation, abandonment rates that depend on a customer's position in queue, and on the length of that queue. We therefore need to prediction the elapsed waiting time of that customer,

given the available state information. To that end, assume that the queue length at an arbitrary time is k, and consider the customer, C_j, who is jth from the end of the line, $1 \leq j \leq k$. If there were no abandonments, then there would have been exactly $j-1$ arrival events since C_j arrived. Assuming that abandonments are relatively rare compared to service completions, a reasonable prediction is that there have been j arrival events since C_j arrived. Since a simple rough prediction for the time between successive arrival events is the reciprocal of the arrival rate, $1/\lambda$, the elapsed waiting time of C_j is approximated by j/λ and his abandonment rate by (3.13). The associated total abandonment rate from the queue in that system state is $\delta_k = \sum_{j=1}^{k} \nu_j = \sum_{j=1}^{k} h(j/\lambda)$, $k \geq 1$, and $\delta_0 \equiv 0$.

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

$$W_Q(n) \approx \sum_{i=0}^{n} X_i \ , \tag{3.14}$$

where X_{n-i} is the time between the ith and $(i+1)$st departure events. There is no difficulty for the first departure: X_n is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and n exponential random variables with rates ν_j, $1 \leq j \leq n$, (corresponding to the abandonment times of the remaining customers waiting in line). That is, X_n has an exponential distribution with rate $s\mu + \sum_{j=1}^{n} \nu_j = s\mu + \delta_n$.

The distribution of the remaining X_i's is more complicated. Since individual customers have different abandonment rates which, in our framework, depend on how long these customers have been waiting in line, we need to consider the dynamics of the system over time to determine, after each departure, who are the remaining

customers and what are their individual abandonment rates (in order to compute the resulting total abandonment rate). To simplify matters, we propose a further approximation, which is a slight modification of the argument in §7 of Whitt (2005b).

Here is what we do: As a further approximation, we assume that successive departure events are either service completions, or abandonments from the head of the line. We also assume that an prediction of the time between successive departures is $1/\lambda$. As a result of these extra assumptions, we approximate the X_i's in (3.14) by exponential random variables. Let X_{n-l}, which is the time between the lth and $(l+1)$st departure events, have an exponential distribution with rate $s\mu + \delta_n - \delta_l$. This is appropriate because it is the minimum of s exponential random variables with rate μ (corresponding to the remaining service times of customers in service), and $n-l$ exponential random variables with rates ν_i, $l+1 \le i \le n$ (corresponding to the abandonment times of the customers waiting in line).

The QL_a delay predictor given to a customer who finds n customers in queue upon arrival is

$$\theta_{QL_a}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}} \ . \tag{3.15}$$

Since QL_a coincides with QL_m in the $GI/GI/s+M$ model, it is the optimal delay predictor in the $GI/M/s+M$ model under the MSE criterion. But, in contrast to QL_m, this new queue-length-based predictor also performs remarkably well in the general $GI/GI/s+GI$ model. The simulation experiments of §3.6 suggest that QL_a is uniformly superior to all other delay predictors, in all models considered.

We emphasize that all queue-length-based predictors apply equally well to steady-state and transient settings. They differ in the amount of information that their implementation requires. It is significant that QL, QL_m, and QL_r^m are all independent

of the arrival process: For these three predictors, the arrival process can be arbitrary, even non-stationary. The QL_r and QL_a predictors require knowledge of the arrival rate λ, which requires some degree of stationarity. (There should not be too much variation over time.)

3.4 Candidate Delay-History-Based Delay predictors

In this section, we briefly describe alternative delay predictors based on recent customer delay history in the system. For a more detailed description, including performance approximations and refinements, see chapters 1 and 2. We emphasize that delay-history-based predictors apply directly to more complex settings, such as models including customer response to delay announcements.

3.4.1 The Last-To-Enter-Service (LES) Delay Predictor

As in Armony et al. (2008), a candidate delay predictor based on recent customer delay history is the delay of the last customer to have entered service, prior to our customer's arrival. That is, letting w_L be the delay of the last customer to have entered service, the corresponding LES delay prediction is: $\theta_{LES}(w_L) \equiv w_L$. As discussed in Whitt (1999a), the possibility of making reliable delay estimations is enhanced by exploiting information about the current state of the system. Thus we anticipate that queue-length-based predictors should be more effective than LES. Nevertheless, simulation experiments in §3.6 show that LES is relatively accurate in all models considered.

3.4.2 Other Delay-History-Based Delay predictors

We can consider alternative delay-history-based predictors, in addition to LES. Closely related 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.

Another alternative delay predictor is the delay of the last customer to have completed service, LCS. We naturally would want to consider this alternative predictor if we only learn customer delay experience after service is completed. That might be the case for customers and outside observers. Under some circumstances, the LCS and LES predictors will be similar, but they typically are very different when s is large, because the last customer to complete service may have experienced his waiting time much before the last customer to enter service, since customers need not depart in order of arrival.

Thus, we are led to propose other candidate delay predictors based on the delay experience of customers that have already completed service. RCS is the delay experienced by the customer that arrived most recently (and thus entered service most recently) among those customers who have already completed service. We found that RCS is far superior to LCS when s is large.

Through analysis and extensive simulation experiments, we conclude that the LES and HOL predictors are very similar, with both being slightly more accurate than RCS and much more accurate than LCS. Here, we only discuss LES.

3.5 Heavy-Traffic Limits for Several Predictors in the $G/M/s + M$ Model

Since we are considering overloaded systems with $\rho > 1$, it is natural to develop analytical approximations for the mean-squared errors of our predictors by considering stochastic-process limits in the ED many-server heavy-traffic limiting regime, as specified in §3.2. As before, we add a subscript s to indicate the dependence upon s and then let $s \to \infty$.

In this section we establish several limits for the $G/M/s + M$ model in the ED regime. Throughout this section we assume that the arrival process satisfies a functional central limit theorem (FCLT): Let $A_s(t)$ count the number of arrivals in the interval $[0, t]$ in model s. We assume that $A_s(t) \equiv A(st)$ for some given arrival process A with arrival rate λ. Let $\bar{A}_s(t) = A_s(t)/s \equiv A(st)/s$ for $t \geq 0$. Let $D \equiv D([0, \infty), \mathbb{R})$ be the function space of all right continuous real-valued functions with left limits, endowed with the usual Skorohod (J_1) topology; e.g., see Whitt (2002). We assume that A satisfies a functional weak law of large numbers (FWLLN) and a FCLT refinement:

$$\bar{A}_s(t) \Rightarrow \lambda t \quad \text{in} \quad D \quad \text{and} \quad \sqrt{s}(\bar{A}_s(t) - \lambda t) \Rightarrow \sqrt{\lambda c_a^2} B(t) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \tag{3.16}$$

where B is a standard Brownian motion. That condition will be satisfied if A is a renewal process with an interarrival-time distribution having finite first and second moments. As usual, the arrival process affects the limits for the other random quantities (the predictors) only via the two normalization constants λ and c_a^2. When A is a renewal counting process, c_a^2 is the SCV of an interarrival time.

We start by considering the Markovian predictor QL_m, which is the best possible

predictor for the $G/M/s+M$ model, under the MSE criterion. It does not depend on the arrival process. Recall that the waiting time for an arrival that finds n customers in queue upon arrival is given by (3.9). We will apply the following lemma, which is Lemma 6.1 of Talreja and Whitt (2008).

Lemma 3.5.1 *For the $G/M/s+M$ model in the ED many-server heavy-traffic regime,*

$$E[W_{Q,s}(\lfloor st \rfloor)] \to c(t), \quad sVar(W_{Q,s}(\lfloor st \rfloor) \to d(t) \tag{3.17}$$

and

$$\hat{W}_{Q,s}(t) \equiv \sqrt{s}\,(W_{Q,s}(\lfloor st \rfloor) - c(t)) \Rightarrow B(d(t)) \quad in \quad D \quad as \quad s \to \infty, \tag{3.18}$$

where B is a standard Brownian motion, while c and d are the deterministic real-valued functions

$$c(t) \equiv \frac{1}{\nu} \ln\left(1 + \frac{\nu t}{\mu}\right) \quad and \quad d(t) \equiv \frac{t}{\mu(\mu + \nu t)}. \tag{3.19}$$

As a consequence of the stochastic-process limit in (3.18), we obtain the one-dimensional limit

$$\sqrt{s}\,(W_{Q,s}(\lfloor st \rfloor) - c(t)) \Rightarrow N(0, d(t)) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad s \to \infty \quad \text{for each} \quad t. \tag{3.20}$$

As a further consequence, we obtain the following result for the best-possible predictors $\theta_{QL_{m,s}}(n)$. We use a random time change by the fluid limit

$$\bar{Q}_s(\infty) \equiv \frac{Q_s(\infty)}{s} \Rightarrow q \equiv \frac{\lambda - \mu}{\nu} \quad \text{as} \quad s \to \infty, \tag{3.21}$$

from Theorem 2.3 of Whitt (2004a) or Theorem 6.1 of Talreja and Whitt (2008).

Theorem 3.5.1 *For the $G/M/s+M$ model in the ED many-server heavy-traffic regime,*

$$sMSE(\theta_{QL_{m,s}}(\lfloor st \rfloor)) \equiv sVar(W_{Q,s}(\lfloor st \rfloor)) \to d(t) \quad as \quad s \to \infty \tag{3.22}$$

for each $t > 0$, where $d(t)$ is given in (3.19) and

$$sMSE(\theta_{QL_{m,s}}(Q_s(\infty))) \equiv sVar(W_{Q,s}(Q_s(\infty)) \Rightarrow d(q) \equiv \frac{q}{\lambda\mu} \equiv \frac{\lambda-\mu}{\lambda\mu\nu} \quad as \quad s \to \infty. \tag{3.23}$$

As a consequence (after establishing appropriate uniform integrability to get convergence of moments from convergence in distribution, which is not difficult at this point), we get associated convergence of moments from the convergence in distribution in (3.23), i.e.,

$$sE[MSE(\theta_{QL_{m,s}}(Q_s(\infty)))] \to d(q) \quad \text{as} \quad s \to \infty. \tag{3.24}$$

From either (3.23) or (3.24), we get the approximation

$$E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))] \approx \frac{\lambda-\mu}{s\lambda\mu\nu}. \tag{3.25}$$

Note that the FCLT normalization constant c_a^2 does not appear in (3.23)–(3.25). Other predictors that do not exploit knowledge of the queue length will fare worse, largely according to c_a^2. First, we can apply an extension of Theorem 6.4 of Talreja and Whitt (2008) to describe the asymptotic behavior of the no-information predictor $W_s(\infty)$. We extend the result for the $M/M/s+M$ model to the $G/M/s+M$

model, which is not difficult, reasoning as in §7.3 of Pang et al. (2007). First, we can extend Theorem 6.1 of Talreja and Whitt (2008) in that way to get an ED stochastic-process limit for the queue-length process in the $G/M/s+M$ model, getting an Ornstein-Uhlenbeck diffusion-process limit with infinitesimal mean $\mu(x) = -\nu x$ and an infinitesimal variance $\sigma^2(x) = \lambda(c_a^2+1)$, which in turn leads to a limit for the steady-state queue lengths. We then apply that result to get a generalization of the limit for the steady-state waiting time in Theorem 6.4 of Talreja and Whitt (2008).

Theorem 3.5.2 *For the $G/M/s+M$ model in the ED many-server heavy-traffic regime,*

$$\hat{Q}_s(\infty) \equiv \sqrt{s}(\bar{Q}_s(\infty) - q) \Rightarrow N\left(0, \frac{\lambda(c_a^2+1)}{2\nu}\right) \quad as \quad s \to \infty \tag{3.26}$$

and

$$\hat{W}_s(\infty) \equiv \sqrt{s}(W_s(\infty) - w) \Rightarrow N\left(0, \sigma_w^2\right) \quad as \quad s \to \infty, \tag{3.27}$$

where $\sigma_w^2 \equiv 1/\nu\mu + (c_a^2-1)/2\lambda\nu$, with w in (3.3) and q in (3.21).

Note that the variance terms in Theorem 3.5.2 simplify when $c_a^2 = 1$. We immediately obtain the limit for the MSE of the no-information (NI) predictor, assuming appropriate uniform integrability. The no-information predictor can be either the mean steady-state waiting time $E[W_s(\infty)]$ or the fluid limit w, because of the fluid limit in (3.3).

Corollary 3.5.1 *In the setting of Theorem 3.5.2, assuming necessary uniform integrability,*

$$sMSE(\theta_{NI,s}) \equiv sVar(W_s(\infty)) \to \frac{1}{\nu\mu} + \frac{c_a^2-1}{2\lambda\nu} \quad as \quad s \to \infty. \tag{3.28}$$

Combining the limits in (3.23) and (3.28), we obtain the following

Corollary 3.5.2 *In the setting of Theorem 3.5.2, assuming necessary uniform integrability,*

$$\frac{MSE(\theta_{NI,s})}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty))]} \to \frac{2\lambda+\mu(c_a^2-1)}{2(\lambda-\mu)} > 1 \quad as \quad s\to\infty. \tag{3.29}$$

We now establish corresponding results for the delay-history-based predictor LES. We exploit the fact that we can represent $W_{LES}(w_L)$ in terms of the random variable $W_{QL_{m,s}}(n)$ in (3.9) and a net-input process $N_s \equiv \{N_s(t) : t \geq 0\}$ over the interval $[0, w_L]$, i.e.,

$$W_{LES,s}(w_L) \approx W_{Q,s}(N_s(w_L)) \equiv \sum_{i=0}^{N_s(w_L)} X_{s,i}, \tag{3.30}$$

where $N_s(w_L)$ counts the number of arrivals in the interval $[0, w_L]$ who do not abandon, in system s. Formula (3.30) is not an exact relation because it does not account for the state change since the last customer entered service and the conditioning on both the LES customer and the new arrival epochs, but that change is clearly asymptotically negligible in the ED many-server limiting regime.

It is significant that the net-input stochastic process N_s has the structure of the number in system in a $G/M/\infty$ infinite-server system, starting out empty, with arrival rate $\lambda_s \equiv \lambda s$ and individual service rate equal to our abandonment rate ν. The Markovian $M/M/\infty$ special case is very well studied; e.g., see Eick et al. (1993a, b). In particular, it is well known that $N_s(t)$ has a Poisson distribution for each s and t with

$$E[N_s(t)] = \frac{s\lambda}{\nu}\left(1-e^{-\nu t}\right), \quad t \geq 0. \tag{3.31}$$

The heavy-traffic limit for more general infinite-server models, starting out empty,

was established by Borovkov (1967), as reviewed on p. 176 of Whitt (1982b).

Theorem 3.5.3 (Borovkov 1967) *For the $G/M/\infty$ models under consideration, with arrival rate $\lambda_s = \lambda s$ and service rate ν,*

$$\bar{N}_s(t) \equiv \frac{N_s(t)}{s} \Rightarrow a(t) \equiv \frac{\lambda}{\nu}\left(1 - e^{-\nu t}\right) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty \tag{3.32}$$

and

$$\hat{N}_s(t) \equiv \sqrt{s}(\bar{N}_s(t) - a(t)) \Rightarrow \hat{G}(t) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \tag{3.33}$$

where $\hat{G} \equiv \{\hat{G}(t) : t \geq 0\}$ is a Gaussian stochastic process with

$$\hat{G}(t) \stackrel{\mathrm{d}}{=} N(0, \sigma_n^2(t)) \quad \text{where} \quad \sigma_n^2(t) \equiv a(t) + \frac{\lambda(c_a^2 - 1)}{2\nu}\left(1 - e^{-2\nu t}\right), \tag{3.34}$$

for $a(t)$ defined in (3.32) and c_a^2 in (3.16).

We apply Theorem 3.5.3 to establish the following results for LES. To go beyond the $M/M/s+M$ model to treat the more general $G/M/s+M$ model, we add an extra assumption here. We assume that the limits for $\hat{N}_s$ in (3.33) and for $\hat{W}_s(\infty)$ in (3.27) hold jointly with independent limits. That holds automatically if the arrival process has independent increments (which is covered by the M case), because the evolution of N_s occurs after the arrival of the customer with the observed LES waiting time $W_s(\infty)$. For renewal processes, that joint convergence with independent limits should also hold because the interarrival times are i.i.d. and the arrivals are very fast. We add this condition to the general FCLT assumed in (3.16).

Theorem 3.5.4 *For the $G/M/s+M$ model in the ED many-server limiting regime (assuming the extra assumption immediately above and the necessary uniform inte-*

grability for the moment convergence), *as* $s \to \infty$,

$$\begin{aligned} \theta_{LES,s}(W_s(\infty)) \equiv W_s(\infty) &\Rightarrow w \equiv \frac{1}{\nu}\left(\ln\left(\frac{\lambda}{\mu}\right)\right), \\ \hat{W}_{LES,s}(W_s(\infty)) \equiv \sqrt{s}\,(W_{LES,s}(W_s(\infty)) - W_s(\infty)) &\Rightarrow N(0,\sigma^2_{LES}), \qquad (3.35) \\ sE[MSE(\theta_{LES,s}(W_s(\infty))] &\to \sigma^2_{LES}, \qquad (3.36) \end{aligned}$$

where

$$\sigma^2_{LES} \equiv d(a(w)) + \frac{\sigma^2_n(w)}{\lambda^2} + \left(\frac{\lambda-\mu}{\lambda}\right)^2 \sigma^2_w = 2d(q) + \frac{(c^2_a-1)(\lambda-\mu)}{\nu\lambda^2}, \quad (3.37)$$

for σ^2_w *in Theorem 2,* $\sigma^2_n(t)$ *in* (3.34), $a(w) = q$ *and* $d(q) = q/\lambda\mu$.

Proof of Theorem 3.5.4. We now prove the convergence in distribution in (3.35). The proof follows the general approach used to prove Theorem 6.4 of Talreja and Whitt (2008), exploiting stochastic-process limits in order to obtain the desired one-dimensional limit in $\mathbb{R}$. As in (6.37) of Talreja and Whitt (2008), we use the continuous mapping theorem with the composition map to treat random time changes. We start with the joint convergence

$$(\hat{W}_{Q,s}(t), \hat{N}_s(t), \hat{W}_s(\infty)) \Rightarrow (B(d(t)), \hat{G}(t), N(0,\sigma^2_w)) \quad \text{in} \quad D^2 \times \mathbb{R} \qquad (3.38)$$

for the processes defined in (3.18), (3.33) and (3.27), where the limits are mutually independent.

For the $M/M/s+M$ model, we can obtain the joint convergence from the individual limits established above, because we can regard the component processes on the left as mutually independent. That requires some comment, however. First in time we have the waiting time for the last customer to enter service, which is distributed asymptotically the same as $W_s(\infty)$. Then we have the buildup of the

queue behind this customer until this customer starts service, given by $\hat{N}_s(t)$. Finally, we have the remaining times between successive departures after the new arrival enters the system, as given by $\hat{W}_{Q,s}(t)$, which involves independent exponential random variables. These are mutually independent with reference to the designated arrival at one fixed time, for whom we are doing the prediction. The processes are well defined as independent random elements of D, but they only correctly apply to describe our system at a single time, as stated in the final one-dimensional limit in (3.35). (In the case of the $G/M/s+M$ model, we assume that the joint limit of $(\hat{N}_s(t), \hat{W}_s(\infty))$ is the same as if these were independent.)

Assuming the limit in (3.38), since $\bar{N}_s$ converges to a deterministic limit, we can append the limit for $\bar{N}_s$ to get

$$(\hat{W}_{Q,s}(t), \hat{N}_s(t), \bar{N}_s(t), \hat{W}_s(\infty)) \Rightarrow (B(d(t)), \hat{G}(t), a(t), N(0, \sigma_w^2)) \quad \text{in} \quad D^3 \times \mathbb{R}. \tag{3.39}$$

We can now apply the continuous mapping theorem with composition to perform a random time change with $\bar{N}_s$ to obtain the limit

$$\hat{W}_{Q,s}(\bar{N}_s(t)) \equiv \sqrt{s}\left(W_{Q,s}(N_s(t)) - c(\bar{N}_s(t))\right) \Rightarrow B(d(a(t))) \quad \text{in} \quad D \quad \text{as} \quad s \to \infty, \tag{3.40}$$

jointly with the limit in (3.39), where B is the given standard Brownian motion and $a(t)$ is defined in (3.32). We can now apply a random-time-change argument one more time with $W_s(\infty)$ to obtain the limit

$$\begin{aligned} \hat{Z}_s &\equiv \sqrt{s}\left(W_{Q,s}(N_s(W_s(\infty))) - c(\bar{N}_s(W_s(\infty)))\right) \\ \Rightarrow B(d(a(w))) &\stackrel{\mathrm{d}}{=} N(0, d(a(w))) \quad \text{in} \quad \mathbb{R} \quad \text{as } s \to \infty, \end{aligned} \tag{3.41}$$

again jointly with the limit in (3.39), where again the limit involves the same Brownian motion B.

We obtain the desired limit in (3.35) by writing

$$\hat{W}_{LES,s}(W_s(\infty)) \equiv \sqrt{s}\,(W_{LES,s}(W_s(\infty)) - W_s(\infty)) \equiv \hat{Z}_s + \hat{Y}_s \tag{3.42}$$

for $\hat{Z}_s$ in (3.41) and

$$\hat{Y}_s \equiv \sqrt{s}\,\left(c(\bar{N}_s(W_s(\infty))) - W_s(\infty)\right) \tag{3.43}$$

and establishing a limit for $\hat{Y}_s$ in (3.43) within the framework of the initial limits in (3.39). In order to make a connection to the given limits for $(\hat{N}_s(t), \hat{W}_s(\infty))$ in (3.39), we exploit a Taylor series expansion for the functions $c(t)$ and $a(t)$ in (3.19) and (3.32). Note that

$$c(q) = w \equiv \frac{1}{\nu}\ln(\rho), \quad a(w) = q \equiv \frac{\lambda-\mu}{\nu} \quad \text{and} \quad d(q) = \frac{q}{\lambda\mu}. \tag{3.44}$$

Hence, $d(a(w)) = d(q) = q/(\lambda\mu)$.

We write

$$\hat{Y}_s \equiv \sqrt{s}\,\left(c(\bar{N}_s(W_s(\infty))) - W_s(\infty)\right) \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3}, \tag{3.45}$$

where

$$\begin{aligned}
\hat{Y}_{s,1} &\equiv \sqrt{s}\,\left(c(\bar{N}_s(W_s(\infty)) - c(a(W_s(\infty)))\right), \\
\hat{Y}_{s,2} &\equiv \sqrt{s}\,(c(a(W_s(\infty))) - c(a(w))), \\
\hat{Y}_{s,3} &\equiv \sqrt{s}\,(c(a(w)) - W_s(\infty)),
\end{aligned} \tag{3.46}$$

Using a Taylor series expansion of c, we see that

$$\hat{Y}_{s,1} - c'(a(w))\sqrt{s}\left(\bar{N}_s(W_s(\infty)) - a(W_s(\infty))\right) \Rightarrow 0, \tag{3.47}$$

where $c'(a(w)) = 1/\lambda$. By Theorem 3.5.3,

$$\hat{Y}_{s,1} \Rightarrow \frac{1}{\lambda}\hat{G}(w) \stackrel{\mathrm{d}}{=} N(0, \sigma_n^2(w)/\lambda^2) \quad \text{as} \quad s \to \infty. \tag{3.48}$$

Using a Taylor series expansion of $c \circ a$, noting that $a'(w) = \mu$, we get

$$\hat{Y}_{s,2} - c'(a(w))a'(w)\sqrt{s}\left(W_s(\infty) - w\right)) \Rightarrow 0, \tag{3.49}$$

so that, by Theorem 3.5.2,

$$\hat{Y}_{s,2} \Rightarrow \frac{\mu}{\lambda}N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty. \tag{3.50}$$

Similarly, using the relation $c(a(w)) = c(q) = w$ and replacing $c(a(w))$ by w, we get

$$\hat{Y}_{s,3} - \sqrt{s}\left(w - W_s(\infty)\right) \Rightarrow 0, \tag{3.51}$$

so that, by Theorem 3.5.2 again,

$$\hat{Y}_{s,3} \Rightarrow N(0, \sigma_w^2) \quad \text{as} \quad s \to \infty, \tag{3.52}$$

where the limiting random variables $N(0, \sigma_w^2)$ in (3.50) and (3.52) are identical. By these constructions, we obtain convergence of the vector $(\hat{Y}_{s,1}, \hat{Y}_{s,2}, \hat{Y}_{s,3})$ jointly with the initial limits in (3.39) and thus also jointly with $\hat{Z}_s$ in (3.41). The processes $\hat{Y}_{s,i}$ are each asymptotically equivalent to processes that are simple functions of the processes in the original limit (3.39).

Hence,

$$\hat{Y}_s \equiv \hat{Y}_{s,1} + \hat{Y}_{s,2} + \hat{Y}_{s,3} \Rightarrow N\left(0, \frac{\sigma_n^2(w)}{\lambda^2} + \frac{(\lambda-\mu)^2\sigma_w^2}{\lambda^2}\right). \tag{3.53}$$

We can thus obtain the limit from (3.41)–(3.43), (3.45), (3.46) and (3.53) by adding the normal components.

Corollary 3.5.3 *Consider the setting of Theorem 3.5.4. For the $M/M/s+M$ model,*

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty)))]}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty)))]} \to 2 \quad as \quad s \to \infty. \tag{3.54}$$

For the $D/M/s+M$ model,

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty)))]}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty)))]} \to (2-\rho^{-1}) > 1 \quad as \quad s \to \infty. \tag{3.55}$$

For the more general $G/M/s+M$ model,

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty)))]}{E[MSE(\theta_{QL_{m,s}}(Q_s(\infty)))]} \to r(LES, QL_m) \quad as \quad s \to \infty, \tag{3.56}$$

where

$$r(LES, QL_m) = 2 \quad (\geq 2 \quad or \quad \leq 2) \quad \text{if and only if} \quad c_a^2 = 1 \quad (\geq 1 \quad or \quad \leq 1).$$

From (3.55), we see that QL_m is only slightly better than LES in the $D/M/s+M$ model when $\rho \equiv \lambda/\mu$ is only slightly greater than 1. Combining the MSE ratio limits in Theorems 3.5.1 and 3.5.4, we obtain

Corollary 3.5.4 *For the $M/M/s+M$ model in the ED many-server limiting regime,*

$$\frac{E[MSE(\theta_{LES,s}(W_s(\infty)))]}{MSE(\theta_{NI,s})} \to \frac{2(\rho-1)}{\rho}, \tag{3.57}$$

so that LES is asymptotically more (*less*) *efficient than NI if* $\rho < 2$ ($\rho > 2$).

We conclude this section by stating a CLT for the steady-state waiting time, and thus the NI delay predictor, in the $M/M/s+GI$ model in the ED regime, which is Theorem 6.1 (e) of Zeltyn and Mandelbaum (2005).

Theorem 3.5.5 (Zeltyn and Mandelbaum 2005) *For the* $M/M/s+GI$ *model in the ED regime, where the abandonment-time cdf* F *has density* f, $W_s(\infty) \Rightarrow w$ *for* w *in* (3.11) *and*

$$\sqrt{s}(W_s(\infty) - w) \Rightarrow N(0, 1/\lambda f(w)) \quad as \quad s \to \infty. \tag{3.58}$$

3.6 Simulation Results for the $M/M/s+GI$ Model

In this section, we present simulation results quantifying the performance of the alternative queue-length-based delay predictors of §3.3, and of the LES delay predictor, with exponential and non-exponential abandonment-time distributions; i.e., we consider the $M/M/s+GI$ model. For the abandonment-time distribution, we consider M (exponential), H_2 distribution (hyperexponential with SCV equal to 4 and balanced means), and E_{10} (Erlang, sum of 10 exponentials) distributions. We consider H_2 (E_{10}) to consider high (low) variability distributions relative to M.

3.6.1 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 do this without loss of generality, since we are free to choose the

time units in our system, and this assumption amounts to measuring time in units of mean service time. We also let the abandonment rate, ν, be equal to 1 because that seems to be a representative value. We consider $\nu = 0.2$ and $\nu = 5.0$ in §3.8. We vary λ to get a fixed value of ρ, for alternative values of s. We let $\rho = 1.4$ in all models. This value is chosen to let our systems be significantly overloaded. Because of abandonment, the congestion is not extraordinarily high. For example, with $s = 100$ servers and exponential abandonments, the mean queue length is about $q \approx (\rho - 1)s/\nu \approx 40$, while the average potential waiting time is about $w \approx q/s\mu \approx 0.4/\mu$ (less than half a mean service time).

Our simulation are steady-state simulations. The simulation results are based on 10 independent replications of 5 million events each, where an event is either a service completion, an arrival event, or an abandonment from the system. In this section we show plots of simulation estimates (see Figures 3.1-3.6) in addition to tables with corresponding 95% confidence intervals (see Tables 3.2-3.4).

3.6.2 Results for the $M/M/s+M$ model

In this model, QL_a coincides with QL_m. Therefore, we do not include separate results for QL_a. Consistent with theory in §3.3, Figure 3.1 shows that QL_m is the best possible, under the MSE criterion. The RRASE for QL_m ranges from about 14% for $s = 100$ to about 4% when $s = 1000$. We see that the accuracy of this predictor improves as the number of servers increases. Figure 3.2 shows that $s \times \mathrm{ASE}(\mathrm{QL}_m)$, the ASE of QL_m multiplied by the number of servers s, is nearly constant for all values of s considered. In particular, Figure 3.2 shows that $s \times \mathrm{ASE}(\mathrm{QL}_m) \approx (\lambda - \mu)/(\lambda\mu\nu)$, as in (3.25) of §3.5. The relative error between the simulation estimates for $\mathrm{ASE}(\mathrm{QL}_m)$ and the numerical value given by (3.25) is

less than 1% throughout.

Note that all predictors are relatively accurate for this model, with the exception of QL. For example, the RRASE of LES ranges from about 22% for $s = 100$ to about 7% for $s = 1000$. Also, Figure 3.2 shows that $s\times$ASE(LES) $\approx \sigma^2_{LES}$, consistent with (3.36). Indeed, the relative error between the simulation estimates and the numerical value given by (3.37) is less than 1% throughout. Figure 3.2 shows that $s\times$ASE(NI) $\approx 1/\nu\mu$, as in (3.28), with $c_a^2 = 1$. The relative error between the simulation estimates for ASE(NI) and the numerical value given by (3.28) is less than 2% throughout. Finally, Figure 3.2 shows that $s \times$ ASE(QL) is monotone increasing in s.

Table 3.2 shows that, consistent with (3.54), the LES predictor performs worse than QL_m, but not greatly so: The relative error between the simulation estimates for ASE(LES)/ASE(QL_m) and the numerical value, 2, given by (3.54) is less than 1% throughout. It is important to note that this is consistent with the results in chapter 2 for the $GI/M/s$ model; e.g., see (2.37) and simulation results in §2.3. Consistent with (3.29), the NI predictor is less efficient than QL_m: The relative error between the simulation estimates for ASE(NI)/ASE(QL_m) and the numerical value, 3.5, given by (3.29) is less than 1% throughout. The NI predictor also performs worse than LES: The ratio ASE(NI)/ASE(LES) is close to 1.75 throughout. The relative error between the simulation estimates for ASE(NI)/ASE(LES) and the numerical value given by (3.57) is less than 2% throughout. Finally, the QL predictor performs significantly worse than all the other predictors, particularly for large values of s. The ratio ASE(QL)/ASE(QL_m) ranges from about 3 when $s = 100$ to nearly 16 when $s = 1000$.

The QL_r^m predictor is nearly identical to QL_m. This can be easily explained: When the number seen in queue upon arrival, n, is large, $\theta_{QL_m}(n)$ can be approximated by

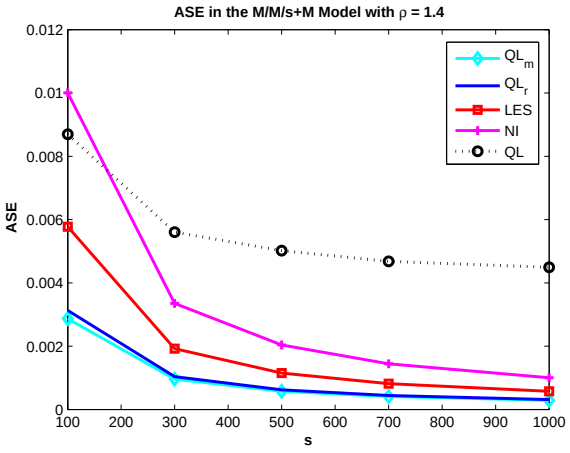

Figure 3.1: ASE of the predictors in the $M/M/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$.

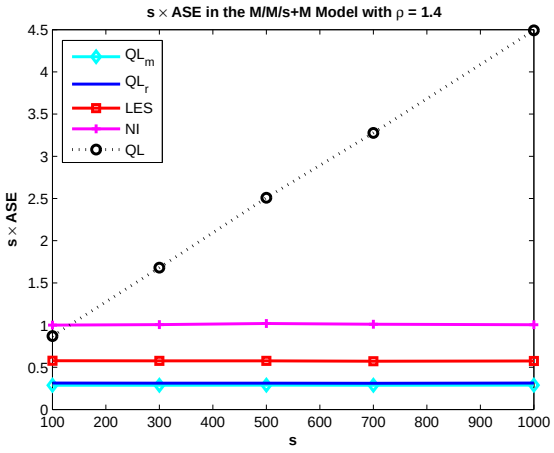

Figure 3.2: $s \times$ ASE of the predictors in the $M/M/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$.

an integral (limit of the Riemann sum)

$$\theta_{QL_m}(n) \approx \int_0^n \frac{1}{s\mu + \nu x} dx = \ln(s\mu + \nu n) - \ln(s\mu) = \frac{1}{\nu} \ln(1 + \nu n/s\mu) \ .$$

On the other hand, we have that

$$\theta_{QL_r^m}(n) \equiv \left[\ln\left(\frac{\nu n}{s\mu} + 1\right) / \left(\frac{\nu n}{s\mu}\right) \right] \times \frac{n+1}{s\mu} \approx \frac{1}{\nu} \ln(1 + \nu n/s\mu) \ .$$

So that, for large n, the two predictors QL_m and QL_r^m should perform nearly the same.

3.6.3 Results for the $M/M/s + H_2$ model

Figure 3.3 and Table 3.3 show that the best delay predictor for this model is QL_a. The corresponding RRASE ranges from about 20% for $s = 100$ to about 6% when $s = 1000$. Once more, we see that the accuracy of this predictor improves as the number of servers increases. The QL_r predictor performs nearly the same as QL_a,

Efficiency of the predictors in the $M/M/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r^m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.867×10^{-3}	2.869×10^{-3}	3.130×10^{-3}	8.693×10^{-3}	5.772×10^{-3}	1.00×10^{-2}
	$\pm1.76\times10^{-5}$	$\pm1.78\times10^{-5}$	$\pm1.89\times10^{-5}$	$\pm3.20\times10^{-5}$	$\pm2.79\times10^{-5}$	$\pm5.97\times10^{-5}$
300	9.587×10^{-4}	9.601×10^{-4}	1.039×10^{-3}	5.602×10^{-3}	1.922×10^{-3}	3.351×10^{-3}
	$\pm6.86\times10^{-6}$	6.92×10^{-6}	$\pm6.41\times10^{-6}$	$\pm2.64\times10^{-5}$	$\pm1.50\times10^{-5}$	$\pm6.03\times10^{-5}$
500	5.761×10^{-4}	5.661×10^{-4}	6.224×10^{-4}	5.017×10^{-3}	1.153×10^{-3}	2.038×10^{-3}
	$\pm1.94\times10^{-6}$	$\pm3.86\times10^{-6}$	$\pm2.94\times10^{-6}$	$\pm2.41\times10^{-5}$	$\pm9.99\times10^{-6}$	$\pm2.26\times10^{-5}$
700	4.104×10^{-4}	4.201×10^{-4}	4.440×10^{-4}	4.682×10^{-3}	8.166×10^{-4}	1.441×10^{-3}
	$\pm1.82\times10^{-6}$	2.839×10^{-4}	$\pm2.71\times10^{-6}$	$\pm2.40\times10^{-5}$	$\pm5.78\times10^{-6}$	$\pm1.57\times10^{-5}$
1000	2.892×10^{-4}	2.839×10^{-4}	3.136×10^{-4}	4.492×10^{-3}	5.752×10^{-4}	1.019×10^{-3}
	$\pm3.48\times10^{-6}$	$\pm3.86\times10^{-6}$	$\pm3.09\times10^{-6}$	$\pm1.54\times10^{-5}$	$\pm6.91\times10^{-6}$	$\pm3.00\times10^{-5}$

Table 3.2: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/M/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

and is only slightly outperformed. The ratio ASE(QL_r)/ASE(QL_a) is close to 1 for all values of s. The QL_m predictor performs well but it is now slightly outperformed by QL_r. The two are nearly the same when $s = 100$; the ratio ASE(QL_m)/ASE(QL_r) is close to 1 when $s = 100$ but closer to 2 when $s = 1000$. The RRASE for QL_m ranges from about 20% when $s = 100$ to about 8% when $s = 1000$.

The LES predictor performs worse than QL_a, QL_m, and QL_r when $s = 100$ but nearly the same as QL_m when $s = 1000$. The RRASE of the LES predictor ranges from about 30% when $s = 100$ to about 10% when $s = 1000$. The ratio ASE(LES)/ASE(QL_{ap}) is close to 2 for all values of s, suggesting that our analytical results of §3.5 should extend to general abandonment-time distributions.

The NI predictor performs worse than LES but not as bad as QL. The ratio ASE(NI)/ ASE(QL_a) is close to 3.5 for all values of s considered. As above, the efficiency of QL is degrading as the number of servers increases. The ratio ASE(QL)/ASE(QL_a)

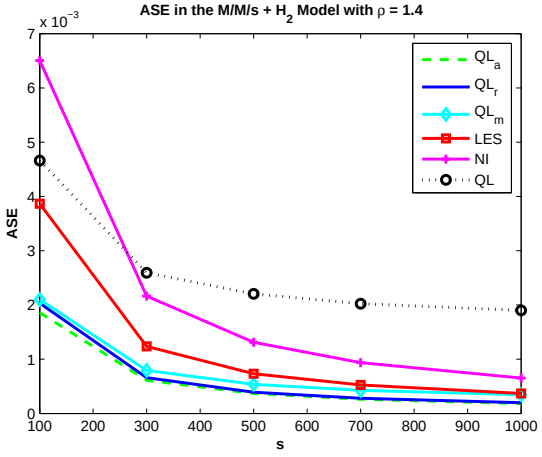

Figure 3.3: ASE of the predictors in the $M/M/s+H_2$ model with $\rho = 1.4$ and $\nu = 1.0$.

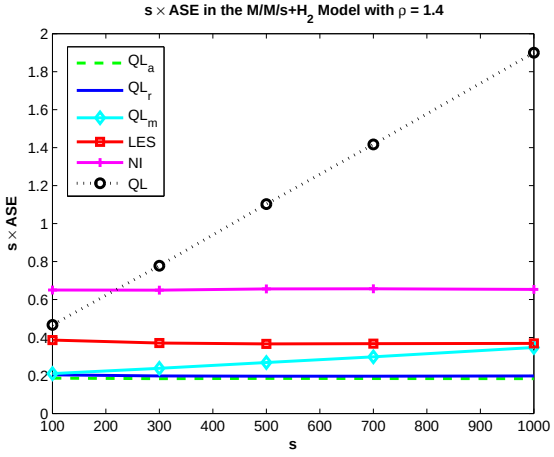

Figure 3.4: $s \times$ ASE of the predictors in the $M/M/s+H_2$ model with $\rho = 1.4$ and $\nu = 1.0$.

ranges from about 2 when $s = 100$ to about 10 when $s = 1000$. Once more, the need to go beyond QL is evident.

Consistent with §3.5, Figure 3.4 shows that all predictors, except QL and QL_m, have an ASE which is inversely proportional to the number of servers, but mathematical support for the predictors (besides NI) has yet to be provided, with non-exponential abandonment distributions. Beyond Theorems 3.5.2 and 3.5.5, the NI behavior is consistent with conjectured stochastic refinements to the fluid limits in Whitt (2006).

3.6.4 Results for the $M/M/s+E_{10}$ model

Figure 3.5 and Table 3.4 show that QL_a is the best possible delay predictor, for this model, except when s is very large (e.g., $s = 700$ or $s = 1000$). The corresponding RRASE ranges from about 10% when $s = 100$ to about 3% when $s = 1000$. The QL_r predictor performs worse than QL_a for smaller values of s, but slightly outperforms QL_a for larger values of s. The ratio $\text{ASE}(\text{QL}_r)/\text{ASE}(\text{QL}_a)$ ranges

Efficiency of the predictors in the $M/M/s+H_2$ model with $\rho = 1.4$ and $\nu = 1.0$

s	$ASE[\theta_{QL_a}]$	$ASE[\theta_{QL_m}]$	$ASE[\theta_{QL_r}]$	$ASE[\theta_{QL}]$	$ASE[\theta_{LES}]$	$ASE[\theta_{NI}]$
100	1.859×10^{-3}	2.100×10^{-3}	2.032×10^{-3}	4.662×10^{-3}	3.866×10^{-3}	6.503×10^{-3}
	$\pm6.52\times10^{-6}$	$\pm5.54\times10^{-6}$	$\pm6.31\times10^{-6}$	$\pm1.83\times10^{-5}$	$\pm8.10\times10^{-6}$	$\pm3.85\times10^{-5}$
300	6.116×10^{-4}	7.933×10^{-4}	6.599×10^{-4}	2.593×10^{-3}	1.236×10^{-3}	2.165×10^{-3}
	$\pm4.64\times10^{-6}$	$\pm7.62\times10^{-6}$	$\pm8.82\times10^{-6}$	$\pm2.25\times10^{-5}$	$\pm1.76\times10^{-5}$	$\pm2.09\times10^{-5}$
500	3.695×10^{-4}	5.367×10^{-4}	3.921×10^{-4}	2.205×10^{-3}	7.331×10^{-4}	1.311×10^{-3}
	$\pm2.19\times10^{-6}$	$\pm2.12\times10^{-6}$	$\pm2.47\times10^{-6}$	$\pm9.97\times10^{-6}$	$\pm5.41\times10^{-6}$	$\pm1.03\times10^{-5}$
700	2.630×10^{-4}	4.257×10^{-4}	2.802×10^{-4}	2.024×10^{-3}	5.250×10^{-4}	9.378×10^{-4}
	$\pm1.43\times10^{-6}$	$\pm1.89\times10^{-6}$	$\pm1.00\times10^{-5}$	$\pm2.35\times10^{-6}$	$\pm2.52\times10^{-6}$	$\pm1.07\times10^{-5}$
1000	1.833×10^{-4}	3.474×10^{-4}	1.978×10^{-4}	1.900×10^{-3}	3.691×10^{-4}	6.533×10^{-4}
	$\pm1.55\times10^{-6}$	$\pm1.43\times10^{-6}$	$\pm6.90\times10^{-7}$	$\pm5.93\times10^{-6}$	$\pm3.00\times10^{-6}$	1.14×10^{-5}

Table 3.3: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/M/s+H_2$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

from nearly 2 when $s = 100$ to nearly 0.9 when $s = 1000$. The QL_m predictor, which was nearly identical to QL_a before, now performs worse particularly when the number of servers is large; e.g., when $s = 1000$, $ASE(QL_m)/ASE(QL_a) \approx 9$. The RRASE of QL_m ranges from about 14% when $s = 100$ to about 10% when $s = 1000$.

In contrast to previous cases, NI is the second or third most effective delay predictor here, depending on the number of servers. It performs nearly as well as QL_a, particularly when s is large. This confirms that NI can be a competitive delay predictor, with customer abandonment. Figure 3.6 shows that the ASE's of QL_a, QL_r, and NI are all inversely proportional to the number of servers s. The LES predictor also fares well but is slightly outperformed by QL_a, QL_r and NI. The corresponding RRASE ranges from about 14% when $s = 100$ to about 3% when $s = 1000$. Nevertheless, Figure 3.6 also shows that $s\times$ ASE(LES) equals a constant, for all

Efficiency of the predictors in the $M/M/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$

s	$ASE[\theta_{QL_a}]$	$ASE[\theta_{QL_m}]$	$ASE[\theta_{QL_r}]$	$ASE[\theta_{QL}]$	$ASE[\theta_{LES}]$	$ASE[\theta_{NI}]$
100	5.388×10^{-3}	9.400×10^{-3}	6.317×10^{-3}	8.097×10^{-2}	8.810×10^{-3}	6.077×10^{-3}
	$\pm1.54\times10^{-5}$	$\pm3.48\times10^{-5}$	4.51×10^{-5}	$\pm2.47\times10^{-4}$	$\pm3.91\times10^{-5}$	$\pm2.63\times10^{-5}$
300	1.955×10^{-3}	7.211×10^{-3}	2.139×10^{-3}	7.211×10^{-2}	2.933×10^{-3}	2.040×10^{-3}
	$\pm5.13\times10^{-6}$	$\pm3.86\times10^{-5}$	$\pm1.83\times10^{-5}$	$\pm3.30\times10^{-4}$	$\pm3.22\times10^{-5}$	$\pm2.23\times10^{-5}$
500	1.244×10^{-3}	6.746×10^{-3}	1.293×10^{-3}	7.049×10^{-2}	1.760×10^{-3}	1.288×10^{-3}
	$\pm1.54\times10^{-5}$	$\pm2.68\times10^{-5}$	$\pm1.35\times10^{-5}$	$\pm2.48\times10^{-4}$	$\pm2.44\times10^{-5}$	$\pm2.61\times10^{-5}$
700	9.572×10^{-4}	6.584×10^{-3}	9.319×10^{-4}	6.975×10^{-2}	1.241×10^{-3}	9.966×10^{-4}
	$\pm8.31\times10^{-6}$	$\pm1.43\times10^{-6}$	$\pm1.00\times10^{-5}$	$\pm1.00\times10^{-5}$	$\pm2.35\times10^{-6}$	$\pm1.30\times10^{-5}$
1000	7.369×10^{-4}	6.454×10^{-3}	6.694×10^{-4}	6.902×10^{-2}	8.830×10^{-4}	8.242×10^{-4}
	$\pm1.96\times10^{-5}$	$\pm1.70\times10^{-5}$	$\pm1.13\times10^{-5}$	$\pm1.68\times10^{-4}$	$\pm1.28\times10^{-5}$	$\pm1.17\times10^{-5}$

Table 3.4: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/M/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

values of s. Finally, QL is yet again the least effective predictor for this model. For example, the ratio ASE(QL)/ASE(QL_a) ranges from about 15 when $s = 100$ to nearly 95 when $s = 1000$. That is why the corresponding ASE curve is not even included in Figures 3.5 and 3.6.

3.7 Simulation Results for the $M/GI/s+M$ Model

In this section, we present simulation results quantifying the performance of the alternative delay predictors with non-exponential service-time distributions; i.e., we consider the $M/GI/s+M$ model. In this model, QL_a coincides with QL_m so we do not include separate results for it. For the service-time distribution, we consider H_2, LN(1, 1) (lognormal with mean and variance equal to 1), D, and E_{10} distributions in Tables 3.5, 3.6, 3.7, and 3.8, respectively. To illustrate problems with D service

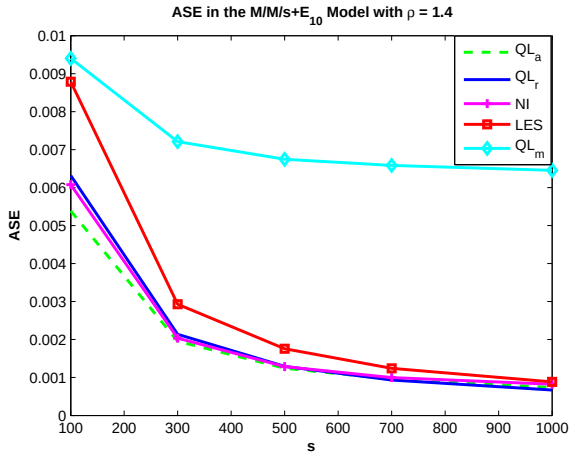

Figure 3.5: ASE of the predictors in the $M/M/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$.

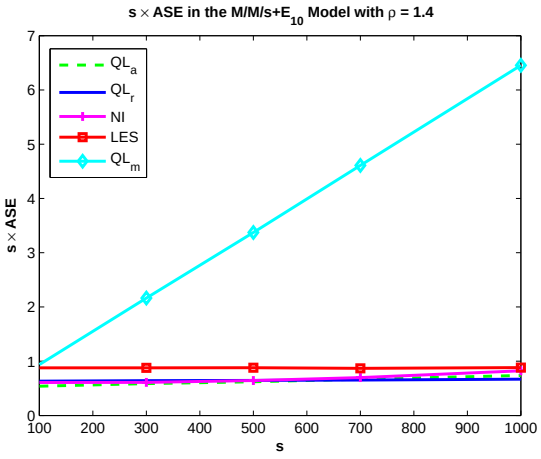

Figure 3.6: $s\times$ASE of the predictors in the $M/M/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$.

Efficiency of the predictors in the $M/H_2/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	3.491×10^{-3}	3.487×10^{-3}	8.720×10^{-3}	6.227×10^{-3}	1.435×10^{-2}
	$\pm2.89\times10^{-5}$	$\pm1.75\times10^{-5}$	$\pm2.66\times10^{-5}$	$\pm3.61\times10^{-5}$	$\pm5.30\times10^{-5}$
300	1.114×10^{-3}	1.117×10^{-3}	5.530×10^{-3}	1.996×10^{-3}	4.893×10^{-3}
	$\pm1.31\times10^{-5}$	$\pm9.83\times10^{-6}$	$\pm2.34\times10^{-5}$	$\pm1.73\times10^{-5}$	$\pm8.37\times10^{-5}$
500	6.660×10^{-4}	6.696×10^{-4}	4.953×10^{-3}	1.190×10^{-3}	2.931×10^{-3}
	$\pm5.68\times10^{-6}$	$\pm5.68\times10^{-6}$	$\pm2.26\times10^{-5}$	$\pm1.14\times10^{-5}$	$\pm5.60\times10^{-5}$
700	4.807×10^{-4}	4.797×10^{-4}	4.672×10^{-3}	8.612×10^{-4}	2.083×10^{-3}
	$\pm6.73\times10^{-6}$	$\pm5.59\times10^{-6}$	$\pm2.46\times10^{-5}$	$\pm1.23\times10^{-5}$	$\pm4.91\times10^{-5}$
1000	3.362×10^{-4}	3.346×10^{-4}	4.489×10^{-3}	6.136×10^{-4}	1.494×10^{-3}
	$\pm3.16\times10^{-6}$	$\pm1.76\times10^{-6}$	$\pm1.62\times10^{-5}$	$\pm9.34\times10^{-6}$	$\pm5.50\times10^{-5}$

Table 3.5: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/H_2/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

times, we also show plots of simulation estimates in Figures 3.7 and 3.8. We let $\mu = \nu = 1.0$, and vary λ, for alternative values of s, to keep $\rho = 1.4$.

Efficiency of the predictors in the $M/LN(1,1)/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.359×10^{-3}	2.596×10^{-3}	8.207×10^{-3}	5.248×10^{-3}	9.089×10^{-3}
	$\pm7.00\times10^{-6}$	$\pm9.02\times10^{-6}$	$\pm4.45\times10^{-5}$	$\pm2.37\times10^{-5}$	$\pm4.80\times10^{-5}$
300	7.810×10^{-4}	8.506×10^{-4}	5.394×10^{-3}	1.716×10^{-3}	3.032×10^{-3}
	$\pm5.14\times10^{-6}$	$\pm5.68\times10^{-6}$	3.36×10^{-5}	$\pm1.25\times10^{-5}$	$\pm5.30\times10^{-5}$
500	4.663×10^{-4}	5.0685×10^{-4}	4.836×10^{-3}	1.029×10^{-3}	1.826×10^{-3}
	$\pm2.04\times10^{-6}$	$\pm2.12\times10^{-6}$	$\pm2.085\times10^{-5}$	$\pm7.29\times10^{-6}$	$\pm8.10\times10^{-6}$
700	3.346×10^{-4}	3.635×10^{-4}	4.615×10^{-3}	7.438×10^{-4}	1.290×10^{-3}
	$\pm2.71\times10^{-6}$	$\pm3.37\times10^{-6}$	$\pm1.77\times10^{-5}$	$\pm6.47\times10^{-6}$	$\pm1.12\times10^{-5}$
1000	2.340×10^{-4}	2.548×10^{-4}	4.443×10^{-3}	5.290×10^{-4}	8.942×10^{-4}
	$\pm1.84\times10^{-6}$	$\pm2.81\times10^{-6}$	$\pm2.54\times10^{-5}$	$\pm5.90\times10^{-6}$	$\pm2.46\times10^{-5}$

Table 3.6: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/LN(1,1)/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

3.7.1 The $M/H_2/s+M$ model

Table 3.5 shows that, with high variability in the service times, the results that we get are not too different from those we get with M service times. The QL_m predictor is the most efficient predictor for this model. The RRASE of QL_m ranges from about 17% when $s = 100$ to about 5% when $s = 1000$. The QL_r predictor is only very slightly outperformed by QL_m (the ratio ASE(QL_m)/ASE(QL_r) is very close to 1 for all values of s). The LES predictor is relatively accurate as well: The ratio ASE(LES)/ASE(QL_m) is close to 1.8 for all values of s, suggesting possible extensions for our analytical results for the $GI/M/s+M$ model to the $M/GI/s+M$ model. The NI predictor is outperformed by QL_m, QL_r, and LES: The ratio ASE(NI)/ASE(QL_m) is close to 4 for all values of s considered.

3.7.2 Results for the $M/LN(1,1)/s+M$ model

We consider the lognormal distribution for the service times because there is empirical evidence suggesting a remarkable fit of the service-time distribution to the lognormal distribution; e.g., see Brown et al. (2005). Results with LN service times are consistent with those corresponding to M service times. Table 3.6 shows that QL_m is the most effective delay predictor for this model. The RRASE for QL_m ranges from approximately 14% when $s = 100$ to approximately 5% when $s = 1000$. The QL_r predictor is slightly less efficient than QL_m: The ratio $\text{ASE}(\text{QL}_r)/\text{ASE}(\text{QL}_m)$ ranges from approximately 1.1 when $s = 100$ to approximately 1.08 when $s = 1000$. The LES predictor is relatively accurate as well: The RRASE of LES ranges from approximately 26% when $s = 100$ to approximately 7% when $s = 1000$. The NI predictor does not perform as well as LES, nor as bad as QL. The QL predictor is the least efficient predictor: the ratio $\text{ASE}(\text{QL})/\text{ASE}(\text{QL}_m)$ ranges from approximately 4 when $s = 100$ to approximately 19 when $s = 1000$.

3.7.3 The $M/D/s+M$ model

Figure 3.7 shows that there is a significant increase in ASE for all predictors with deterministic (constant) service times, with performance tending to be independent of s. However, even very low variability in the service times, e.g., the E_{10} distribution with SCV equal to 0.1, is enough for our delay predictors to be relatively accurate (see §3.7.4). With D service times, Figure 3.8 shows that $s\times\text{ASE}$ for all predictors increases with s. That is, we do not see an improvement in performance in large systems. In chapter 5, we will show that with time-varying demand and capacity, our delay prediction techniques remain ineffective even with low variability in the service times. Alternative delay prediction procedures, appropriate for deterministic

service times, remain to be investigated.

Table 3.7 shows that the NI predictor, which uses no information at all beyond the model, is the most effective delay predictor when $s \geq 300$. (For $s = 100$, QL_m slightly outperforms NI.) But, even the NI predictor is not very accurate: The RRASE for NI is roughly equal to 25% for all values of s considered. The ASE's for QL_m, QL_r, QL, and LES do not vary much in this model; e.g., ASE(QL_m) varies little about 0.01, for all values of s considered.

3.7.4 The $M/E_{10}/s+M$ model

Table 3.8 shows that the proposed delay predictors remain effective even with very low variability in the service times. The QL_m predictor is the most effective delay predictor for the $M/E_{10}/s+M$ model. The QL_r predictor is nearly identical to QL_m, particularly when s is large enough ($s \geq 300$). Once more, the relative accuracy of the delay predictors improves as s increases. For example, the RRASE for QL_m ranges from approximately 13% when $s = 100$ to approximately 4% when $s = 1000$. The LES predictor is relatively accurate as well. The RRASE of LES ranges from approximately 21% when $s = 100$ to approximately 7% when $s = 1000$.

The NI predictor does not perform as well as LES, nor as bad as QL. The QL predictor is the least efficient predictor: The ratio ASE(QL)/ASE(QL_m) ranges from approximately 4 when $s = 100$ to approximately 22 when $s = 1000$. Consistent with §3.5, Table 3.8 shows that all predictors, except QL, have an ASE which is inversely proportional to the number of servers, but mathematical support for the predictors has yet to be provided with non-exponential service-time distributions.

Efficiency of the predictors in the $M/D/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	9.171×10^{-3}	1.066×10^{-2}	1.772×10^{-2}	1.525×10^{-2}	9.316×10^{-3}
	$\pm3.08\times10^{-3}$	$\pm3.56\times10^{-3}$	$\pm4.08\times10^{-3}$	$\pm4.12\times10^{-3}$	$\pm1.59\times10^{-3}$
300	1.492×10^{-2}	1.698×10^{-2}	2.400×10^{-2}	2.511×10^{-2}	8.553×10^{-3}
	$\pm1.91\times10^{-3}$	$\pm2.15\times10^{-3}$	$\pm2.48\times10^{-3}$	$\pm3.48\times10^{-3}$	$\pm1.084\times10^{-3}$
500	1.560×10^{-2}	1.771×10^{-2}	2.469×10^{-2}	2.585×10^{-2}	7.806×10^{-3}
	$\pm2.85\times10^{-3}$	$\pm3.23\times10^{-3}$	$\pm3.72\times10^{-3}$	$\pm4.64\times10^{-3}$	$\pm6.00\times10^{-4}$
700	1.259×10^{-2}	1.433×10^{-2}	2.071×10^{-2}	2.015×10^{-2}	8.232×10^{-3}
	1.590×10^{-3}	1.797×10^{-3}	2.053×10^{-3}	2.566×10^{-3}	$\pm9.059\times10^{-4}$
1000	1.417×10^{-2}	1.611×10^{-2}	2.267×10^{-2}	2.246×10^{-2}	7.566×10^{-3}
	$\pm1.515\times10^{-3}$	$\pm1.706\times10^{-3}$	$\pm1.964\times10^{-3}$	$\pm2.64\times10^{-3}$	$\pm4.711\times10^{-4}$

Table 3.7: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/D/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $M/E_{10}/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.024×10^{-3}	2.405×10^{-3}	8.249×10^{-3}	5.052×10^{-3}	6.284×10^{-3}
	$\pm6.51\times10^{-6}$	$\pm9.91\times10^{-6}$	$\pm4.27\times10^{-5}$	$\pm2.12\times10^{-5}$	$\pm2.63\times10^{-5}$
300	6.790×10^{-4}	7.972×10^{-4}	5.439×10^{-3}	1.687×10^{-3}	2.111×10^{-3}
	$\pm2.48\times10^{-6}$	$\pm2.71\times10^{-6}$	$\pm2.39\times10^{-5}$	$\pm8.44\times10^{-6}$	$\pm2.51\times10^{-5}$
500	4.072×10^{-4}	4.775×10^{-4}	4.857×10^{-3}	1.001×10^{-3}	1.266×10^{-3}
	$\pm2.81\times10^{-6}$	$\pm3.48\times10^{-6}$	$\pm2.04\times10^{-5}$	$\pm7.67\times10^{-6}$	$\pm1.81\times10^{-5}$
700	2.946×10^{-4}	3.449×10^{-4}	4.632×10^{-3}	7.147×10^{-4}	9.006×10^{-4}
	$\pm1.41\times10^{-6}$	$\pm1.84\times10^{-6}$	$\pm2.20\times10^{-5}$	$\pm7.31\times10^{-6}$	$\pm1.64\times10^{-5}$
1000	2.063×10^{-4}	2.408×10^{-4}	4.440×10^{-3}	5.073×10^{-4}	6.480×10^{-4}
	$\pm2.37\times10^{-6}$	$\pm2.89\times10^{-6}$	$\pm2.653\times10^{-5}$	$\pm3.95\times10^{-6}$	$\pm1.55\times10^{-5}$

Table 3.8: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/E_{10}/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

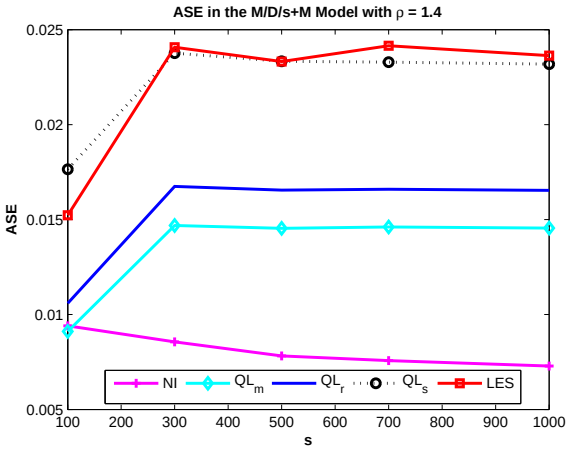

Figure 3.7: ASE of the predictors in the $M/D/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$.

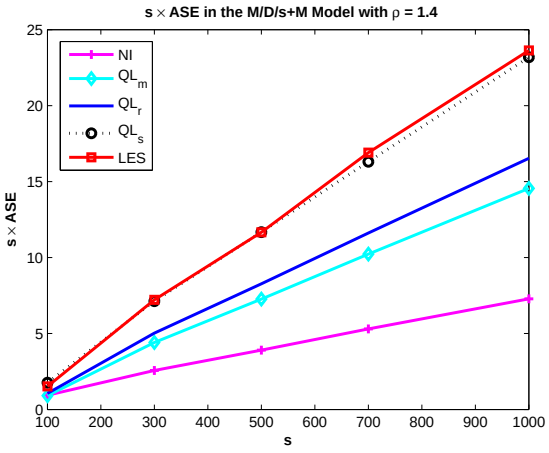

Figure 3.8: $s \times$ ASE of the predictors in the $M/D/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$.

3.8 Simulations Results for the $GI/M/s+M$ Model

In this section, we present simulation results quantifying the performance of the alternative delay predictors with non-exponential interarrival-time distributions; i.e., we consider the $GI/M/s+M$ model. For the interarrival-time distribution, we consider D and H_2 distributions. The simulation results of this section provide further support to the theoretical results in §3.5. We also consider different abandonment rates; specifically we let $\nu = 0.2$ and $\nu = 5.0$. As indicated by Formulas (3.3) and (3.7), the queue length and delay tend to be inversely proportional to ν. Thus, changing ν from 1.0 to 0.2 or 5.0 tends to change congestion by a factor of 5. The system is very heavily overloaded when $\nu = 0.2$, but relatively lightly loaded when $\nu = 5.0$. We consider the same values of s as before and we let $\mu = 1$. We vary λ to get a fixed value of ρ ($\rho = 1.4$), for alternative values of s. We present additional simulation results for the $GI/M/s+M$ model in §A.2.

3.8.1 Results for the $D/M/s+M$ model with $\nu = 0.2$

Table 3.9 compares the efficiencies of the alternative delay predictors in the $D/M/s+M$ model with $\nu = 0.2$. Consistent with theory, QL_m is the optimal delay predictor for this model, under the MSE criterion. The RRASE of QL_m ranges from approximately 35% when $s = 100$ to approximately 11% when $s = 1000$. The QL_r predictor is slightly less efficient than QL_m: $ASE(QL_r)/ASE(QL_m)$ is less than 1.05 for all values of s considered. The LES predictor is slightly less accurate, with an RRASE ranging from approximately 40% when $s = 100$ to approximately 13% when $s = 1000$. The NI predictor is less accurate than LES, but not as bad as QL. The QL predictor is, once more, the least effective predictor: The ratio $ASE(QL)/ASE(QL_m)$ ranges from approximately 8 when $s = 100$ to approximately 71 when $s = 1000$.

Table 3.9 substantiates (3.55) and (3.29) of §3.5, that compare the performances of QL_m, LES and NI in the $D/M/s+M$ model. Consistent with (3.55), Table 3.9 shows that the performance of LES is close to that of QL_m, when the arrival process is deterministic. The simulation estimates of $ASE(LES)/ASE(QL_m)$, for alternative values of s, are remarkably close to the numerical value, approximately 1.286, predicted by (3.55); the relative error (RE) observed is less than 1% for all values of s considered. Consistent with (3.29), Table 3.9 shows that the performance of NI is worse than that of LES and QL_m. The simulation estimates of $ASE(NI)/ASE(QL_m)$ are also remarkably close to the numerical value, 2.25, predicted by (3.29); the RE observed is less than 4% for all values of s considered.

Efficiency of the predictors in the $D/M/s+M$ model with $\rho = 1.4$ and $\nu = 0.2$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	1.436×10^{-2}	1.492×10^{-2}	1.192×10^{-1}	1.863×10^{-2}	3.266×10^{-2}
	$\pm9.78\times10^{-5}$	$\pm9.40\times10^{-5}$	$\pm1.57\times10^{-4}$	$\pm1.64\times10^{-4}$	$\pm5.33\times10^{-4}$
300	4.798×10^{-3}	5.005×10^{-3}	1.071×10^{-1}	6.172×10^{-3}	1.056×10^{-2}
	$\pm5.99\times10^{-5}$	$\pm6.08\times10^{-5}$	$\pm1.41\times10^{-4}$	$\pm7.45\times10^{-5}$	$\pm1.92\times10^{-4}$
500	2.865×10^{-3}	2.966×10^{-3}	1.044×10^{-1}	3.672×10^{-3}	6.641×10^{-3}
	$\pm5.43\times10^{-5}$	$\pm5.24\times10^{-5}$	$\pm1.07\times10^{-4}$	$\pm6.67\times10^{-5}$	$\pm2.93\times10^{-4}$
700	2.091×10^{-3}	2.170×10^{-3}	1.033×10^{-1}	2.691×10^{-3}	4.802×10^{-3}
	$\pm2.39\times10^{-5}$	$\pm1.90\times10^{-5}$	$\pm1.54\times10^{-4}$	$\pm3.23\times10^{-5}$	$\pm2.26\times10^{-4}$
1000	1.435×10^{-3}	1.507×10^{-3}	1.026×10^{-1}	1.859×10^{-3}	3.030×10^{-3}
	$\pm1.15\times10^{-5}$	$\pm1.52\times10^{-5}$	$\pm1.20\times10^{-4}$	$\pm2.06\times10^{-5}$	$\pm1.05\times10^{-4}$

Table 3.9: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $D/M/s+M$ model with $\rho = 1.4$ and $\nu = 0.2$. The ASE's are measured in units of mean service time squared per customer.

3.8.2 Results for the $H_2/M/s+M$ model with $\nu = 5.0$

Table 3.10 compares the efficiencies of the alternative delay predictors in the $H_2/M/s+M$ model with $\nu = 5.0$, which makes the model more lightly loaded. Consistent with theory, QL_m is the optimal delay predictor for this model, under the MSE criterion. The RRASE of QL_m ranges from approximately 8% when $s = 100$ to approximately 2% when $s = 1000$.

In this more lightly loaded setting, the ASE's of all the predictors are relatively low, being smaller than for the $M/M/s+M$ model with $\nu = 1.0$ in Table 3.2 by a factor of about 4, despite having $c_a^2 = 4.0$ instead of $c_a^2 = 1.0$. However, the lighter loading makes the ED heavy-traffic approximations less appropriate.

The QL_r predictor is less efficient than QL_m: ASE(QL_r)/ASE(QL_m) ranges from approximately 1.5 when $s = 100$ to approximately 1.25 when $s = 1000$. The LES

Efficiency of the predictors in the $H_2/M/s+M$ model with $\rho = 1.4$ and $\nu = 5.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	7.193×10^{-4}	1.059×10^{-3}	2.217×10^{-3}	2.393×10^{-3}	3.101×10^{-3}
	$\pm2.63\times10^{-6}$	$\pm4.47\times10^{-6}$	$\pm1.01\times10^{-5}$	$\pm6.72\times10^{-6}$	$\pm1.42\times10^{-5}$
300	2.008×10^{-4}	2.675×10^{-4}	7.240×10^{-4}	7.569×10^{-4}	1.169×10^{-3}
	$\pm7.85\times10^{-7}$	$\pm1.28\times10^{-6}$	$\pm2.63\times10^{-6}$	$\pm2.70\times10^{-6}$	$\pm5.82\times10^{-6}$
500	1.167×10^{-4}	1.495×10^{-4}	4.792×10^{-4}	4.540×10^{-4}	7.624×10^{-4}
	$\pm7.05\times10^{-7}$	$\pm8.78\times10^{-7}$	$\pm2.68\times10^{-6}$	$\pm1.71\times10^{-6}$	$\pm6.07\times10^{-6}$
700	8.277×10^{-5}	1.042×10^{-4}	3.856×10^{-4}	3.280×10^{-4}	5.714×10^{-4}
	$\pm4.12\times10^{-7}$	$\pm6.52\times10^{-7}$	$\pm2.50\times10^{-6}$	$\pm1.27\times10^{-6}$	$\pm4.72\times10^{-6}$
1000	5.733×10^{-5}	7.141×10^{-5}	3.184×10^{-4}	2.302×10^{-4}	4.0951×10^{-4}
	$\pm2.48\times10^{-7}$	$\pm2.44\times10^{-7}$	$\pm1.34\times10^{-6}$	$\pm1.19\times10^{-6}$	$\pm4.15\times10^{-6}$

Table 3.10: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $H_2/M/s+M$ model with $\rho = 1.4$ and $\nu = 5.0$. The ASE's are measured in units of mean service time squared per customer.

predictor is less accurate, with an RRASE ranging from approximately 14% when $s = 100$ to approximately 4% when $s = 1000$. The QL predictor performs slightly worse than LES: The ratio ASE(QL)/ASE(QL_m) ranges from about 3 when $s = 100$ to about 5 when $s = 1000$. The NI predictor is the least efficient predictor for this model.

Table 3.10 substantiates (3.56) and (3.29) of §3.5, which compare the performances of QL_m, LES and NI in the $H_2/M/s+M$ model. Consistent with (3.56), Table 3.10 shows that the performance of LES is significantly worse than that of QL_m, when the arrival process is highly variable. The simulation estimates of ASE(LES)/ASE(QL_m), for alternative values of s, are close to the numerical value, approximately 4.143, predicted by (3.56), especially for large values of s; the RE observed ranges from approximately -20% for $s = 100$ to approximately -3% when $s = 1000$. We observe a relatively poor performance of the approximation in (3.56)

when the number of servers is small. That is understandable because the system is not very heavily loaded when $\nu = 5.0$. Consistent with (3.29), Table 3.10 shows that the performance of NI is much worse than that of QL_m, when the arrival process is highly variable. The approximation in (3.29) performs poorly when $s = 100$ (RE $\approx -40\%$) but becomes remarkably accurate when $s = 1000$ (RE $\approx -1.5\%$).

3.9 Simulation Results for the $M/GI/s+GI$ Model

In this section, we present simulation results quantifying the performance of the alternative predictors in the $M/GI/s+GI$ model, i.e., we consider different combinations of service-time and abandon-time distributions. We consider D and E_{10} distributions in Tables 3.11 and 3.12. We do not consider D abandonment times because our QL_a predictor requires a density. Constant service times cause a problem in all cases, but otherwise the predictors perform well. Results with higher variability distributions, such as H_2 for example, are not displayed here because they are consistent with previous results obtained with M service or abandonment times.

3.9.1 The $M/D/s+E_{10}$ model

Table 3.11 shows that we get slightly better results with deterministic service times and low-variability abandonment times (Erlang with SCV = 0.1), than those we get with the $M/D/s+M$ model. The LES predictor is the most efficient predictor when the number of servers s is large enough ($s \geq 500$). The RRASE for LES ranges from about 13% when $s = 100$ to about 9% when $s = 1000$, so we see a slight improvement in performance as s increases. The NI predictor is competitive as well, and is the second most efficient predictor when $s \geq 500$. The QL_a predictor is

the most efficient predictor when $s \leq 300$, but not otherwise. The QL_m predictor performs poorly, but not as bad as QL which is the least efficient predictor. Table 3.11 shows that $s \times$ ASE of all predictors increase nearly linearly with s in the $M/D/s+E_{10}$ model.

3.9.2 The $M/E_{10}/s+E_{10}$ model

Table 3.12 shows that the proposed delay predictors remain effective, with very low variability in the service times, even if combined with low-variability abandonment times. The QL_a predictor is the most effective delay predictor for the $M/E_{10}/s+E_{10}$ model. The NI predictor is competitive as well, and is the second most effective predictor in this model. The LES predictor is relatively accurate as well. The NI predictor does not perform as well as LES, nor as bad as QL. The QL predictor is the least efficient predictor. Except for QL_m, the relative accuracy of the delay predictors improves as s increases. Indeed, the products $s \times$ ASE are nearly constant for all predictors, except QL_m, but mathematical support for the predictors has yet to be provided with non-exponential service-time distributions.

3.10 Concluding Remarks

In this chapter, we proposed several delay predictors and showed that they are effective in the multiserver $GI/GI/s+GI$ model. As a frame of reference, we considered the classical delay predictor based on the queue length, QL, treated extensively in chapter 2. We showed that QL performs poorly when there is significant customer abandonment in the system, thus establishing the need to propose better ways of making delay predictions.

Efficiency of the predictors in the $M/D/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_a}]$	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	8.678×10^{-3}	1.286×10^{-2}	1.038×10^{-2}	8.464×10^{-2}	1.174×10^{-2}	9.016×10^{-3}
	$\pm3.04\times10^{-3}$	$\pm3.05\times10^{-3}$	$\pm3.49\times10^{-3}$	$\pm3.17\times10^{-3}$	$\pm3.97\times10^{-3}$	$\pm2.92\times10^{-3}$
300	7.377×10^{-3}	1.384×10^{-2}	9.827×10^{-3}	7.902×10^{-2}	7.809×10^{-3}	8.230×10^{-3}
	$\pm9.69\times10^{-4}$	$\pm1.60\times10^{-3}$	$\pm1.81\times10^{-3}$	$\pm1.56\times10^{-3}$	$\pm1.13\times10^{-3}$	$\pm1.63\times10^{-3}$
500	7.344×10^{-3}	1.318×10^{-2}	8.821×10^{-3}	7.714×10^{-2}	6.763×10^{-3}	7.088×10^{-3}
	$\pm1.18\times10^{-3}$	$\pm9.99\times10^{-4}$	$\pm1.10\times10^{-3}$	$\pm9.55\times10^{-4}$	$\pm5.09\times10^{-4}$	$\pm1.11\times10^{-3}$
700	7.336×10^{-3}	1.296×10^{-2}	8.412×10^{-3}	7.628×10^{-2}	5.718×10^{-3}	6.805×10^{-3}
	$\pm9.29\times10^{-4}$	$\pm1.06\times10^{-3}$	$\pm1.18\times10^{-3}$	$\pm8.65\times10^{-4}$	$\pm4.15\times10^{-4}$	$\pm1.11\times10^{-3}$
1000	7.269×10^{-3}	1.303×10^{-2}	8.327×10^{-3}	7.575×10^{-2}	5.316×10^{-3}	6.828×10^{-3}
	$\pm6.57\times10^{-4}$	$\pm8.15\times10^{-4}$	$\pm9.03\times10^{-4}$	$\pm8.63\times10^{-4}$	$\pm3.16\times10^{-4}$	$\pm8.64\times10^{-4}$

Table 3.11: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/D/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $M/E_{10}/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_a}]$	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	3.539×10^{-3}	7.632×10^{-3}	4.457×10^{-3}	7.940×10^{-2}	6.348×10^{-3}	4.011×10^{-3}
	$\pm1.91\times10^{-5}$	$\pm1.44\times10^{-5}$	$\pm2.25\times10^{-5}$	$\pm2.75\times10^{-4}$	$\pm2.35\times10^{-5}$	$\pm1.78\times10^{-5}$
300	1.295×10^{-3}	6.603×10^{-3}	1.502×10^{-3}	7.181×10^{-2}	2.102×10^{-3}	1.364×10^{-3}
	$\pm7.50\times10^{-6}$	$\pm1.53\times10^{-5}$	$\pm1.45\times10^{-5}$	$\pm2.90\times10^{-4}$	$\pm1.75\times10^{-5}$	$\pm1.52\times10^{-5}$
500	8.642×10^{-4}	6.440×10^{-3}	8.984×10^{-4}	7.001×10^{-2}	1.260×10^{-3}	8.660×10^{-4}
	$\pm1.16\times10^{-5}$	$\pm1.88\times10^{-5}$	$\pm6.61\times10^{-6}$	$\pm2.56\times10^{-4}$	$\pm1.17\times10^{-5}$	$\pm1.33\times10^{-5}$
700	6.752×10^{-4}	6.326×10^{-3}	6.440×10^{-4}	6.923×10^{-2}	9.068×10^{-4}	6.771×10^{-4}
	$\pm9.87\times10^{-6}$	$\pm9.13\times10^{-6}$	$\pm9.15\times10^{-6}$	$\pm1.84\times10^{-4}$	$\pm1.27\times10^{-5}$	$\pm1.15\times10^{-5}$
1000	5.413×10^{-4}	6.230×10^{-3}	4.592×10^{-4}	6.890×10^{-2}	6.406×10^{-4}	5.547×10^{-4}
	$\pm8.62\times10^{-6}$	$\pm2.03\times10^{-5}$	$\pm4.29\times10^{-6}$	$\pm2.70\times10^{-4}$	$\pm6.66\times10^{-6}$	$\pm1.37\times10^{-5}$

Table 3.12: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/E_{10}/s+E_{10}$ model with $\rho = 1.4$ and $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

The Markovian queue-length-based predictor, QL_m, is a variant of QL that accounts for customer abandonment by assuming that waiting customers have i.i.d. exponential abandonment times with rate ν. It also assumes that service times are i.i.d. with an exponential distribution. In §3.3, we showed that QL_m is the most accurate predictor, under the MSE criterion, in the $GI/M/s+M$ model. We established heavy-traffic limits that generated an approximation for the expected MSE of QL_m in the $GI/M/s+M$ model in §3.5.

In practice, the QL_m predictor is effective whenever the abandonment-time and service-time distributions in the actual service system are well modeled by an exponential distribution. The abandonment rate ν, which is required for the implementation of QL_m, can be estimated from system data as the ratio of the proportion of abandoning customers to the average waiting time in the queue; see §5 of Garnett et al. (2002). The average waiting time in the queue and the proportion of abandoning customers are fairly standard system data outputs. In the context of call centers, for example, they can be easily obtained from the Automatic Call Distributor's (ACD) data.

But, QL_m is not always so good for the more general $GI/GI/s+GI$ model. In Figures 3.3 and 3.5, we showed that it can be inferior to all other predictors (except QL) with a non-exponential abandonment-time distribution. Since non-exponential service-time and abandonment-time distributions are commonly observed in practice, it is important to propose other queue-length-based delay predictors that effectively cope with non-exponential distributions. Approximations are needed because direct mathematical analysis is difficult.

We proposed the simple-refined QL_r predictor, which multiplies the QL prediction by a model-dependent constant, based on fluid approximations in the ED heavy-traffic limiting regime. Simulation results in §3.6 and the e-companion show that QL_r

performs remarkably well. The QL_r predictor is competitive whenever the actual service system is large and overloaded, i.e., whenever the fluid approximations are appropriate. The QL_r predictor performs significantly better than QL_m (and QL) when the abandonment-time distribution is not nearly exponential; e.g., see Tables 3.3 and 3.4.

Our most promising delay predictor is the new approximation-based predictor, QL_a. Simulation results in §3.6 show that QL_a is consistently the most effective predictor (with the exception of D service). It is a variant of QL_m which assumes that abandonment times are independent, exponential, with state-dependent abandonment rates. The QL_a predictor coincides with QL_m in the setting of the $GI/M/s+M$ model, and is thus the most accurate for that model, under the MSE criterion. It also performs remarkably well for non-exponential abandonment-time distributions.

The QL_a delay prediction, $\theta_{QL_a}(n)$ in (3.15), requires knowledge of the abandonment-time hazard-rate function, h. That is convenient from a practical point of view, because it is relatively easy to estimate hazard rates from system data; see Brown et al. (2005). It is significant that QL_r and QL_a require knowledge of the arrival rate, λ, which requires some degree of stationarity. These predictors should be effective whenever the arrival rate in the actual service system does not vary too rapidly.

Unlike without abandonments, the NI predictor, announcing the deterministic heavy-traffic fluid limit w of the waiting time, is an effective predictor in the overloaded $GI/GI/s+GI$ model. It is best possible for D service, but not otherwise. Nevertheless, it is remarkably effective, especially when the abandonment-time distribution has low variability. The NI predictor is a competitive predictor whenever the actual service system is large and overloaded, and the service and abandonment times are not highly variable.

Finally we considered the LES predictor, which is appealing because it only depends on the history of delays in the system. Intuitively, we should expect that LES will perform worse than queue-length-based predictors when the queue length and model parameters are known, because it does not exploit information about system state. Simulation shows that this is usually true. Nevertheless, the LES predictor is quite effective in all models considered. In §3.5, we showed that the expected MSE of LES in the $GI/M/s + M$ model increases with the squared coefficient of variation of the interarrival times, c_a^2. The practical significance of this result is that reliability of LES increases as the variability in the arrival process decreases.

4

Delay-History-Based Predictors with Time-Varying Arrivals

4.1 Introduction

In this chapter, we study real-time delay predictors in heavily loaded many-server service systems with time-varying arrival rates. Our main contributions are: (i) to show that time-varying arrival rates can cause prediction bias for delay-history-based delay predictors, (ii) to propose new and easily implementable delay predictors, based on the history of delays in the system, that effectively cope with time-varying arrivals and general service-time and abandon-time distributions, (iii) to provide analytical results quantifying the performance of some delay predictors, and (iv) to describe results of a wide range of simulation experiments evaluating alternative delay predictors, with time-varying arrivals. This chapter is an edited version of a paper currently under revision, Ibrahim and Whitt (2010a).

4.1.1 Delay-History-Based Predictors

In this chapter, we examine alternative predictors based on recent customer delay history in the system. For completeness, we now briefly review those predictors. As in Armony et al. (2008), a candidate delay predictor based on recent customer delay history is the delay of the last customer to have entered service, prior to our customer's arrival at time t, denoted by LES. That is, letting w be the delay of the last customer to have entered service, the corresponding LES delay prediction is: $\theta_{LES}(t, w) \equiv w$. Armony et al. (2008) studied delay announcements in many-server queues with customer abandonment, focusing on customer response to the announcements, leading to balking and new abandonment behavior. They developed ways to approximately describe the equilibrium system performance using LES delay announcements.

Closely related to LES is the elapsed waiting time of the customer at the head of the line (HOL), assuming that there is at least one customer waiting at the new arrival epoch. The HOL delay predictor was mentioned as a candidate delay announcement by Nakibly (2002). For a more detailed discussion of the HOL and LES predictors, see chapter 2. Experience indicates that the LES and HOL predictors have very similar performance. In complex systems, the LES delay is more likely to be observable than the HOL delay, because arrival and service completion times are more likely to be known than the experience of customers who have not yet completed their service; e.g., customers may have abandoned and that might not be known. Nevertheless, here we focus on HOL, because it is easier to analyze. However, we do so with the understanding that similar results will hold for LES.

4.1.2 The Case of a Stationary Arrival Process

In chapters 2 and 3, we studied the performance of the LES and HOL delay predictors in many-server systems, both with and without customer abandonment, by studying conventional stationary queueing models. In chapter 2, we studied the performance of HOL in the $GI/M/s$ queueing model and showed that HOL is an effective predictor in that model. As a frame of reference, we considered the classical delay predictor based on the queue length, denoted by QL, which multiplies the queue length plus one times the mean interval between successive service completions, ignoring customer abandonment. In the $GI/M/s$ model, the QL predictor is provably the most effective predictor, under the mean squared error (MSE) criterion; see §2.2. (The argument is reviewed in §4.4 below.) The HOL predictor performs worse than QL, because it does not exploit queue-length information. Nevertheless, we showed that the difference in performance need not be too great, particularly when the arrival process has low variability. Because the model is highly structured, we were able to obtain analytical results.

In chapter 3, we considered the $GI/GI/s + GI$ model. As one would expect, QL can overestimate customer delay when there is significant customer abandonment in the system. We showed that QL performs poorly in a heavily loaded $GI/GI/s + GI$ model, while HOL remains effective predictor. When customer abandonment is a serious issue, it is possible to refine the queue-length-based delay predictor by using the exact expected conditional delay, given the queue length, in the $G/M/s + M$ model; we denoted this by QL_m. However, for non-exponential service-time and abandonment distributions, the delay-history-based predictors can also outperform this refined queue-length-based predictor QL_m, even when the queue length and the model are known; e.g., see Figures 3.5 and 3.6 of chapter 2.

However, we do not mean to suggest that the queue length does not provide useful information when it is known. Indeed, as shown in chapter 3, our best predictor for the $GI/GI/s+GI$ model is an approximation-based predictor, referred to as QL_a, which exploits the queue length as well as model parameters; we also will make use of QL_a here for the $M(t)/GI/s+GI$ model in §4.8.

4.1.3 Time-Varying Arrival Rates

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

The HOL predictor can perform poorly when the delays vary systematically over time, as can occur when there are alternating periods of significant overload and underload. Then the delay of a new arrival may not be like the HOL delay. To demonstrate potential problems with the HOL predictor, we plot simulation sample paths of HOL delay predictions given, and actual delays observed, as a function of time, in simulation runs from two different heavily-loaded many-server systems. In Figure 4.1, we consider the stationary $M/M/100$ model with traffic intensity $\rho = 0.95$ and mean service time 5 minutes; in Figure 4.2, we consider the $M(t)/M/100$ model with sinusoidal arrival rates, again with traffic intensity $\rho = 0.95$, but now defined as the long-run average, and mean service time 5 minutes. We consider a daily cycle, so that there is one peak during the day. We let the relative amplitude be $\alpha = 0.5$. (The ratio of the peak arrival rate to the average arrival rate is $1+\alpha$.) We measure time and, thus, the delays in units of mean service times. The overall plotted time interval of length 500 mean service times is slightly less than two days, so we see two peaks.

For Figure 2, we deliberately chose an extreme case in which the system alternates between extreme overload and underload, while the number of servers remains fixed. In that setting, the maximum delays themselves are about 40 mean service times or 200 minutes, about 60 times greater than in the stationary environment. Delay prediction tends to be especially important with such large delays. Figure 4.2 shows that, with time-varying arrival rates, the HOL curve is clearly shifted to the right of the actual-delay curve; i.e., there is a time lag between the HOL predictions and the actual delays observed; leading to big errors.

Figure 4.2 also shows a third plot, the plot of a modified HOL predictor, denoted by HOL_m, which we develop in §4.4. Clearly, it eliminates the time lag; visually the HOL_m plot falls on top of the actual delays. The ratio of the average squared errors $\text{ASE(HOL)}/\text{ASE}(\text{HOL}_m)$, defined in §4.3 below, is about 95 in Figure 2. (If we would reduce the relative amplitude from 0.5 to 0.1, then the ratio would be only 1.3; it then requires careful analysis to see the improvement provided by HOL_m over HOL; see Figure 4.4 for the plot.)

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

However, the improved performance of the refined HOL predictors comes at the expense of exploiting more information about the system, such as the arrival rate, the number of service times and the mean service time. That requirement greatly reduces the advantage over queue-length-based delay predictors. Indeed, our strategy for obtaining the refined HOL predictors involves two steps: (i) representing

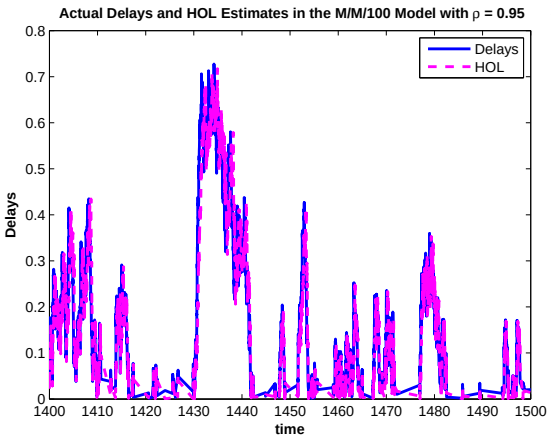

Figure 4.1: Sample paths of actual delays and HOL delay predictions with constant arrival rate

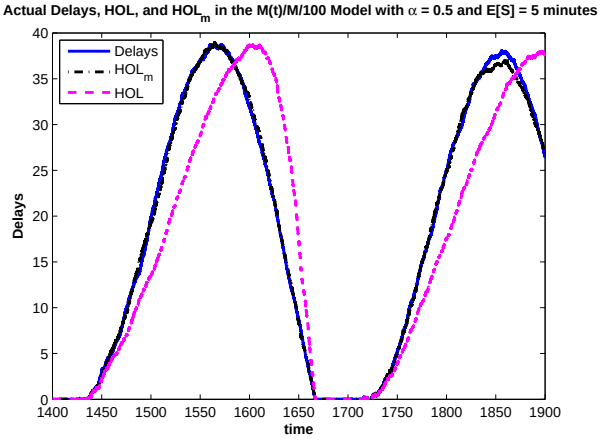

Figure 4.2: Sample paths of actual delays and delay predictions using HOL and HOL_m with a sinusoidal arrival rate and $\alpha = 0.5$

or approximating the expected conditional delay given the queue length, and (ii) estimating the queue length, given the observed HOL delay and the model parameters. Hence, the refined HOL predictors are valuable only when the queue length is not known. However, such cases are not uncommon, as in Web chat and the ticket queues, when we directly observe arrivals and service completions, but not the queue, because we do not observe the customer abandonments.

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

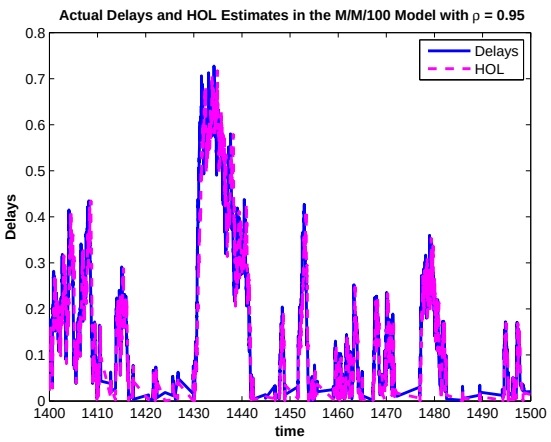

Figure 4.3: Sample paths of actual delays and HOL delay predictions with constant arrival rate

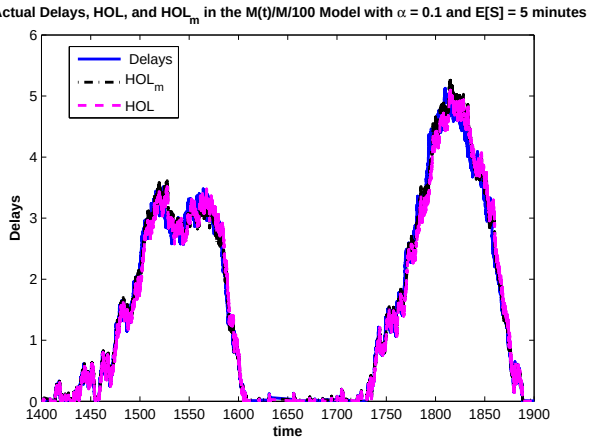

Figure 4.4: Sample paths of actual delays and HOL delay predictions with a sinusoidal arrival rate and $\alpha = 0.1$

4.1.4 Organization

The rest of this chapter is organized as follows: In §4.2, we describe the queueing models considered. In §4.3, we describe measures quantifying the performance of our candidate delay predictors. In §4.4, we introduce a new delay predictor for the $M(t)/GI/s$ model. In §4.5, we provide analytical results for the performance of this predictor in the $M(t)/M/s$ model. In §4.6, we present simulation results showing that it is effective in the $M(t)/GI/s$ model. In §4.7, we propose ways of obtaining the additional system information required for implementing the new delay predictor of §4.4. In §4.8, we develop a new delay predictor for the $M(t)/GI/s+GI$ model. In §4.9, we present simulation results showing that it is effective. We make concluding remarks in §4.10. Additional simulation results with time-varying arrival rates appear in the appendix.

4.2 The Modeling Framework

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

In §§4.4-4.6, we consider the $M(t)/GI/s$ model, which has a nonhomogeneous Poisson arrival process, i.i.d. service times distributed as a random variable S with a general distribution, having mean $E[S] = \mu^{-1}$ and no customer abandonment. Motivated by large service systems, we are primarily interested in the case of large s, which we take to be fixed. It is possible to choose appropriate time-varying staffing (making s a function of time) so that delays are stabilized at low levels; e.g., see Green et al. (2007). However, in practice there often is not adequate flexibility in setting staffing levels. Our fixed staffing assumption captures the spirit of such situations. We leave to future research the important extension to time-varying

staffing levels. In chapter 5 we consider the case of time-varying staffing.

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

In §4.8 and §4.9 we consider the $M(t)/GI/s + GI$ model, which adds customer abandonment. The abandonment times are i.i.d. with mean ν^{-1} and a general cumulative distribution function (cdf) F. As in chapter 3, we see that the abandonment distribution has a significant impact.

4.3 Performance Measures for the Delay Predictors

For completeness, we now indicate how we evaluate the performance of our candidate delay predictors. Once more, we use computer simulation to do the actual estimation. In our simulation experiments, we quantify the performance of a delay predictor by computing the *average squared error* (ASE), defined by:

$$ASE \equiv \frac{1}{k}\sum_{i=1}^{k}(w_i - p_i)^2 \ , \tag{4.1}$$

where $w_i > 0$ is the potential waiting time of delayed customer i, p_i is the delay prediction given to customer i, and k is the number of customers in our sample.

The potential waiting time of a delayed customer is the waiting time he would experience if he had infinite patience. In our simulation experiments, we measure w_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 chapters 2 and 3, the ASE approximates the expected *mean squared error* (MSE) for a system in steady state with a constant arrival rate, but the situation is more complicated with time-varying arrivals. We regard ASE as directly meaningful, but now we indicate how it relates to the MSE. Let $W_{HOL}(t, w)$ represent a random variable with the conditional distribution of the potential delay of an arriving customer, given that this customer must wait before starting service, and given that the elapsed delay of the customer at the head of the line at the time of his arrival, t, is equal to w. Let $\theta_{HOL}(t, w)$ be some given single-number delay prediction which is based on the HOL delay, w, and the time of arrival, t. Then, the MSE of the corresponding delay predictor is given by:

$$MSE(\theta_{HOL}(t, w)) \equiv E[(W_{HOL}(t, w) - \theta_{HOL}(t, w))^2] \ , \tag{4.2}$$

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

$$MSE(t) \equiv E[MSE(\theta_{HOL}(t, W_{HOL}(t)))]. \tag{4.3}$$

Finally, in order to relate the ASE in (4.1) to the MSE, we need to average $MSE(t)$ defined in (4.3) appropriately over time, but since the ASE represents a customer

average instead of a time average, we need to use a weighted time average of the time-dependent MSE in (4.2) in order to relate it to the ASE. In particular, if T is the cycle length, then

$$ASE \approx \frac{\int_0^T \lambda(u) MSE(u) du}{\int_0^T \lambda(u) du} , \tag{4.4}$$

where $MSE(t)$ is defined in (4.3); for supporting theory see the appendix of Massey and Whitt (1994). The right-hand side of (4.4) is the *weighted MSE* (WMSE).

As in chapters 2 and 3, we also quantify the performance of a delay predictor by computing the *root relative average squared error* (RRASE), defined by

$$RRASE \equiv \frac{\sqrt{ASE}}{(1/k)\sum_{i=1}^{k} p_i} , \tag{4.5}$$

using the same notation as in (4.1). The denominator in (4.5) is the average potential waiting time of customers who must wait.

4.4 Delay Predictors for the $M(t)/GI/s$ Model

In this section, we propose a modified HOL-based delay predictor, HOL_m, for the $M(t)/GI/s$ model. Our idea is to use the refined predictor $\theta^r_{HOL}(t, w) \equiv E[W_{HOL}(t, w)]$ instead of the HOL predictor $\theta_{HOL}(t, w) \equiv w$, because the mean necessarily minimizes the MSE based on this information. However, this mean is difficult to compute, so we propose an approximation. We approximate the mean in the given $M(t)/GI/s$ model by its exact value in the corresponding $M(t)/M/s$ model, with exponential service time having the given mean $E[S]$.

For the $M(t)/M/s$ model, we have the representation:

$$W_{HOL}(t, w) \equiv \sum_{i=1}^{A(t)-A(t-w)+2} S_i/s \ , \tag{4.6}$$

where $\{A(t) : t \geq 0\}$ denotes the arrival (counting) process. We have division by s in (4.6) because the times between successive service completions, when all servers are busy, are i.i.d. random variables distributed as the minimum of s exponential random variables, each with rate μ, which makes the minimum exponential with rate $s\mu$. The random variable $A(t) - A(t-w)$ has a Poisson distribution with mean $\int_{t-w}^{t} \lambda(u)du$. Since $W_{HOL}(t, w)$ in (4.6) is a random sum of i.i.d. random variables, where $A(t) - A(t-w)$ is independent of the summands S_i/s, we can easily compute this mean. Hence our refined HOL predictor for the $M(t)/GI/s$ model is this mean

$$\theta_{HOL_r}(t, w) \equiv E[W_{HOL,M(t)/M/s}(t, w)] = \frac{1}{s\mu}\left(2 + \int_{t-w}^{t} \lambda(u)du\right) \ . \tag{4.7}$$

In general, with a non-exponential service-time distribution, $\theta_{HOL_r}(t, w)$ in (4.7) does not equal $E[W_{HOL}(t, w)]$, because many of the remaining service times at time t are residual service times for service times begun prior to time t. Consequently, these service times have a different distribution than the original service time. However, we can make stochastic comparisons. A cumulative distribution function (cdf) G of a nonnegative random variable is said to be new better (worse) than used - NBU (NWU) - if $G_t^c(x) \equiv G^c(t + x)/G^c(t) \leq G^c(x)$ for all $t \geq 0$ and $x \geq 0$, where $G^c(x) \equiv 1 - G(x)$; see p. 159 of Barlow and Proschan (1975). In the parlance of survival analysis, a cdf is NBU (NWU) if the probability of surviving for an additional x time units, given survival up to time t, decreases (increases) with t.

Proposition 1 *If the service-time cdf is NBU (NWU), then*

$$\theta_{HOL_r}(t, w) \geq (\leq) E[W_{HOL}(t, w)].$$

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

More importantly, simulation shows that HOL_m provides a good approximation even when the service-time distribution is not nearly exponential; see §4.6.

We conclude this section by reviewing the QL predictor, previously considered in chapters 2 and 3, using slightly different notation to account for a nonstationary arrival process. Let $W_Q(t, n)$ represent a random variable with the conditional distribution of the potential delay of an arriving customer, given that this customer must wait before starting service, and given that the queue length seen upon arrival, at time t, is equal to n. Again, the QL predictor is obtained by using the exact expected value $E[W_Q(t, n)]$ for the corresponding $M(t)/M/s$ model with the same mean service time.

In the $M(t)/M/s$ model, $W_Q(t, n)$ is the sum of $n + 1$ i.i.d. exponential random variables, each with rate $s\mu$. The QL prediction given to a customer who finds n other customers in queue upon arrival is: $\theta_{QL}(t, n) \equiv E[W_Q(t, n)] = (n + 1)/s\mu$, which depends on t only through n, which is directly observable. The optimal delay

predictor, using the MSE criterion, is the one announcing the mean, $E[W_Q(t,n)]$, conditional on the number of customers, n, seen in line at time t. That is why the QL predictor is the optimal delay predictor, under the MSE criterion, in the $M(t)/M/s$ model.

By essentially the same reasoning as for Proposition 1, we can obtain bounds for

Proposition 2 *If the service-time cdf is NBU (NWU), then*

$$\theta_{QL}(t,n) \geq (\leq) E[W_Q(t,n)].$$

Fortunately, again simulation shows that QL remains effective in the $M(t)/GI/s$ model, even when the service-time distribution is not nearly exponential; see §4.6. For the $M(t)/M/s$ model, we obtain analytical results quantifying the difference in performance between QL and HOL_m in the next section.

4.5 Analytical Expressions for the $M(t)/M/s$ Model

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

$$W_Q(t,n) \equiv \sum_{i=1}^{n+1} S_i/s \ . \tag{4.8}$$

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

$$E[W_Q(t,n)] = \frac{n+1}{s\mu}, \quad Var[W_Q(t,n)] = \frac{n+1}{s^2\mu^2},$$
$$\text{and} \quad c^2_{W_Q(t,n)} \equiv \frac{Var[W_Q(t,n)]}{(E[W_Q(t,n)])^2} = \frac{1}{n+1}\ , \tag{4.9}$$

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

To treat HOL_m, we use the representation in (4.6), which allows us to characterize the probability distribution of the random variable $W_{HOL}(t,w)$, in the $M(t)/M/s$ model.

Proposition 3 *For the $M(t)/M/s$ model,*

$$Var[W_{HOL}(t,w)] = \frac{2}{s^2\mu^2}(1 + \int_{t-w}^{t} \lambda(u)du)\ , \tag{4.10}$$

which, combined with (4.7), yields

$$c^2_{W_{HOL}(t,w)} = \frac{Var[W_{HOL}(t,w)]}{(E[W_{HOL}(t,w)])^2} = 2 \times \frac{1 + \int_{t-w}^{t} \lambda(u)du}{(2 + \int_{t-w}^{t} \lambda(u)du)^2}\ . \tag{4.11}$$

Proof. Formula (4.10) follows from the conditional variance formula, e.g., p.51 of Ross (1996). Formula (4.11) immediately follows from (4.7) and (4.10). ■.

Since $\theta_{HOL_r}(t,w) \equiv E[W_{HOL}(t,w)]$ and $\theta_{QL}(t,n) \equiv E[W_Q(t,n)]$, we can compare the performance of HOL_m and QL by comparing the respective SCV's in (4.9) and (4.11). (When the delay prediction equals the conditional mean, the MSE coincides with the variance.)

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

$$\lambda(u) = \bar{\lambda} + \beta \sin(\gamma u) \equiv \bar{\lambda} + \bar{\lambda}\alpha \sin(2\pi u/\Gamma), \quad \text{for } -\infty < u < \infty\ , \tag{4.12}$$

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

With sinusoidal arrival rates, we obtain analytical results comparing the performance of the QL and HOL_m predictors. We determine the limit of the ratio of the SCV's as $n \to \infty$. Formula (4.13) below coincides with formula (2.37) of chapter 2 for the stationary $GI/M/s$ model. As before, the condition $n \to \infty$ arises naturally in heavy traffic, either with fixed s or as $s \to \infty$; e.g., see Garnett et al. (2002). (When $s \to \infty$ along with the arrival rate, the queue length is of order s and $\sqrt{s}$ in the ED and QED regimes.) Recall that $\rho \equiv \bar{\lambda}/s\mu$.

Proposition 4 *For the $M(t)/M/s$ model with sinusoidal arrival rates,*

$$\frac{c^2_{W_{HOL(t,w)}}}{c^2_{W_Q(n)}} \to \frac{2}{\rho} \text{ as } n \to \infty \ , \tag{4.13}$$

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

Proof. Using Equations (4.7), (4.10), (4.11) and (4.12), we get the following expressions for the mean, variance, and SCV of $W_{HOL}(t, w)$, in the $M(t)/M/s$ model with sinusoidal arrivals:

$$E[W_{HOL}(t, w)] = \frac{2 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t))}{s\mu} \ , \tag{4.14}$$

and,

$$Var[W_{HOL}(t, w)] = 2 \times \frac{1 + \bar{\lambda}w + (\beta/\gamma)(\cos(\gamma t - \gamma w) - \cos(\gamma t)))}{s^2\mu^2} \ , \tag{4.15}$$

which yields

$$c^2_{W_{HOL}(t,w)} = \frac{Var[W_{HOL}(t,w)]}{(E[W_{HOL}(t,w)])^2} = 2\times\frac{1+\bar{\lambda}w+(\beta/\gamma)(\cos(\gamma t-\gamma w)-\cos(\gamma t))}{[2+\bar{\lambda}w+(\beta/\gamma)(\cos(\gamma t-\gamma w)-\cos(\gamma t))]^2}\,, \tag{4.16}$$

for $0 \le t \le \Gamma$. Using (4.16), and recalling that $-1 \le \cos(u) \le 1$ for all u, we obtain the following bounds for the SCV of $W_{HOL}(t,w)$:

$$\frac{2+2\bar{\lambda}w-4\beta/\gamma}{(2+\bar{\lambda}w+2\beta/\gamma)^2} \le c^2_{W_{HOL}(t,w)} \le \frac{2+2\bar{\lambda}w+4\beta/\gamma}{(2+\bar{\lambda}w-2\beta/\gamma)^2}\,. \tag{4.17}$$

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

$$W(t) = \sum_{i=1}^{Q(t)+1} S_i/s\,, \tag{4.18}$$

where $Q(t)$ is the number of customers waiting in queue upon arrival at t, the law of large numbers implies that $W(t)/Q(t) \to 1/s\mu$ as $Q(t) \to \infty$. Thus, when $Q(t)$ is large, we have $W(t) \approx Q(t)/s\mu$. Assuming that n in (4.9) is large with $w = n/s\mu + o(n)$ as $n \to \infty$, where $o(n)$ denotes a quantity that is asymptotically negligible when divided by n, and combining that with (4.17), for large n we get

$$\frac{(2+2\rho(n+o(n))-4\beta/\gamma)(n+1)}{(2+\rho(n+o(n))+2\beta/\gamma)^2} \le \frac{c^2_{W_{HOL}(t,w)}}{c^2_{W_Q(n)}} \le \frac{(2+2\rho(n+o(n))+4\beta\gamma)(n+1)}{(2+\rho(n+o(n))-2\beta/\gamma)^2}\,, \tag{4.19}$$

for all t. By a sandwiching argument, (4.19) yields (4.13) as $n \to \infty$. ∎

4.6 Simulations Experiments for the $M(t)/GI/s$ Model

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

Description of the Experiments. We fix the number of servers, $s = 100$, because we are interested in large service systems. We consider nonhomogeneous Poisson arrival processes with the sinusoidal arrival rate functions in (4.12). We vary $\bar{\lambda}$ to get alternative values of ρ, for fixed s. We consider values of ρ ranging from 0.90 to 0.98. These values of ρ are chosen to let our systems alternate between periods of overload and underload. We consider two values of the relative amplitude: $\alpha = 0.1$, and $\alpha = 0.5$. Simulation point and 95% confidence interval estimates are based on 10 independent replications of 5 million events each, where an event is either an arrival or a service completion. That is, each simulation run terminates when the sum of the number of arrivals and the number of service completions is equal to 5 the appendix for more.

The parameters of the arrival-rate intensity function, $\lambda(u)$ in (4.12), should be interpreted relative to the mean service time, $E[S]$. As in §4.1.3, we measure time in units of mean service times; hence $\mu = 1$. Then, we refer to γ in (4.12) as the relative frequency. Table 4.1 displays values of the relative frequency as a function of $E[S]$, assuming a daily cycle. For interpretation, we also will specify the associated

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

Table 4.1: 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 $E[S] = 1$.

mean service time in minutes, given a daily cycle.

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

In Table 4.2, we present simulation (point and 95% confidence interval estimates) quantifying the performance of QL, HOL$_m$, and HOL in the $M(t)/GI/s$ model with M, $LN(1, 4)$, and D service-time distributions. We discuss these results next.

Comparing HOL$_m$ and HOL.

Table 4.2 shows that, for $\alpha = 0.1$ and $E[S] = 30$ minutes, HOL$_m$ performs better than HOL, particularly for high values of ρ. We get consistent results with M,

$LN(1,4)$, and D service times: ASE(HOL)/ASE(HOL$_m$) is roughly equal to 1 for $\rho = 0.9$, and roughly equal to 1.4 for $\rho = 0.98$. The case with high ρ corresponds to extreme fluctuations between phases of underload and overload, in which case HOL performs relatively poorly.

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

Comparing HOL$_m$ and QL.

In the $M(t)/M/s$ model, QL is provably the optimal predictor, under the MSE criterion; see §4.4. With $\alpha = 0.1$, $E[S] = 30$ minutes, and M service times, Table 4.2 shows that RRASE(QL) ranges from about 21% for $\rho = 0.9$ to about 10% for $\rho = 0.98$. With non-exponential service times, QL remains the most effective predictor, under the MSE criterion. It is relatively accurate, in all models considered. For example, with $\alpha = 0.5$, $E[S] = 6$ hours, and $LN(1,4)$ service times, RRASE(QL) ranges from about 20% for $\rho = 0.9$ to about 12% for $\rho = 0.98$.

Consistent with §4.5, the approximation for the ratio of the SCV's in (4.13) provides a remarkably accurate approximation for the ratio of the ASE's with M service times, particularly for high values of ρ, as we would expect. (The distortion caused by the customer average in (4.4) is evidently minor.) For example, with $E[S] = 30$ minutes and $\alpha = 0.1$, Table 4.2 shows that the relative error between simulation

point estimates for ASE(HOL_m)/ASE(QL) and numerical values given by (4.13), is less than 3% for $\rho = 0.98$.

With $LN(1, 4)$ service times, $E[S] = 30$ minutes, and $\alpha = 0.1$, Table 4.2 shows that ASE(HOL_m)/ASE(QL) ranges from about 1.7 for $\rho = 0.9$ to about 1.5 for $\rho = 0.98$, which is less than predicted by (4.13). Similarly, with D service times, $E[S] = 6$ hours, and $\alpha = 0.5$, Table 4.2 shows that ASE(HOL_m)/ASE(QL) is approximately equal to 1.5 for all ρ.

4.7 Estimating the Required Additional Information for HOL_m

We have shown, both analytically and using simulation, that the HOL predictor can perform poorly when the arrival rate varies considerably over time while the staffing is fixed. We showed that the new refined HOL predictor, HOL_m, performs remarkably better than HOL in the $M(t)/GI/s$ queueing model, with time-varying arrival rates; see §4.6.

However, the statistical accuracy of HOL_m is obtained at the expense of ease of implementation. In addition to the HOL delay, w, HOL_m depends on the arrival-rate function, $\lambda(t)$, and the mean time between successive service completions (which equals $1/s\mu$ with s simultaneously busy servers and i.i.d. exponential service times with rate μ); see (4.7). In practice, the implementation of HOL_m requires knowledge of those system parameters, which may require estimation from data. Any estimation procedure inevitably produces some estimation error, which would affect the performance of HOL_m.

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

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

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

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

Table 4.2: A comparison of the efficiency of QL, HOL_m, and HOL in the $M(t)/GI/100$ model, as a function of the traffic intensity, ρ. Point and 95% confidence interval estimates of the average squared error (ASE) are shown. Estimated ASE's are in units of 10^{-3}.

time between successive service completions in real-life service systems. Further, we quantify the estimation error resulting from those procedures, and its impact on the performance of HOL_m; see Table 4.3. We show that the HOL_m predictor remains effective even with imperfect information about system parameters.

To prediction the arrival-rate function, $\lambda(t)$, we propose relying on forecasts relying on data from previous days, and observations over the current day, up to date. For $\theta_{HOL_m}(t, w)$ in (4.7), we need estimates of the arrival-rate function over the interval $[t - w, t]$. Here, we assume that the arrival process is a nonhomogeneous Poisson process with an integrable arrival-rate function. Since we observe customer arrival times, but not the arrival rates, we need to forecast future rates based on historical call volumes. For ways of forecasting future arrival rates, we refer the reader to recent work on forecasting arrival rates to service systems such as call centers. For one example, Shen and Huang (2008b) propose an approach to forecast the time series of an inhomogeneous Poisson process by first building a factor model for the arrival rates, and then forecasting the time series of factor scores. As another example, Aldor-Noiman et al. (2009) propose an arrival count model which is based on a mixed Poisson process approach incorporating day-of-week, periodic, and exogenous effects. For other related work, see Avramidis et al. (2004), Brown et al. (2005), and Shen and Huang (2008a).

We might also rely on historical data from previous days to prediction the mean time between successive service completions, combined with real-time data over the recent past. However, we consider a procedure based on real-time estimation alone, and investigate its feasibility. As a real-time predictor, we propose computing the sample average, $\hat{m}$, of (recent) time intervals between successive service completions in the system. In doing so, as an approximation, we assume (i) that all servers are simultaneously busy, and (ii) that the times between successive service comple-

tions are i.i.d. (Since we are interested in systems which are heavily loaded, the assumption of busy servers is not too restrictive. The second assumption is exact for exponential service times, but not more generally.) Given that assumption, we can apply elementary statistics to compute the sample size, $n(x)$, needed to obtain a desired margin of relative error, x, at a given confidence level. (Specifically, the half width of a confidence interval is a function of the number of observations used. Therefore, we can obtain a desired margin of relative error by changing the number of observations, thus leading to a different half width.) The error, x, measures the relative error between the actual mean and the sample mean.

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

There remains to study the effect of the estimation error, x, on the performance of the HOL_m predictor. To that end, we consider modified HOL_m delay predictors, denoted by $\mathrm{HOL}_m(x)$, depending on the relative error, x, in estimating $1/s\mu$. That is, the $\mathrm{HOL}_m(x)$ predictors use the following delay prediction:

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

where $-1 < x < 1$. We study the performance of $\text{HOL}_m(x)$ for alternative small values of x. Clearly, the performance of $\text{HOL}_m(x)$ should degrade as $|x|$ increases, but we would like to know by how much.

In Table 4.3, we study the performance of $\text{HOL}_m(x)$ as a function of the traffic intensity, ρ, in the $M(t)/M/100$ queueing model, with $\alpha = 0.5$ and $E[S] = 5$ minutes. We also include the sample sizes needed to obtain system parameter estimates within that error margin and, in parentheses, the corresponding required length of the estimation interval (under our model assumptions). We consider values of x between -0.1 and 0.1. For these values, we find that HOL_m still performs considerably better than HOL. For example, for $x = 0.05$, the ratio ASE(HOL)/ASE($\text{HOL}_m(x)$) ranges from about 14 to about 23 for values of ρ between 0.9 and 0.98. For $x = -0.05$, ASE(HOL)/ASE($\text{HOL}_m(x)$) ranges from about 16 to about 27 for ρ between 0.9 and 0.98. That is, simulation shows that HOL_m remains remarkably more effective than HOL, even with imperfect information about system parameters, as would commonly occur in practice.

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

$M(t)/M/100$, $\alpha = 0.5$, $E[S]$ = 5 min

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

Table 4.3: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/M/100$ queueing model with $\alpha = 0.5$ and $E[S] = 5$ minutes. Sample sizes needed and length of estimation intervals required are also included.

4.8 Delay Predictors for the $M(t)/GI/s+GI$ Model

In this section, we propose a new delay predictor for the $M(t)/GI/s+GI$ model, based on the HOL delay observed upon arrival to the system. In §4.9 we show that this new predictor, HOL_a, performs remarkably well. In particular, HOL_a effectively copes with both time-varying arrivals and non-exponential abandonment-time distributions. As a frame of reference, we also consider a classical delay predictor based on the queue-length seen upon arrival to the system. This predictor, QL_m, was previously considered in chapter 3. For completeness, we now provide a short description of QL_m.

The Markovian Queue-Length-Based Delay Predictor (QL_m).

The QL_m predictor approximates the expected conditional delay, given the queue length, in the $M(t)/GI/s+GI$ model by the expected conditional delay, given the queue length, in the corresponding $M(t)/M/s+M$ model with the same service-time

and abandon-time means. For the $M(t)/M/s+M$ model, we have the representation:

$$W_Q(t,n) \equiv \sum_{i=0}^{n} Y_i \ , \tag{4.20}$$

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

$$E[W_Q(t,n)] = \sum_{i=0}^{n} E[Y_i] = \sum_{i=0}^{n} \frac{1}{s\mu + i\nu} \ . \tag{4.21}$$

The QL_m predictor given to a customer who finds n customers in queue upon arrival is $\theta_{QL_m}(t,n) \equiv E[W_Q(t,n)]$. Under the MSE criterion, QL_m is the best possible predictor in the $M(t)/M/s+M$ model, but we found that it is not always so good for the more general $M(t)/GI/s+GI$ model; see §4.9. Thus, there is a need to go beyond QL_m, in practice.

The Approximation-Based QL-Based Delay Predictor (QL_a).

In chapter 3, we introduced an approximation-based queue-length-based delay predictor, QL_a, which exploits established approximations for performance measures in the $M/GI/s+GI$ model, developed by Whitt (2005b). We showed that QL_a consistently outperforms all other predictors considered in the $GI/GI/s+GI$ model, with a stationary arrival process. Here, we propose an analog of QL_a that uses the observed HOL delay, and effectively copes with time-varying arrival rates. For completeness, we begin by briefly reviewing the QL_a predictor for the $GI/GI/s+GI$ model.

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

$$\psi_j \equiv h(j/\lambda), \quad 1 \leq j \leq k \ ; \tag{4.22}$$

k is the current queue length, λ is the arrival rate (assumed constant), and h is the abandonment-time hazard-rate function, defined as $h(t) \equiv f(t)/(1-F(t))$, $t \geq 0$, where f is the corresponding density function (assumed to exist). Here is how (4.22) 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 prediction the elapsed waiting time of that customer, given the available state information. Assuming that abandonments are relatively rare compared to service completions, it is reasonable to act as if there have been j arrival events since our customer arrived. Since a simple rough prediction for the time between successive arrival events is the reciprocal of the arrival rate, $1/\lambda$, the elapsed waiting time of is approximated by j/λ and the corresponding abandonment rate by (4.22).

For the $GI/M/s+M(n)$ model, we need to make further approximations in order to describe the potential waiting time of a customer who finds n other customers waiting in line, upon arrival. Let $W_Q(n)$ represent a random variable with the conditional distribution of the potential delay of an arriving customer, given that this customer must wait before starting service, and given that the queue-length seen

upon arrival, is equal to n. We have the approximate representation:

$$W_Q(n) \approx \sum_{i=0}^{n} X_i \ , \tag{4.23}$$

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

The QL_a delay prediction given to a customer who finds n customers in queue upon arrival is

$$\theta_{QL_a}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}} \ ; \tag{4.24}$$

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

The HOL_a Predictor.

We are now ready to propose a new delay predictor for the $M(t)/GI/s+GI$ model, which we refer to as HOL_a. This predictor requires knowledge of the abandonment-time hazard- rate function, h. That is convenient from a practical point of view,

because it is relatively easy to prediction hazard rates from system data; see Brown et al. (2005).

We proceed in two steps: (i) we use the observed HOL delay, w, to prediction the queue length seen upon arrival, and (ii) we use this queue-length prediction to implement a new delay predictor, paralleling (4.24). Unlike QL_a, HOL_a exploits the HOL delay, and does not assume knowledge of the queue length seen upon arrival.

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

$$m(t, w) \equiv E[N_w(t)] = \int_{t-w}^{t} \lambda(s)(1 - F(t-s))ds \ , \tag{4.25}$$

where F is the abandonment-time cdf.

For step (ii), we use $m(t, w)+1$ as an prediction of the queue length seen upon arrival, at time t. In (4.22), we replace λ by $\hat{\lambda}$, where $\hat{\lambda}$ is defined as the average arrival rate over the interval $[t-w, t]$, i.e., $\hat{\lambda} \equiv (1/w)\int_{t-w}^{t} \lambda(s)ds$. We do so because approximating the arrival process we now have a nonstationary arrival process instead of a stationary arrival process. Paralleling (4.24), 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:

$$\theta_{HOL_a}(t,w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s\mu + \hat{\delta}_n - \hat{\delta}_{n-i}} , \tag{4.26}$$

for $m(t,w)$ in (4.25), $\hat{\delta_k} = \sum_{j=1}^{k} h(j/\hat{\lambda})$, and $\hat{\delta_0} = 0$. If we actually know the queue length, then we can replace $m(t,w)$ by $Q(t)$, i.e., we can use QL_a. There remains to investigate ways of estimating the abandonment-time distribution needed to implement QL_h. We envision that such estimates will be based on long-term estimates of customer time-to-abandon distribution, instead of real-time information about customer abandonment times. Providing additional details relating to this estimation is outside the scope of this thesis, and is left for future research.

4.9 Simulation Results for the $M(t)/M/s+GI$ Model

In this section, we present simulation results for the $M(t)/M/s+GI$ model, with sinusoidal arrival rates. For the abandonment-time distribution, we considered M (exponential), H_2 (hyperexponential with SCV equal to 4 and balanced means), and E_{10} (Erlang, sum of 10 exponentials). We consider non-exponential service-time distributions in the appendix. In this section, we show plots of the simulation results and tables with estimates of 95% confidence intervals.

Description of the Experiments.

We vary the number of servers, s, but consider only relatively large values ($s \geq 100$), because we are interested in large service systems. We let the service rate, μ, be equal to 1. For the arrival rate function, $\lambda(u)$ in (4.12), we fix the relative frequency, $\gamma = 1.571$. This value of γ corresponds to a mean service time $E[S] = 6$ hours, for

daily arrival-rate cycles; see Table 4.1.

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

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

Consistent with theory in §4.8, Table 4.4 and Figure 4.5 shows that QL_m is the best possible predictor, under the MSE criterion. The RRASE of QL_m ranges from about 14% for $s = 100$ to about 4% when $s = 1000$. Figure 4.6 shows that $s \times \mathrm{ASE}(\mathrm{QL}_m)$, the ASE of QL_m multiplied by the number of servers s, is nearly constant for all values of s considered. This shows that QL_m is asymptotically correct as s increases, i.e., $\mathrm{ASE}(\mathrm{QL}_m)$ approaches 0 as s increases.

Table 4.4 shows that HOL_a is the second best predictor for this model. The RRASE of HOL_a ranges from about 20% for $s = 100$ to about 6% for $s = 1000$. That is, HOL_a is relatively accurate for this model. The difference in performance between HOL_a and QL_m is not too great: $\mathrm{ASE}(\mathrm{HOL}_a)/\mathrm{ASE}(\mathrm{QL}_m)$ is close to 1.6, for all s. Moreover, Figure 4.6 shows that HOL_a is asymptotically correct: $s \times \mathrm{ASE}(\mathrm{HOL}_a)$ is also roughly equal to a constant, for all s.

The HOL predictor performs much worse than QL_m and HOL_a. For example, the ratio $\mathrm{ASE}(\mathrm{HOL})/\mathrm{ASE}(\mathrm{HOL}_a)$ 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 a considerable improvement in the

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/M/s+M$ Model

s	QL_m	HOL_a	HOL
100	3.059×10^{-3}	5.556×10^{-3}	1.623×10^{-2}
	$\pm1.95\times10^{-4}$	$\pm4.23\times10^{-4}$	$\pm9.52\times10^{-4}$
300	9.911×10^{-4}	1.630×10^{-3}	1.114×10^{-2}
	$\pm7.07\times10^{-5}$	$\pm1.43\times10^{-4}$	$\pm4.33\times10^{-4}$
500	5.474×10^{-4}	9.653×10^{-4}	1.033×10^{-2}
	$\pm4.42\times10^{-5}$	$\pm6.37\times10^{-5}$	$\pm2.34\times10^{-4}$
700	4.076×10^{-4}	6.780×10^{-4}	9.866×10^{-3}
	$\pm2.08\times10^{-5}$	$\pm3.09\times10^{-5}$	$\pm2.26\times10^{-4}$
1000	2.853×10^{-4}	4.907×10^{-4}	9.806×10^{-3}
	$\pm2.48\times10^{-5}$	$\pm1.90\times10^{-5}$	$\pm1.75\times10^{-4}$

Table 4.4: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

performance of HOL, as s increases. That is confirmed by Figure 4.6, where we see that $s \times$ ASE(HOL) increases linearly, as s increases.

4.9.1 Results for the $M(t)/M/s+H_2$ model

With H_2 abandonment, Table 4.5 and Figure 4.7 shows that HOL_a is the best possible predictor, under the MSE criterion, for large values of s. In particular, HOL_a outperforms QL_m for $s \geq 300$. The ratio ASE(QL_m)/ASE(HOL_a) ranges from about 0.9 for $s = 100$ to about 3 for $s = 1000$. The RRASE of HOL_a ranges from about 20% for $s = 100$ to about 6% for $s = 1000$. However, the QL_m predictor remains relatively accurate for this model: RRASE(QL_m) ranges from about 20% for $s = 100$ to about 11% for $s = 1000$.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/M/s + H_2$ Model

s	QL_m	HOL_a	HOL
100	3.166×10^{-3}	3.710×10^{-3}	7.951×10^{-3}
	$\pm1.52\times10^{-4}$	$\pm1.77\times10^{-4}$	$\pm5.38\times10^{-4}$
300	1.488×10^{-3}	1.148×10^{-3}	4.768×10^{-3}
	$\pm5.61\times10^{-5}$	$\pm7.93\times10^{-5}$	$\pm2.16\times10^{-4}$
500	1.192×10^{-3}	7.139×10^{-4}	4.227×10^{-3}
	$\pm5.26\times10^{-5}$	$\pm4.74\times10^{-5}$	$\pm1.85\times10^{-4}$
700	1.067×10^{-3}	5.180×10^{-4}	3.960×10^{-3}
	$\pm4.18\times10^{-5}$	$\pm2.92\times10^{-5}$	$\pm1.31\times10^{-4}$
1000	9.590×10^{-4}	3.363×10^{-4}	3.827×10^{-3}
	$\pm2.24\times10^{-5}$	$\pm1.86\times10^{-5}$	$\pm5.75\times10^{-5}$

Table 4.5: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

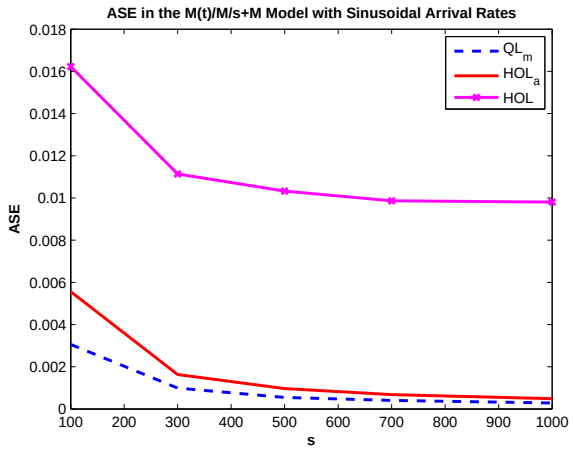

Figure 4.5: ASE in the $M(t)/M/s + M$ model, $E[S] = 6$ hours, $\alpha = 0.5$

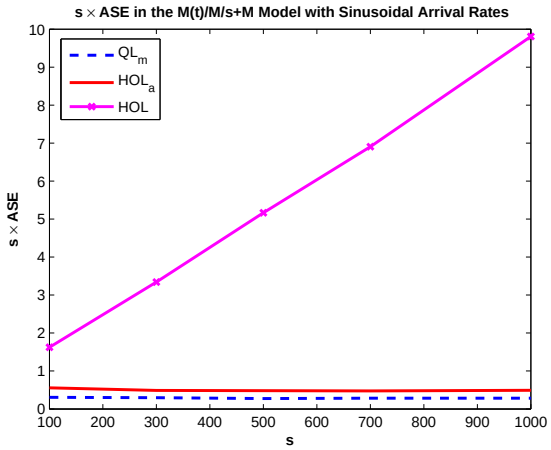

Figure 4.6: $s \times$ ASE in the $M(t)/M/s + M$ model, $E[S] = 6$ hours, $\alpha = 0.5$

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/M/s + E_{10}$ Model

s	QL_m	HOL_a	HOL
100	9.542×10^{-3}	6.481×10^{-3}	4.531×10^{-2}
	$\pm4.26\times10^{-4}$	$\pm2.80\times10^{-4}$	$\pm1.79\times10^{-3}$
300	7.711×10^{-3}	2.551×10^{-3}	3.744×10^{-2}
	$\pm2.94\times10^{-4}$	$\pm9.64\times10^{-5}$	$\pm9.55\times10^{-4}$
500	7.083×10^{-3}	1.666×10^{-3}	3.652×10^{-2}
	$\pm2.63\times10^{-4}$	$\pm8.36\times10^{-5}$	$\pm8.56\times10^{-4}$
700	6.875×10^{-3}	1.360×10^{-3}	3.622×10^{-2}
	$\pm1.73\times10^{-4}$	$\pm6.38\times10^{-5}$	$\pm5.22\times10^{-4}$
1000	6.858×10^{-3}	1.070×10^{-3}	3.582×10^{-2}
	$\pm1.17\times10^{-4}$	$\pm4.27\times10^{-5}$	$\pm6.03\times10^{-4}$

Table 4.6: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

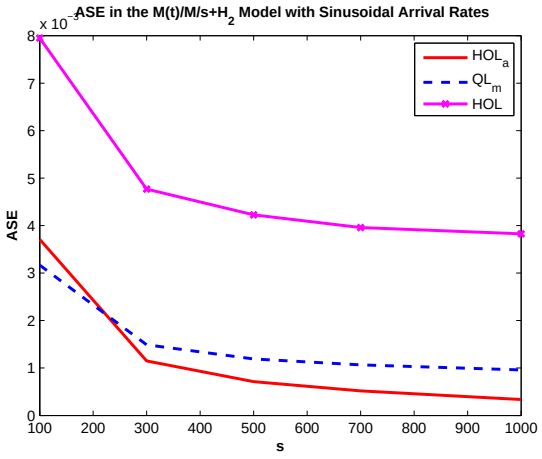

Figure 4.7: ASE in the $M(t)/M/s + H_2$ model, $E[S] = 6$ hours, $\alpha = 0.5$

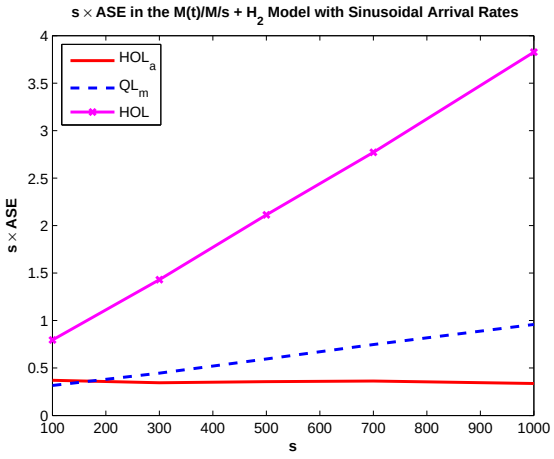

Figure 4.8: $s \times$ ASE in the $M(t)/M/s + H_2$ model, $E[S] = 6$ hours, $\alpha = 0.5$

The difference in performance between QL_m and HOL_a is particularly significant as the number of servers, s, increases. Figure 4.8 also shows that HOL_a is asymptotically correct as s increases: $s \times \text{ASE}(\text{HOL}_a)$ is roughly constant for all s. In contrast, $s \times \text{ASE}(\text{QL}_m)$ increases roughly linearly, as s increases, which shows that the performance of QL_m deteriorates as s increases.

Figure 4.7 shows that HOL is, once more, the least effective predictor for this model. The RRASE of HOL ranges from about 31% for $s = 100$ to about 22% for $s = 1000$. The HOL predictor performs slightly better than with M abandonment: $\text{ASE}(\text{HOL})/\text{ASE}(\text{HOL}_a)$ ranges from about 2 for $s = 100$ to about 11 for $s = 1000$. Once more, we do not see an improvement in the performance of HOL, as s increases: Figure 4.8 shows that $s \times \text{ASE}(\text{HOL})$ increases roughly linearly as s increases. The slope of the $s \times \text{ASE}(\text{HOL})$ curve is substantially greater than that of the $s \times \text{ASE}(\text{QL}_m)$ curve.

Results for the $M(t)/M/s + E_{10}$ model.

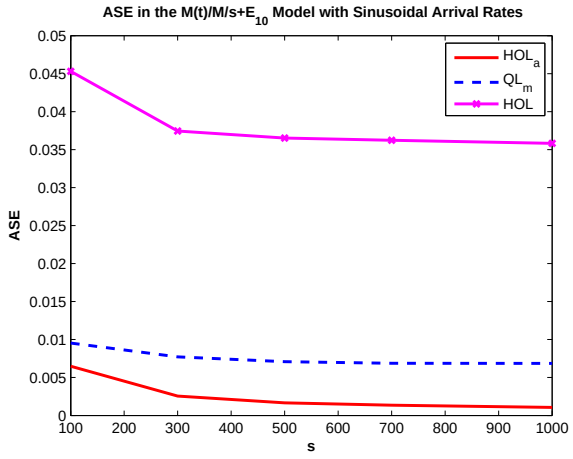

Figure 4.9: ASE in the $M(t)/M/s + E_{10}$ model, $E[S] = 6$ hours, $\alpha = 0.5$

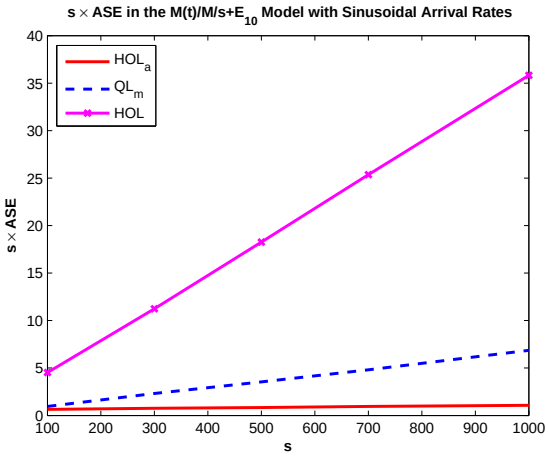

Figure 4.10: $s \times$ ASE in the $M(t)/M/s+E_{10}$ model, $E[S] = 6$ hours, $\alpha = 0.5$

Table 4.6 and Figure 4.9 shows that HOL_a is the most effective predictor, under the MSE criterion, for this model. 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 4.10 shows that HOL_a is asymptotically correct: $s \times \text{ASE}(\text{HOL}_a)$ is roughly equal to a constant for all values of s considered.

The QL_m predictor performs significantly worse than HOL_a, with E_{10} abandonment. The ratio $\text{ASE}(\text{QL}_m)/\text{ASE}(\text{HOL}_a)$ ranges from about 1.5 for $s = 100$ to about 6.5 for $s = 1000$. The RRASE of QL_m ranges from about 13% for $s = 100$ to about 10% for $s = 1000$. Figure 4.10 shows that QL_m is not asymptotically correct as s increases.

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: $\text{ASE}(\text{HOL})/\text{ASE}(\text{HOL}_a)$ ranges from roughly 7 for $s = 100$ to roughly 33 for $s = 1000$. Figure 4.6 shows that $s \times \text{ASE}(\text{HOL})$ increases linearly (and steeply) as s increases.

4.10 Concluding Remarks

In this chapter, we studied the performance of alternative delay-history-based delay predictors in the $M(t)/GI/s$ and $M(t)/GI/s + GI$ queueing models, which have a nonhomogeneous Poisson process. A main conclusion is that the performance of these delay-history-based delay predictors can degrade in face of time-varying arrivals, which often occurs in practice; that is dramatically shown in Figure 4.2.

As a consequence, we developed refinements of HOL, in particular, the HOL_m predictor in (4.7) for the $M(t)/GI/s$ model and the HOL_a predictor in (4.26) for the $M(t)/GI/s+GI$ model. Through simulation experiments, we showed that these proposed predictors effectively cope with both time-varying arrivals and non-exponential service-time and abandon-time distributions. We also established analytical results supporting HOL_m in §4.5. In particular, we quantified the difference in performance between QL and HOL_m and found that the ratio of their respective MSE's is roughly equal to 2, particularly for high values of the traffic intensity, ρ; see (4.13).

However, the new predictors lose some of their appeal compared to the simple HOL and LES predictors, because they require information about the model, in particular, the arrival-rate function and the mean time between successive departures. Hence, in §4.7 we proposed ways to estimate the required information. Even if we rely on real-time estimation of the mean time between successive departures, we showed that we can obtain suitably accurate estimates without requiring that the observation interval be too long. Table 4.3 showed that the HOL_m predictor remains effective even if the information is known imperfectly.

Our general strategy for creating the refined HOL predictors has been to approximate the mean conditional delay, given the observed HOL delay by (i) approximating the queue length, given the observed HOL delay, and (ii) approximating the expected

delay given the queue length. As a consequence, direct queue-length-based delay predictors would be preferred if the queue length is known. We would use QL, studied in chapter 2, instead of HOL_m without abandonment, and we would use QL_a, studied in chapter 3, instead of HOL_a with customer abandonment. However, in the introduction we observed that there are complex service systems such as Web chat and ticket queues for which the queue length is not known.

5

Time-Varying Demand and Capacity

5.1 Introduction

In this chapter, 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. We develop four new predictors, two of which exploit an established deterministic fluid approximation for a many-server queueing model with those features. This chapter is an edited version of Ibrahim and Whitt (2010b).

5.1.1 Recap of Previous Chapters

For completeness, we now briefly review relevant results from earlier chapters. We started with the $GI/M/s$ model (chapter 2), and extended to the $GI/GI/s + GI$

model (chapter 3). We showed that standard queue-length-based predictors, which are commonly used in practice, may perform poorly. We proposed new, more accurate, queue-length-based predictors that effectively cope with non-exponential 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 (2005b): it approximates the $GI/GI/s + GI$ model by the corresponding $GI/M/s + M(n)$ model, with state-dependent Markovian abandonment rates; see §5.3. Since QL_a assumes a stationary arrival process and a constant number of servers, it may 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 (chapter 4) with time-varying arrival rates and a constant number of servers. We focused on the HOL delay predictor. We showed that HOL may 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 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 §5.3. 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 may perform poorly when the number of servers varies significantly over time. Therefore, there is a need to go beyond HOL_a.

5.1.2 Main Contributions

In this chapter, we consider the $M(t)/M/s(t) + GI$ model, which we describe in §5.2. Since direct analysis of customer delay is complicated in this model, we propose two different approaches: (i) in §5.3, we propose modified versions of QL_a and HOL_a to account for a time-varying number of servers, and (ii) in §5.5, 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 simulation results show that those new predictors have a superior performance in the $M(t)/M/s(t) + GI$ model.

In Figure 5.1, 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 leading to significant delays. We assume that $\lambda(t)$ and $s(t)$ have a period equal to 4 times the mean service time; see §5.6.1. (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 5.1.

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 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 5.1, 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_{rt}, and a refined queue-length-based predictor, QL_{rt}. The HOL_{rt} and QL_{rt} predictors are based on the fluid model in Liu and Whitt (2010). (Subscript "t" indicates that those predictors are based on the fluid model with time-varying arrivals; that is to distinguish them from the refined predictors of chapter 3 which are based on the fluid model with a stationary arrival process.) Figure 5.1 nicely illustrates the improvement in performance resulting from our proposed refinements: We plot simulation estimates of the average differences between HOL_{rt} (QL_{rt}) delay predictions and actual delays observed in the system, as a function of time (solid curves).

5.1.3 Organization of the Chapter

The rest of this chapter is organized as follows: In §5.2, we describe our general framework. In §5.3, we briefly describe the QL_a and HOL_a predictors, considered in §5.1, and propose modified predictors, QL_a^m and HOL_a^m, that cope with a time-varying number of servers. In §5.4, we review a deterministic fluid model, developed in Liu and Whitt (2010), for multiserver queues with time-varying arrival rates and customer abandonment. In §5.5, we use these fluid approximations to develop new, refined, delay predictors. In §5.6, we present simulation results showing that these

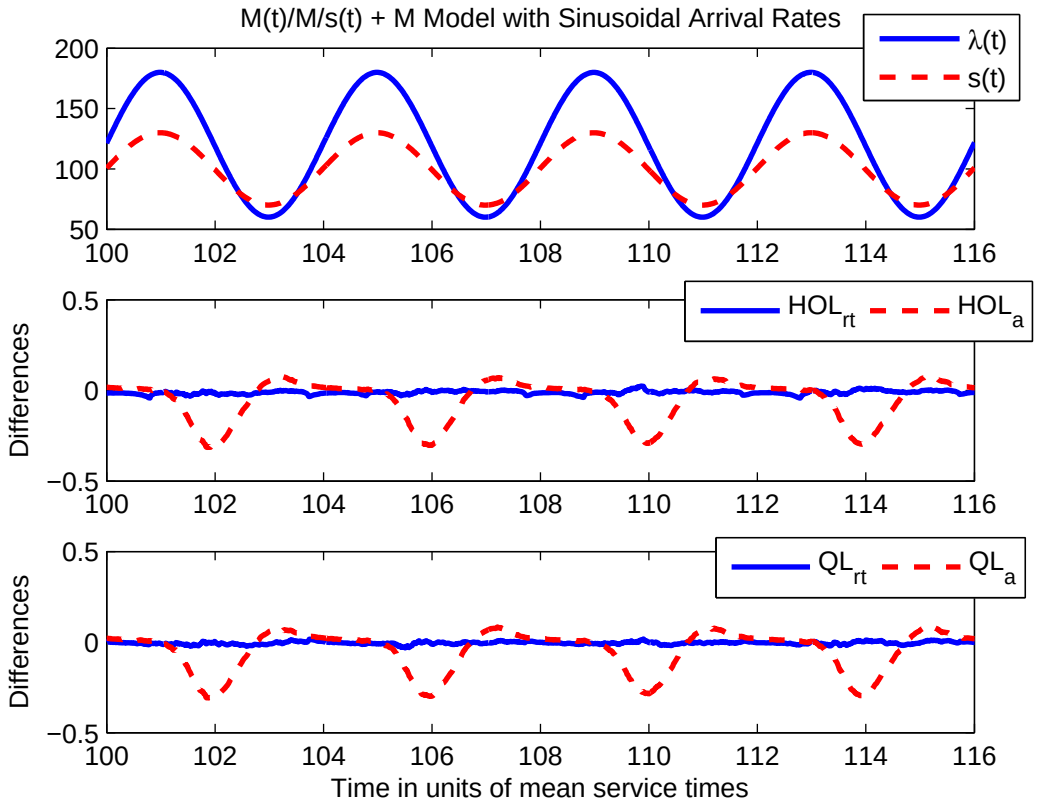

Figure 5.1: Bias of standard and refined delay predictors in the $M(t)/M/s(t) + M$ model with sinusoidal arrival rates (for model in §6.1). The differences between delay predictions and actual (potential) delays observed are based on averaging 100 independent simulation replications.

new predictors are effective in the $M(t)/M/s(t)+M$ model. We make concluding remarks in §5.7. In the appendix, we describe additional simulation results. In particular, we consider the $M(t)/M/s(t)+GI$ model with non-exponential abandonment-time distributions, and the $M(t)/GI/s(t)+GI$ model with non-exponential service and abandonment-time distributions. Finally, in the appendix, we also propose a simple modified QL_a-based delay predictor, QL_a^{sm}, and study its performance in the $M(t)/M/s(t)+M$ model.

5.2 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.

5.2.1 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 $\lambda \equiv \{\lambda(u) : -\infty < u < \infty\}$. 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: $s \equiv \{s(u) : -\infty < u < \infty\}$.

5.2.2 Performance measures

For completeness, we now indicate how we evaluate the performance of our candidate delay predictors. Once more, we use computer simulation to do the actual estimation.

5.2.2.1 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:

$$ASE \equiv \frac{1}{k}\sum_{i=1}^{k}(p_i - a_i)^2 \ , \tag{5.1}$$

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 previous chapters, 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.

5.2.2.2 Weighted Mean Squared Error (WMSE).

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:

$$MSE(\theta_{QL}(t, n)) \equiv E[(W_{QL}(t, n) - \theta_{QL}(t, n))^2] \ , \tag{5.2}$$

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.,

$$MSE(t) \equiv E[MSE(\theta_{QL}(t, Q(t)))] \ . \tag{5.3}$$

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

$$WMSE \equiv \frac{\int_0^T \lambda(t) MSE(t) dt}{\int_0^T \lambda(t) dt} \ . \tag{5.4}$$

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

5.3 Modified Delay Predictors: QL_a^m and HOL_a^m

Figure 5.1 shows that QL_a and HOL_a may 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 §5.6, 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.

5.3.1 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:

$$W_{QL}(t, n) \equiv \sum_{i=0}^{n} Y_i \ , \tag{5.5}$$

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 (2005b). 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 §3 of Whitt (2005b). 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

$$\psi_j \equiv h(j/\lambda), \quad 1 \le j \le k \ ; \tag{5.6}$$

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 $t \ge 0$, where f is the

corresponding density function (assumed to exist).

Here is how (5.6) 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 (5.6).

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 $\delta_k = \sum_{j=1}^{k} \psi_j = \sum_{j=1}^{k} h(j/\hat{\lambda})$, $k \geq 1$, 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

$$\theta_{QL_a}(n) = \sum_{i=0}^{n} \frac{1}{s\mu + \delta_n - \delta_{n-i}} \ ; \tag{5.7}$$

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 (5.7) 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 (5.7).

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

$$m(t, w) \equiv E[N_w(t)] = \int_{t-w}^{t} \lambda(s)(1 - F(t-s))ds \ , \tag{5.8}$$

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 (5.7), 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:

$$\theta_{HOL_a}(t, w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s\mu + \delta_n - \delta_{n-i}} , \tag{5.9}$$

for $m(t, w)$ in (5.8). 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 (5.9) by $\bar{s}$.

5.3.2 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 (5.7) 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:

$$\theta_{QL_a^m}(t, n) = \sum_{i=0}^{n} \frac{1}{s(t_{i+1})\mu + \delta_n - \delta_{n-i}} , \tag{5.10}$$

where

$$t_i = t_{i+1} + \frac{1}{s(t_{i+1})\mu + \delta_n - \delta_{n-i}} \text{ for } 0 \le i \le n \quad , \tag{5.11}$$

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

$$\theta_{HOL_a^m}(t, w) \equiv \sum_{i=0}^{m(t,w)+1} \frac{1}{s(t_{i+1})\mu + \delta_n - \delta_{n-i}} , \tag{5.12}$$

where t_i is given by (5.11) 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 §5.5, we take a different approach and propose new delay predictors based on fluid approximations, which we now review.

5.4 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 $x > w(t)$:

$$w(t) = \inf\{x > 0 : q(t,y) = 0 \text{ for all } y > x\} \ . \tag{5.13}$$

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 non-zero initial queue content, see §5 of Liu and Whitt (2010).

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:

$$q(t+u, x+u) = q(t,x)\frac{\bar{F}(x+u)}{\bar{F}(x)} \ , 0 \le x \le w(t) \ , \text{ and,} \tag{5.14}$$

$$b(t+u, x+u) = b(t,x)\frac{\bar{G}(x+u)}{\bar{G}(x)} \ . \tag{5.15}$$

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

$$Q_f(t) = \int_0^{w(t)} q(t,y)dy = \int_0^{w(t)} \lambda(t-x)\bar{F}(x)dx \ , \tag{5.16}$$

where we use (5.14) 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

$$v(t - w(t)) = w(t) \ . \tag{5.17}$$

Therefore, for a given feasible boundary waiting time process, $\{w(t) : t \geq 0\}$, we can determine the associated potential waiting time process, $\{v(t) : t \geq 0\}$, using (5.17).

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):

$$w'(t) = 1 - \frac{b(t, 0)}{q(t, w(t))} \ , \tag{5.18}$$

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

$$s(t)\mu + s'(t) \geq 0 \text{ for all } t \ . \tag{5.19}$$

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

exponential service times,

$$w'(t) = 1 - \frac{s(t)\mu + s'(t)}{\lambda(t - w(t))\bar{F}(w(t))} \ . \tag{5.20}$$

Note that (5.20) 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 (5.14) and (5.15), together with (5.20), 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, 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

$$B_f(t) = B_f(t_0)e^{-\mu(t-t_0)} + \int_{t_0}^{t} \lambda(t-x)e^{-\mu x}dx \ . \tag{5.21}$$

The first term in (5.21) 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: $t_1 = \inf\{t > 0 : B_f(t) \geq s(t)\}$, for $B_f(t)$ in (5.21). 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 (5.20). 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 (5.17), and $Q_f(t)$ via (5.16), for

all t.

Liu and Whitt (2010) also treat the case of non-exponential 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 solely depends 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 Theorem 5.1 of Liu and Whitt (2010), 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.

5.5 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.

5.5.1 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. Since we do not have a convenient form for the mean, we use the fluid model of §5.4 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

$$\theta_{NIF}(t_0) \equiv v(t_0) \ , \tag{5.22}$$

where $v(t_0)$ is the fluid approximation for the potential waiting time, at t_0. To compute $v(t_0)$, we use (5.17) and proceed as described in §5.4. 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.

5.5.2 The Refined-Queue-Length-Based (QL_{rt}) Delay Predictor

We now propose a predictor based on the queue length seen upon arrival to the system. Let QL_{rt} refer to this refined-queue-length-based predictor. The derivation of QL_{rt} is based on that of the simple queue-length-based predictor, QL_s, which was considered in chapter 3. (Note that QL_{rt} is similar to QL_r, of chapter 3, which was based on the fluid model for a multiserver queue with a stationary arrival process and a constant number of servers.) 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

$$\theta_{QL}(t, n) = \frac{n+1}{s(t)\mu} \ . \tag{5.23}$$

In chapter 2, we showed 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 (5.23), and its lack of predictive power with customer abandonment, we propose a simple refinement of QL_s, QL_{rt}, which makes use of the fluid model in §5.4. 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 (5.16). As a result, $\mathrm{QL}_{s,f}$ 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 (5.17). For QL_{rt}, we propose computing the ratio

$$\beta(t) = v(t)/((Q_f(t)+1)/s(t)\mu) = v(t)s(t)\mu/(Q_f(t)+1) \ , \tag{5.24}$$

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:

$$\theta_{QL_{rt}}(t,n) \equiv \beta(t) \times \theta_{QL}(t,n) = v(t) \times \frac{n+1}{Q_f(t)+1} \ , \tag{5.25}$$

for $\beta(t)$ in (5.24). It is significant that $\theta_{QL_{rt}}$ 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.

5.5.3 The Refined HOL (HOL_{rt}) 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 (5.13). 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_{rt} denote this refined HOL delay predictor. The delay prediction, as a function of w and the time of arrival t, is defined as

$$\theta_{HOL_{rt}}(t, w) \equiv \frac{v(t)}{w(t)} \times \theta_{HOL}(t, w) = \frac{v(t)}{w(t)} \times w \ . \tag{5.26}$$

The QL_{rt} and HOL_{rt} predictors reduce to the $GI/GI/s + GI$ model, considered in chapter 3, so that we have "version consistency", as with QL_a^m and HOL_a^m.

5.6 Simulation Experiments for the $M(t)/M/s(t)+M$ Model

In this section, we describe simulation results quantifying the performance of all candidate delay predictors in the $M(t)/M/s(t) + M$ queueing model. Our methods apply to general time-varying functions. To illustrate, we consider sinusoidal functions which are similar to what is observed with daily cycles. Additional simulation results with H_2 (hyperexponential) and E_{10} (Erlang) abandonment-time distributions are described in the appendix. We first vary the number of servers (from tens to

hundreds) while holding all other system parameters fixed; see Figures 5.2 and 5.3. We then vary the frequency of the arrival process (from slow variation to fast) while holding all other system parameters fixed; see Table 5.2.

5.6.1 Description of the Experiments

We consider a sinusoidal arrival-rate intensity function given by

$$\lambda(u) \equiv \bar{\lambda} + \bar{\lambda}\alpha_a \sin(\gamma_a u), \quad -\infty < u < \infty \ , \tag{5.27}$$

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. (1993b), the parameters of $\lambda(u)$ in (5.27) 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 5.1 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 5.1 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

$$s(t) = \bar{s} + \bar{s}\alpha_s \sin(\gamma_s t) \ , \tag{5.28}$$

where $\bar{s}$ is the average number of servers. As in (5.27), γ_s is the frequency and α_s is the amplitude.

In this section, we let $\alpha_a = 0.5$ and $\alpha_s = 0.3$. That is, we assume that $\lambda(t)$

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

Table 5.1: The relative frequency, γ, as a function of the mean service time, $E[S]$, for a daily (24 hour) cycle.

fluctuates more extremely than $s(t)$. 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 chapter, 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. (1993a, b), 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 the less ideal case where the service provider has limited ability to respond to unexpected demand fluctuations. In that setting, (i) customers may experience significant delays which motivates 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 com-

puting the *root relative average squared error* (RRASE), defined by

$$RRASE \equiv \frac{\sqrt{ASE}}{(1/k)\sum_{i=1}^{k} p_i} , \tag{5.29}$$

using the same notation as in (5.1). The denominator in (5.29) 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 a few months each (depending on the model), assuming a 24 hour cycle.

5.6.2 Simulation Results

5.6.2.1 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 5.1. We consider this relatively large value of $E[S]$ to describe the experience of waiting patients in hospital ED's where treatment times are typically long (hours or even days in some cases). We study the impact of changing $E[S]$ in §5.6.2.2. We study the performance of our predictors as a function of $\bar{s}$. In particular, we let $\bar{s}$ range from 10 to 1000. 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 (5.28) 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 $\bar{s} = 1000$, $s(t)$ fluctuates

between 700 and 1300.

In this section, we present plots of $\bar{s} \times$ ASE (the average number of servers times the ASE) of the candidate predictors as a function of $\bar{s}$; see Figures 5.2 and 5.3. 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 §5.6.2.2). We present corresponding tables with estimates (for all predictors) of the 95% confidence intervals in the appendix.

Overview of performance as a function of $\bar{s}$.

From §2.4 of chapter 2, and §3.5 of chapter 3, 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 5.2 and 5.3 show that, for fluid-based predictors, $\bar{s} \times$ ASE is roughly constant, particularly for large $\bar{s}$. This means that the ASE of fluid-based predictors is inversely proportional to $\bar{s}$, and thus converges to 0 in large systems. For example, ASE(QL_{rt}) ranges from about 0.1 for $\bar{s} = 10$ to about 7×10^{-4} for $\bar{s} = 1000$. That is, fluid-based predictors are *asymptotically correct*. Additionally, the ratio ASE(HOL_{rt})/ASE(QL_{rt}) is roughly equal to a constant (equal to 1.3), particularly for large $\bar{s}$. Figures 5.2 and 5.3 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 (non-zero) constant for large systems.

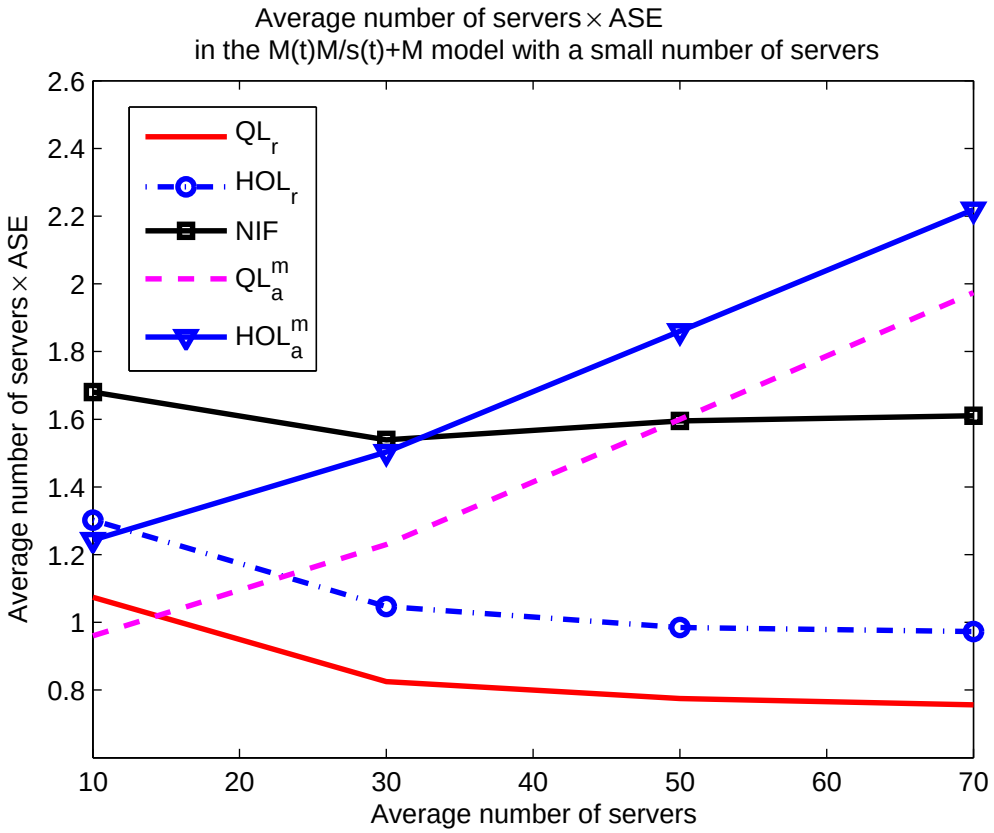

Figure 5.2: ASE of the alternative predictors in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a small average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S]$ = 6 hours with a 24 hour arrival-rate cycle.

Additionally, Figures 5.2 and 5.3 show that the ASE's of all delay predictors decrease as $\bar{s}$ increases. For example, the ASE of QL_{rt} decreases by a factor of 150 in going from $\bar{s} = 10$ to $\bar{s} = 1000$. (That is not surprising since the fluid model is a remarkably accurate approximation of large systems.) Moreover, the RRASE's 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 $\bar{s} = 1000$. (Note that QL_a^m is not a very accurate

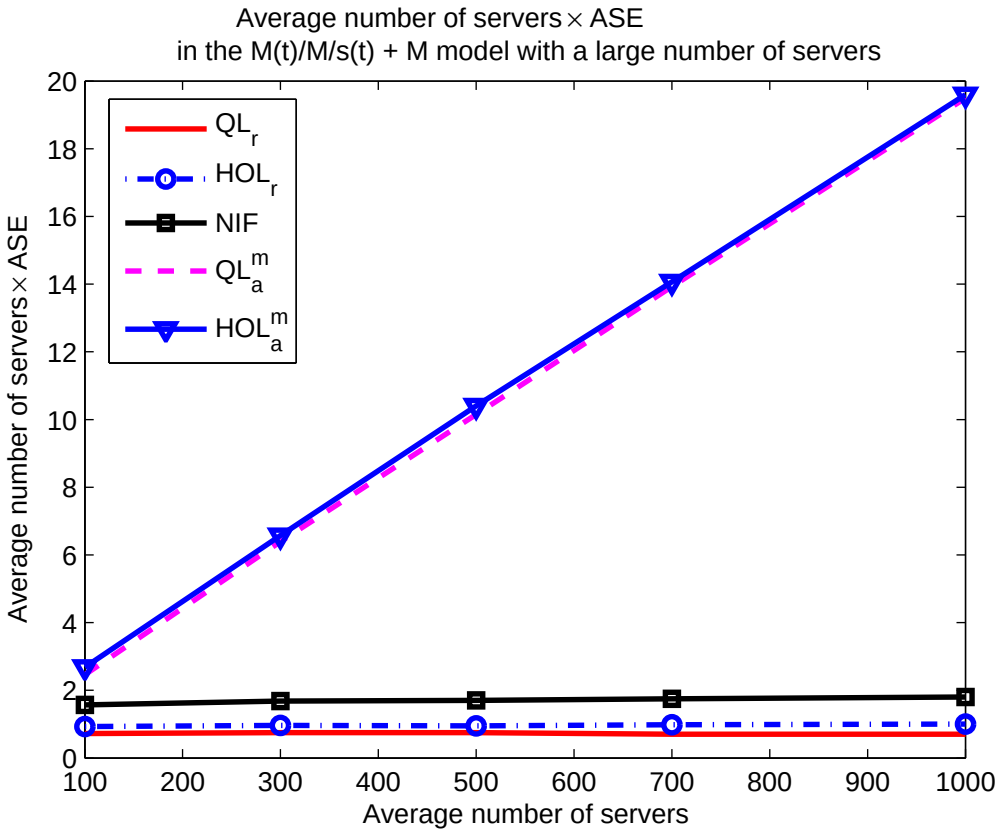

Figure 5.3: ASE of the alternative predictors in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a large average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S] = 6$ hours with a 24 hour arrival-rate cycle.

predictor in this model, even when the number of servers is large.) Although all predictors perform better in large systems, the corresponding ASE's decrease at different rates. Indeed, Figure 5.2 and 5.3 clearly show the superiority of fluid-based predictors (i.e., QL_{rt}, HOL_{rt}, 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 ASE's. For small values of $\bar{s}$, Figure 5.2 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_{rt}) 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 5.2 and 5.3 show that QL_{rt} and HOL_{rt} are more accurate than the rest of the predictors for $\bar{s} > 30$ (with QL_{rt} being the most accurate predictor). For example, the RRASE of QL_{rt} decreases from roughly 67% for $\bar{s} = 10$ to roughly 8% for $\bar{s} = 1000$. The NIF predictor is competitive for $\bar{s} \geq 50$. Indeed, the RRASE of NIF ranges from about 84% for $\bar{s} = 10$ to about 12% for $\bar{s} = 1000$. 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 $\bar{s} = 1000$. They are both significantly outperformed by fluid-based predictors. Indeed, ASE(QL_a^m)/ASE(QL_{rt}) ranges from about 0.9 for $\bar{s} = 10$ to about 27 for $\bar{s} = 1000$. Also, ASE(QL_a^m)/ASE(NIF) ranges from about 0.6 for $\bar{s} = 10$ to about 11 for $\bar{s} = 1000$.

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_{rt})/ASE(NIF) ranges from about 1.5 for $\bar{s} = 10$ to about 2.5 for $\bar{s} = 1000$. Also, ASE(QL_{rt})/ASE(NIF) ranges from about 1.3 for $\bar{s} = 10$ to about 1.8 for $\bar{s} = 1000$. These ratios are even greater for smaller values of $E[S]$; see §5.6.2.2.

ASE of the predictors in the $M(t)/M/s(t)+M$ model as a function of $E[S]$

$E[S]$	QL_{rt}	HOL_{rt}	NIF	QI_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} $\pm5.6\times10^{-4}$

Table 5.2: Performance of the alternative predictors, as a function of $E[S]$, in the $M(t)/M/s(t)+M$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\bar{s}=100$. Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

5.6.2.2 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 5.1. 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 (5.27) and (5.28), 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 which use $\lambda(t)$ and $s(t)$ corresponding to each point in time, such as QL_a^m and HOL_a^m (see (5.10) and (5.12)), will be accurate for small $E[S]$.

Table 5.2 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, the ratio $ASE(HOL_a)/ASE(HOL_a^m)$ 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, the value of $RRASE(HOL_{rt})$ 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 5.2 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 ASE's. The QL_a^m predictor is the most accurate predictor for small $E[S]$, slightly outperforming QL_{rt} (which is the second most accurate predictor in that case). Indeed, Table 5.2 shows that $ASE(QL_{rt})/ASE(QL_a^m)$ 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 since QL_a^m exploits additional information about the queue length seen upon arrival, unlike HOL_a^m.

For $E[S] \geq 2$ hours, however, QL_{rt} is more accurate than QL_a^m (and all remaining predictors); e.g., $\text{ASE}(\text{QL}_{rt})/\text{ASE}(\text{QL}_a^m)$ is roughly equal to 0.85 for $E[S] = 6$ hours. In larger systems, QL_{rt} is more accurate than QL_a^m for even smaller $E[S]$. For example, with $\bar{s} = 1000$, $\text{ASE}(\text{QL}_a^m)$ is slightly larger than $\text{ASE}(\text{QL}_{rt})$ for $E[S] = 30$ minutes, and $\text{ASE}(\text{QL}_a^m)/\text{ASE}(\text{QL}_{rt})$ is roughly equal to 4.2 for $E[S] = 2$ hours.

The QL_a^m and HOL_a^m predictors both make systematic errors which cause their ASE's to increase dramatically with $E[S]$. They are, therefore, significantly less accurate than fluid-based predictors for large $E[S]$. For example, $\text{RRASE}(\text{QL}_a)$ ranges from about 27% for $E[S] = 5$ minutes to about 52% for $E[S] = 6$ hours, whereas $\text{RRASE}(\text{QL}_{rt})$ ranges from about 20% for $E[S] = 5$ minutes to about 25% for $E[S] = 6$ hours. Also, $\text{RRASE}(\text{HOL}_a^m)$ ranges from about 22% for $E[S] = 5$ minutes to about 49% for $E[S] = 6$ hours, whereas $\text{RRASE}(\text{HOL}_{rt})$ ranges from about 25% for $E[S] = 5$ minutes to about 29% for $E[S] = 6$ hours. Additionally, Table 5.2 shows that $\text{ASE}(\text{QL}_a^m)/\text{ASE}(\text{QL}_{rt})$ ranges from roughly 0.8 for $E[S] = 5$ minutes to roughly 3.4 for $E[S] = 6$ hours, and $\text{ASE}(\text{HOL}_a^m)/\text{ASE}(\text{HOL}_{rt})$ 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 §5.6.2.1.

Finally, we now compare the performance of NIF to that of other fluid-based predictors. Table 5.2 shows that NIF remains less accurate than QL_{rt} and HOL_{rt}. For example, $\text{ASE}(\text{NIF})/\text{ASE}(\text{QL}_{rt})$ ranges from about 3.1 for $E[S] = 5$ minutes to about 2.1 for $E[S] = 6$ hours. Also, $\text{ASE}(\text{HOL}_{rt})/\text{ASE}(\text{NIF})$ 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 $E[S] \leq 2$ hours, yet it performs better as $E[S]$ increases. Indeed, it is more accurate than QL_a^m and HOL_a^m for large enough $E[S]$. For example, $\text{ASE}(\text{QL}_a^m)/\text{ASE}(\text{NIF})$ ranges from about 0.25 for $E[S] = 5$ minutes to about 1.6 for $E[S] = 6$ hours.

5.6.2.3 Results for Non-Exponential Distributions.

In the appendix, 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 $\bar{s} = 1000$, $\text{ASE}(\text{QL}_a^m/\text{ASE}(\text{QL}_{rt})$ is roughly equal to 18.

We also study the performance of all delay predictors with both non-exponential 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 §A.6 of the appendix. 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 chapter 3, we treated the case of deterministic service times, and found that QL_a is not reliable in the $GI/D/s+GI$ model. 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.

5.7 Concluding Remarks

In this chapter, 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.

Figure 5.1 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 §5.3, we proposed the modified predictors, QL_a^m and HOL_a^m. Then, in §5.5, 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_{rt}, HOL_{rt}, and NIF. All new delay predictors proposed in this chapter reduce to prior ones which 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, $E[S]$.

QL_{rt} is consistently more accurate than both HOL_{rt} and NIF. In terms of efficiency (low ASE), fluid-based predictors are ordered by $\mathrm{QL}_{rt} < \mathrm{HOL}_{rt} <$ NIF. Consistent with prior theoretical results in chapters 2 and 3, simulation showed that $\mathrm{ASE}(\mathrm{HOL}_{rt})/\mathrm{ASE}(\mathrm{QL}_{rt})$ is roughly equal to a constant between 1 and 2; e.g., see Figures 5.2 and 5.3. Although NIF is relatively accurate, particularly in large systems, it performs worse than both QL_{rt} 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 5.3 showed that QL_{rt}, HOL_{rt}, 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 5.2 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 5.2 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 5.2 also showed that the accuracy of QL_a^m and HOL_a^m decreases steadily as $E[S]$ increases. Indeed, both $\text{RRASE}(\text{QL}_a^m)$ and $\text{RRASE}(\text{HOL}_a^m)$ increase with increasing $E[S]$. Although fluid-based predictors perform worse for large $E[S]$ as well, their RRASE's increase much slower than $\text{RRASE}(\text{QL}_a^m)$ and $\text{RRASE}(\text{HOL}_a^m)$.

In some cases, there is not too much difference in performance between the delay predictors. Figure 5.2 showed that QL_a^m is only slightly more accurate than QL_{rt} 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_{rt} for $\bar{s}=1000$ and $E[S]=5$ minutes. In those cases, all delay predictors proposed are relatively accurate.

6

Conclusions

Motivated by interest in making delay announcements in service systems, we studied alternative ways of predicting customer delay, in real time, in queueing models with several realistic features. We started with the $GI/M/s$ model (exponential service times), and extended to $GI/GI/s+GI$ (non-exponential service and abandonment times), $M(t)/GI/s+GI$ (time-varying demand), and $M(t)/GI/s(t)+GI$ (time-varying demand and capacity). We proposed several real-time delay predictors and used mathematical analysis and computer simulation to evaluate their accuracy in each model. We measured accuracy by the *mean-squared error* (MSE) which we estimated via simulation by the *average-squared error* (ASE). Service systems are typically much more complex than the stylized queueing models considered. Nevertheless, the main results of this thesis indicate what to expect more generally.

We considered both delay-history-based and queue-length-based predictors. In Table 6.1, we summarize the information needed for the implementation of each predictor. An important insight, which applies broadly, is that simplicity and ease of implementation are often obtained at the expense of statistical accuracy. In Table 6.2, we

show how the different predictors are related. We have "version consistency" when a given predictor reduces to a simpler one in a more elementary model. Version consistency is important because it shows that the same delay predictor can apply to a wide variety of models. We now summarize how different model characteristics impact the performance of the predictors. The following general conclusions supplement the specific remarks concluding each of the previous chapters. We conclude this thesis by describing future research directions that remain to be investigated.

6.1 Arrival Process

Queue-length-based predictors do not depend on the arrival process. We showed that, conditional on the queue length (number of waiting customers) seen upon arrival, the potential waiting time of a new arrival depends solely on the times between future departures (either service completions or abandonments from the queue); e.g., see (2.5) and (3.9) which give expressions for $W_Q(n)$ in the $GI/M/s$ and $GI/M/s+M$ models, respectively. Consequently, queue-length-based predictors do not depend on the arrival process. In particular, their accuracy does not degrade in face of variability in the arrival process. (Queue-length-based predictors which rely on approximations to the system, such as QL_a, QL_r, or QL_{rt}, exploit knowledge of the arrival-rate intensity function, $\lambda(t)$, which we assume to be a deterministic function of time.) Moreover, we showed that $W_Q(n)$ has the desirable property that the prediction gets relatively more accurate as the observed queue length increases; e.g., see (2.6) for $W_Q(n)$ in the $GI/M/s$ model. Therefore, when the queue length is large, as occurs in heavily-loaded systems, queue-length-based predictions are relatively accurate.

Comparison of delay-history-based and queue-length-based predictors. To a

Predictor	Information About the Model	Defined in
QL	$Q(t), s(t), \mu$	§2.2.2
QL_r^m	$Q(t), s(t), \mu, \nu$	§3.3.4
QL_r, QL_{rt}	$Q(t), s(t), \mu, F(x), \lambda(t)$	§3.3.3,§5.5.2
QL_m	$Q(t), s(t), \mu, \nu$	§3.3.2
QL_a, QL_a^m	$Q(t), s(t), \mu, F(x), \lambda(t)$	§3.3.5,§5.3
HOL_m	$w, s(t), \mu, \lambda(t)$	§4.4
HOL_a, HOL_a^m	$w, s(t), \mu, F(x), \lambda(t)$	§4.8, §5.3
HOL_{rt}	$w, s(t), \mu, F(x), \lambda(t)$	§5.5.3
NIF	$s(t), \mu, F(x), \lambda(t)$	§5.5.1

Table 6.1: Summary of the information required for the implementation of each delay predictor. Also included is the number for the section where the predictor was defined.

large extent, delay-history-based predictors can be regarded as queue-length-based predictors modified by replacing the known queue length by an estimate of that queue length. The conditional mean delay given the observed delay information (e.g., LES delay) minimizes the MSE. Therefore, the conditional mean, or an approximation of it, serves as a refined delay-history-based predictor. (For example, $\theta^d_{LES}(w_L) \equiv w_L$ is the direct predictor and $\theta^r_{LES}(w) \equiv E[W_{LES}(w_L)]$ is a refined predictor, both based on the LES delay, w_L.) Refined predictors can remove all or nearly all of the bias, but non-negligible variance remains. For a refined predictor, the queue length is estimated by the expected number of arrivals who do not abandon during the observed waiting time. For one example, compare (2.5) and (2.9) which give expressions for $W_Q(n)$ and $W_{LES}(w_L)$ in the $GI/M/s$ model. For another example, compare (3.9) and (3.30) for similar expressions in the $GI/M/s + M$ model. Thus, the increase in MSE in going from queue-length-based to refined delay-history-based predictors is primarily because of variability in the arrival process. With a stationary arrival process, we found that refined and direct delay-history-based predictors perform nearly the same, particularly when the number of servers is large; e.g., see Table 2.5 for a comparison of the performance of alternative HOL-based predictors in the

$M(t)/GI/s(t)+GI$	$M(t)/GI/s+GI$	$GI/GI/s+GI$	$GI/M/s+M$	$GI/M/s$
QL_a^m	QL_a	QL_a	QL_m	QL
QL_{rt}	QL_{rt}	QL_r	QL_r	$--$
HOL_a^m	HOL_a	HOL_a	$--$	$--$
HOL_{rt}	HOL_{rt}	HOL	HOL	HOL
NIF	NIF	NI	NI	NI

Table 6.2: Version consistency of the predictors in the alternative queueing models. In a given row, the predictor on the left, in the more complex model, reduces to the predictor on the right in the simpler model. An empty entry indicates that we did not consider an equivalent predictor in that particular model.

$H_2/M/s$ model. In the $GI/M/s$ model, we established the asymptotic equivalence of direct and refined delay-history-based predictors in the classical and many-server heavy-traffic limiting regimes; see Theorems 2.6.1 and 2.6.2.

Recap of main theoretical results. We derived several theoretical results quantifying the difference in performance between queue-length-based and delay-history-based predictors. We found that predictors that do not exploit knowledge of the queue length fare worse than queue-length-based predictors, largely according to c_a^2; c_a^2 is the SCV of an interarrival time, a common measure of variability for a renewal arrival process. In the $GI/M/s$ model, we showed that the MSE tends to be larger for LES and HOL than QL by the constant factor (c_a^2+1) in both the classical and many-server (quality-and-efficiency-driven, QED) heavy-traffic limiting regimes (see §2.4). In §3.5, we established similar results for the $GI/M/s+M$ model. For example, we showed that MSE(LES) increases with c_a^2 in the efficiency-driven (ED) many-server limiting regime; see (3.37). As a result, we proved in Corollary 3.5.3 that the difference in performance between QL_m and LES, in the ED limiting regime, depends strongly on the variability of the arrival process: The two predictors perform nearly the same for low c_a^2, but not otherwise. Additionally, we described results of simulation experiments substantiating our analysis; e.g., see

§A.2 for performance in the $GI/M/s+M$ model. Consistent with results for the $M/M/s$ model, we showed in Proposition 4 that the ratio MSE(HOL$_m$)/MSE(QL) is asymptotically (in heavy load) equal to $2/\rho$ in the $M(t)/M/s$ model (ρ denotes the traffic intensity). Additional simulation results showed that similar conclusions should hold more generally as well. For one example, Table 4.2 showed that ASE(HOL$_m$)/ASE(QL) is roughly equal to a constant in the $M(t)/GI/s$ model (depending on the service-time distribution). For another example, Figures 5.2 and 5.3 showed that ASE(HOL$_{rt}$)/ASE(QL$_{rt}$) is roughly equal to a constant between 1 and 2 in the $M(t)/M/s(t)+GI$ model.

Time-varying arrival rates. We showed in chapter 4 that the performance of delay-history-based predictors also degrades in face of time-varying arrivals; that was dramatically shown in Figure 4.2. Intuitively, when the delays vary systematically over time, as can occur with alternating periods of significant underload and overload, the delay of a new arrival may not be like a recently observed delay. In chapter 4, we considered the HOL predictor, but did so with the understanding that results for HOL should apply equally well to LES and other delay-history-based predictors (e.g., see Theorem 2.4.4). With time-varying arrivals, we showed that refined predictors, such as HOL$_m$, can perform significantly better than direct predictors, such as HOL (that is different than with a stationary arrival process); e.g., see Table 4.2 for a comparison of the ASE's of HOL and HOL$_m$ in the $M(t)/GI/s$ model. However, the new refined predictors lose some of their appeal compared to simple delay-history-based predictors, because they require additional information about system parameters. We proposed estimation procedures for alternative system parameters and quantified the estimation error resulting from those procedures in §4.7.

6.2 Customer Abandonment

Abandonment rate. As indicated by formulas (3.3) and (3.7) for the $M/M/s+M$ model in the ED regime, the steady-state queue length and delay in the system generally tend to be inversely proportional to the abandonment rate, ν. (Similar results hold more generally for the $GI/GI/s+GI$ model; see (3.6) and (3.7) of Whitt (2006).) We usually let $\nu = 1.0$. But, we also considered other values of ν, such as $\nu = 0.2$ and $\nu = 5.0$. Changing ν from 1.0 to 5.0 or 0.2 tends to change congestion by a factor of 5. That is, the system is very heavily loaded when $\nu = 0.2$, but relatively lightly loaded when $\nu = 5.0$. As a result, we found that larger (smaller) ν leads to smaller (larger) ASE's for all delay predictors considered; e.g., compare Tables A.8 and 3.2 for performance in the $M/M/s+M$ model with $\nu = 5.0$ and $\nu = 1.0$, respectively.

Ignoring customer abandonment. Queue-length-based predictors which fail to take abandonment into account can perform very poorly in face of significant customer abandonment. For example, we showed in chapter 3 that the QL predictor (which has superior performance in the $GI/M/s$ model) consistently overestimates delays in the $GI/GI/s+GI$ model; e.g., see Figures 3.1-3.4 for performance in the $M/M/s+GI$ model. We used steady-state fluid approximations for the $M/M/s+M$ model in the ED regime to quantify the resulting prediction error in that model; see (3.8). We found that this error increases with ρ.

Approximation by an exponential. Prior theoretical results in Whitt (2005b, 2006) suggest that performance measures in the system typically depend on the abandonment-time distribution beyond its mean. Consistent with those results, we found that queue-length-based predictors which approximate the abandonment-time distribution by an exponential (with the same mean) have low predictive power

when the abandonment-time distribution is not nearly exponential. To consider both higher and lower variability relative to the exponential distribution, we considered H_2 (hyperexponential with balanced means and SCV equal to 4), and E_{10} (Erlang, sum of 10 exponentials) abandonment-time distributions.

Paralleling QL, we proposed in §3.3.2 a new predictor, QL_m, which approximates the $GI/GI/s+GI$ model by a $GI/M/s+M$ model with the same service-time and abandon-time means. Under the MSE criterion, QL_m is the most accurate predictor in the $GI/M/s+M$ model since it coincides, in that model, with the conditional mean delay given the queue length. But, with non-exponential abandonment times, QL_m can perform very poorly; e.g., see Figures 3.5 and 3.6 for performance in a heavily-loaded $M/M/s+E_{10}$ model with large values of s.

Coping with non-exponential abandonment times. To effectively cope with non-exponential abandonment-time distributions, often observed in practice (Brown et al. (2005)), we developed new queue-length-based predictors exploiting established approximations for the system. We proposed in §3.3.3 a refined predictor, QL_r, which draws on a deterministic fluid approximation for the $GI/GI/s+GI$ model, developed in Whitt (2006). Extensive simulation results showed that QL_r performs remarkably better than QL_m with non-exponential abandonment times. For example, see Figures 3.1-3.6 for performance in the $M/M/s+GI$ model. With time-varying demand and capacity, we drew on a deterministic fluid approximation for a many-server queueing model with those features, developed in Liu and Whitt (2010), to obtain a similar predictor, QL_{rt}, which reduces to QL_r in the $GI/GI/s+GI$ model; see Table 6.2.

We proposed in §3.3.5 an approximation-based predictor, QL_a, which approximates the abandonment-time distribution $(+GI)$ by state-dependent Markovian abandonment $(+M(n))$, drawing on results in Whitt (2005b). Extensive simulation results

showed that QL_a has superior performance in the $GI/M/s+GI$ model. For example, Figures 3.3-3.6 showed that QL_a performs remarkably well in the $M/M/s+GI$ model with H_2 and E_{10} abandonment times. Since QL_a assumes a constant arrival rate and a constant number of servers, we proposed in §5.3 a modified version of QL_a, QL_a^m, which we showed effectively copes with non-exponential abandonment times, time-varying arrivals, and a time-varying number of servers; e.g., see Figures A.5-A.8 for performance in the $M(t)/M/s(t)+M$ model. The QL_a^m predictor reduces to QL_a in the $GI/GI/s+GI$ model; ; see Table 6.2.

Robustness of delay-history-based predictors. Delay-history-based predictors are appealing because they are robust: They do not require knowledge of the model or its parameters, and they rely solely on the history of recent delays in the system. As a result, they are accurate in models with customer abandonment, irrespective of the abandonment-time distribution (they make no assumptions about that distribution). For example, Figures 3.3-3.6 showed that LES and HOL can significantly outperform QL_m in heavily-loaded $M/M/s+H_2$ and $M/M/s+E_{10}$ models with a large number of servers.

6.3 Service Times

Deterministic service times. In this thesis, we considered H_2, LN (lognormal), M, E_{10}, and D (deterministic) service times. In all models, we implemented the candidate predictors by approximating the service-time distribution by an exponential distribution with the same mean service time. Simulation results with exponential and non-exponential service times are generally consistent, with one notable exception. There is a significant increase in ASE for all predictors with deterministic (constant) service times, with performance tending to be independent of the number

of servers. That indicates a need for new methods for this special case; e.g., see Table 3.7 for performance in the $M/D/s+M$ model. (In that case, we found that NI is the most accurate predictor.) With stationary arrivals, even very low variability in the service times, e.g., the E_{10} distribution with SCV equal to 0.1, is enough for our delay predictors to be relatively accurate; e.g., see Table 3.8 for results in the $M/E_{10}/s+M$ model.

With time-varying arrivals and a time-varying number of servers, the situation is more complicated. In that case, we found that low-variability service-time distributions can be problematic for some predictors, but not for others. For example, Tables A.36 and A.38 showed that fluid-based predictors, such as QL_{rt} and HOL_{rt}, are ineffective with E_{10} service times, whereas QL_a^m and HOL_a^m remained relatively effective in that case. The fluid model proposed in Lui and Whitt (2010) extends to non-exponential service times. Therefore, there remains the possibility to develop new fluid-based predictors based on the more general, and significantly more complicated, fluid model. We leave such extensions to future research.

Approximation by an exponential. Delay predictors which are equal to the conditional mean delay (given some state information) with M service times tend to overestimate or underestimate delays with GI service times. That is primarily because many remaining service times at the new arrival epoch, t, are residual service times for service times begun prior to time t. With a new-better-than-used (NBU) distribution, such as E_{10}, approximation by an exponential leads to overestimating the residual service times and thus the overall delay. In contrast, with a new-worse-than-used (NWU) distribution, such as H_2, approximation by an exponential leads to underestimating the residual service times and thus the overall delay. We proved those results for HOL_m and QL in the $M(t)/GI/s$ model; see Propositions 1 and 2. Simulation showed that similar conclusions hold in more general models as well.

6.4 Number of Servers

Throughout this thesis, we primarily focused on a large number of servers (hundreds) because we are interested in large service systems. Nevertheless, we also studied systems with a medium (tens) or small (less than ten) number of servers. We established several results quantifying the asymptotic (i.e., in large systems) accuracy of the alternative predictors. Those results confirm that our proposed predictors are remarkably accurate in large systems.

Asymptotic results for large systems. A predictor is asymptotically relatively consistent if the ratio of the predictor to the quantity being predicted (here, the potential delay) converges to 1; a predictor is asymptotically relatively efficient if the ratio of the MSE to the square of the mean converges to 0. In chapter 2, we established asymptotic results (for the performance of several predictors) in the $GI/M/s$ model. In Theorem 2.6.2, we showed that QL and HOL are both relatively efficient and consistent in the (many-server) QED limiting regime. We also showed that QL is asymptotically more efficient than HOL, in that regime, by the constant factor (c_a^2+1). In §2.6.3, we showed that RCS and RCS$-\sqrt{s}$ are relatively efficient in the QED regime, whereas LCS is not. We also showed that all delay-history-based predictors are asymptotically equivalent in the classical heavy-traffic regime.

A predictor is asymptotically correct if it is effective in large systems. In particular, the corresponding MSE (ASE) converges to 0 as the number of servers increases. In simulation experiments, we computed $s \times$ ASE (the number of servers times the ASE) for the alternative predictors. A predictor is asymptotically correct if $s \times$ ASE is roughly equal to a constant for large values of s. In §3.5, we established asymptotic results for several predictors in the $GI/M/s+M$ model, in the ED regime. In particular, we showed in Corollary 3.5.1 that the NI predictor is asymptotically

correct in that limiting regime. That indicates that all predictors need be asymptotically correct as well to be worth serious consideration. In Theorems 3.5.1 and 3.5.4, we showed that QL_m and LES are asymptotically correct in that setting as well. Simulation results showed that similar properties should hold for QL_a and QL_r, in addition to LES, in the $M/GI/s + GI$ model (with the exception of D service times). For example, Figures 3.3-3.6 showed that QL_a and QL_r (as well as NI and LES) are asymptotically correct in the $M/M/s + GI$ model with H_2 and E_{10} abandonment-time distributions.

In Proposition 3, we showed that HOL_m is asymptotically correct in the $M(t)/M/s$ model (the variance of the actual delay, $Var[W_{HOL}(t, w)]$, is equal to the MSE of HOL_m). Also, simulation results suggested that HOL_a is asymptotically correct in the $M(t)/M/s + GI$ model; e.g., see Figures 4.5-4.10 for performance with M, H_2, and E_{10} abandonment.

In the $M(t)/M/s(t) + GI$ model, we used simulation to show that several candidate predictors are asymptotically correct. For example, Figure 5.3 showed that fluid-based predictors (i.e., NIF, QL_{rt} and HOL_{rt}) are asymptotically correct in the $M(t)/M/s(t) + M$ model. However, the same does not hold for the modified predictors (i.e., QL_a^m and HOL_a^m). Additional simulation results (in the appendix) showed that the same conclusions hold with general abandonment-time distributions as well; e.g., see Figures A.2 and A.4 for performance in the $M(t)/M/s(t) + H_2$ and $M(t)/M/s(t) + E_{10}$ models, respectively.

Small number of servers. Asymptotic results which hold, in the limit, as the number of servers increases do not adequately describe performance in small systems. For example, Figures 5.2 ($M(t)/M/s(t) + M$), A.1 ($M(t)/M/s(t) + H_2$), and A.3 ($M(t)/M/s(t) + E_{10}$) showed that all predictors are not very accurate when the number of servers is relatively small. With the same traffic intensity, congestion in

the system tends to be higher with a small number of servers. (To illustrate, consider the $M/M/s$ model special case: The steady-state delay given that the delay is positive has an exponential distribution with mean $1/s(1-\rho)$; clearly the magnitude of the delays increases as s decreases, for any fixed ρ.) Consequently, we found that the ASE's of all predictors tend to be higher in small systems; e.g., compare Tables 2.3, 2.2, and 2.1 for performance in the $GI/M/s$ model with $s = 1, 10$, and 100, respectively.

Time-varying number of servers. Figure 5.1 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 §5.3, we proposed the modified predictors, QL_a^m and HOL_a^m. Then, in §5.5, we proposed the new fluid-based delay predictors, QL_{rt}, HOL_{rt}, and NIF. Since direct analysis is difficult in that setting, we relied on computer simulation to study the performance of the alternative predictors proposed. We showed that the new predictors are effective in the $M(t)/M/s(t) + GI$ queueing model with time-varying arrival rates and a time-varying number of servers. As explained in chapter 4, delay-history-based predictors perform particularly poorly in face of such time variation.

6.5 Future Research Directions

There are many interesting research problems, closely related to this thesis, that remain to be investigated. First, we mostly focused in our work on systems with a relatively large number of servers. As suggested above, results with a small number of servers could be dramatically different. Therefore, one future research direction is to study delay predictions in smaller service systems. Second, a natural extension to our work would explicitly model customer reactions to delay announcements,

which is a phenomenon often observed in practice. Incorporating insights from behavioral studies in waiting situations into modeling those reactions is an especially promising direction. Third, it would be interesting to study the performance of the alternative predictors proposed in an actual service system. That could be done through experiments with real-life system data. With real-life data, it is important to obtain all information needed for the implementation of the candidate predictors. Fourth, as mentioned above, there remains to study ways of making effective delay predictions when service times are deterministic or nearly so. From a practical perspective, nearly deterministic service times are observed in amusement parks or in subway systems. Finally, an important future research direction is to study delay predictions in more complicated models with additional realistic features. For example, one could consider models incorporating multiple customer classes and some form of skill-based routing (such as in call centers), or networks of queues where customers go through several service stations (such as in hospitals).

A

Additional Simulation Results

A.1 Additional Simulation Experiments for the $GI/M/s$ Model

In this section, we present additional simulation results quantifying the performance of the alternative predictors in the $GI/M/s$ model.

In Tables A.1-A.4, we present estimates of the ASE of several predictors conditional on the level of actual delay in the system in the $M/M/100$ model. In particular, we consider actual delays that fall in one of the following intervals: $(E[W|W > 0], 2E[W|W > 0])$, $(2E[W|W > 0], 4E[W|W > 0])$, $(4E[W|W > 0], 6E[W|W > 0])$ and $(6E[W|W > 0], \infty)$, where $E[W|W > 0]$ denotes the expected waiting time (in steady state) given that the wait is positive. (We often use a simulation point estimate of $E[W|W > 0]$.) In Tables A.5-A.7, we consider the RCS-based predictor, RCS$-f(s)$, in the $M/M/100$ model, $D/M/100$, and $H_2/M/100$ models, respectively. For RCS$-f(s)$, the delay prediction is equal to the RCS delay among

the last $f(s)$ customers that have completed service. We consider different values of $f(s)$ to deduce the amount of data that need be examined to determine the overall RCS customer. More simulation results appear in an online supplement to Ibrahim and Whitt (2009a), available on the authors' webpages.

A.1.1 Conditional Performance of the Predictors

Evaluating the conditional performance of the predictors is interesting. The relative performance of the alternative predictors is roughly the same as before, but there are some differences. For large delays, relatively accurate predictors become increasingly accurate, and relatively inaccurate predictors become increasingly inaccurate. For example, Table A.1 shows that, conditional on the level of delay falling in $(E[W|W > 0], 2E[W|W > 0])$, the relative ASE (RASE, equal to the square root of the ASE divided by the mean waiting time) for QL is roughly equal to 11% and RASE(NI) is roughly equal to 16%, for $\rho = 0.99$. That is, the difference in performance between QL and NI is not too great in that case. On the other hand, Table A.3 shows that, conditional on the level of delay falling in $(4E[W|W > 0], 6E[W|W > 0])$, RASE(QL) is roughly equal to 6% whereas RASE(NI) is roughly equal to 600%, for $\rho = 0.99$. The difference in performance between the two predictors is now remarkable. In particular, Table A.3 shows that ASE(NI)/ASE(QL) is roughly equal to 275 for $\rho = 0.99$ which considerably exceeds the theoretical value of $1/(1-\rho) = 100$ (ratio of the MSE's in the $M/M/s$ model).

When restricting attention to relatively small actual delays, the NI predictor is more accurate than LCS. For example, Table A.1 shows that ASE(LCS)/ASE(NI) ranges from about 5 for $\rho = 0.9$ to about 2 for $\rho = 0.99$. That is, the NI predictor is significantly more accurate than LCS when the system is lightly loaded. That is

significant because NI does not exploit any information about the system beyond the model.

Finally, the RCS, LES, and HOL predictors are more accurate when conditioning on large delays. For example, Table A.1 shows that all three predictors have an RASE which is roughly equal to 17% for $\rho = 0.99$. In contrast, Table A.2 shows that the RASE reduces to about 7% for the same value of ρ. We get even better results when restricting attention to higher delays. In general, the performance of these three predictors also improves when considering higher ρ. That substantiates theoretical results which show that all three predictors are asymptotically correct.

A.1.2 The Effect of Delay Information: RCS$-f(s)$

We now describe simulation results quantifying the performance of RCS$-f(s)$ in the $M/M/100$ model. (In Table A.6 and Table A.7, we present corresponding results for the $D/M/100$ and $H_2/M/100$ models, respectively. We do not describe those results separately here because they are largely consistent with the M case.) The objective is to determine how much data need be examined to determine the identity of the RCS customer. Theoretical results suggest that the RCS customer is very likely to be among the last $c\sqrt{s}$ customers to have completed service. Extensive simulation results, of which we show a sample here, show that this is indeed the case.

The RCS predictor, exploiting all past delay information is, naturally, the most accurate predictor among all RCS-$-f(s)$ predictors considered. However, Table A.5 shows that RCS-$4\sqrt{s}$, which looks for the RCS customer among the last $4\sqrt{s}$ customers to have completed service, performs exactly the same as RCS. Therefore, it suffices to look back among the last $4\sqrt{s}$ customers to have completed service

in order to find the RCS customer. That can be significant. For example, with $s = 100$, the service system manager need only examine the waiting times of the last $4\times\sqrt{1}00 = 40$ customers to have completed service, instead of the entire data set. Table A.5 also shows that RCS-$c\sqrt{s}$ differs by at most 1% (from RCS) when $c = 2$ and differs by at most 6% when $c = 1$.

A.2 Additional Simulation Results for the $GI/M/s + M$ Model

In this section, we present simulation results for the $GI/M/s + M$ model which substantiate the heavy traffic limits of §3.5. For the interarrival-time distribution, we consider M, D, and H_2. We consider the same values of s as before: $s = 100, 300, 500, 700,$ and 1000. We let the service rate be $\mu = 1.0$, and consider three different values of the abandonment rate, $\nu = 1.0, 5.0,$ and 0.2. We vary the arrival rate λ to get a fixed value of ρ for alternative values of s, $\rho = 1.4$. In this model, QL_a coincides with QL_m so we do not include separate results for it. For more simulation results of the $GI/M/s + M$ model, see §3.8.

A.2.1 The $M/M/s + M$ Model with $\nu = 5.0$ and $\nu = 0.2$

In §3.6, we presented simulation results for the $M/M/s+M$ model with $\nu = 1.0$. We now consider different abandonment rates; specifically we let $\nu = 5.0$ and $\nu = 0.2$. As indicated by formulas (3.3) and (3.7), the queue length and delay tend to be inversely proportional to ν. Thus, changing ν from 1.0 to 5.0 or 0.2 tends to change congestion by a factor of 5. The system is very heavily loaded when $\nu = 0.2$, but relatively lightly loaded when $\nu = 5.0$.

Table A.8 compares the efficiencies of the alternative predictors with $\nu = 5.0$, which makes the model more lightly loaded. In this more lightly loaded setting, the ASE's of all the predictors are relatively low, being smaller than for the $M/M/s+M$ model with $\nu = 1.0$, in Table 3.2, by a factor of about 5.

The lighter loading makes the ED approximations less appropriate. Simulation estimates in Table A.8 allows us to compare the ASE's of QL_m, LES, and NI to the expected MSE's in formulas (3.25), (3.36), and (3.29), respectively; the relative errors (RE) reported are higher than with $\nu = 1.0$, especially when the number of servers is small (e.g., in Table A.8, with $s = 100$, the RE reported exceeds 20%). But, the formulas are much more accurate with a larger number of servers (e.g., the RE's are close to 1 or 2%, with $s = 1000$).

Table A.8 shows that the ratio ASE(LES)/ASE(QL_m) is well approximated by the numerical value, 2.0, predicted by equation (3.54), except when the number of servers is small (e.g., with $s = 100$, RE $\approx 10\%$). Similarly, Table A.8 shows that the ratio ASE(NI)/ASE(QL_m) is well approximated by the numerical value, 3.5, given by (3.29), except when s is small: The RE reported when $s = 100$ is close to 20%.

Table A.9 compares the efficiencies of the alternative predictors with $\nu = 0.2$. In this more heavily loaded setting, the ASE's of the alternative predictors are higher than with $\nu = 1.0$, by a factor of about 5, especially when the number of servers is large.

Simulation estimates in Table A.9 allows us to compare the ASE's of QL_m, LES, and NI to the expected MSE's in formulas (3.25), (3.36), and (3.29), respectively. These formulas are remarkably accurate: The RE's reported are less than 2% throughout. The ratios ASE(LES)/ASE(QL_m) and ASE(NI)/ASE(QL_m) agree closely with the

values predicted by formulas (3.54) and (3.29): The RE's reported are less than 5% throughout.

A.2.2 The $D/M/s+M$ Model with $\nu = 1.0$ and $\nu = 5.0$

Simulation results for the $D/M/s+M$ model with $\nu = 0.2$ were described in §3.8.1. The observations made above for the $M/M/s+M$ model with $\nu = 5.0$ still apply, essentially, to the $D/M/s+M$ model with the same value of ν (see Table A.11), so we will not treat this case separately here. In the following, we describe simulation results for the $D/M/s+M$ model with $\nu = 1.0$; see Table A.10.

Table A.10 shows that, consistent with theory, QL_m is the best possible delay predictor, under the MSE criterion. The RRASE of QL_m ranges from about 16% when $s = 100$ to about 5% when $s = 1000$. All predictors are relatively accurate as well; e.g., the RRASE of LES ranges from about 24% when $s = 100$ to about 6% when $s = 1000$. The QL_r predictor is nearly as efficient as QL_m.

Table A.10 also shows that, consistent with equation (3.54), the LES predictor performs slightly worse than QL_m: The RE between the simulation estimates for ASE(LES)/ASE(QL_m) and the numerical value, 1.286, given by (3.55) is less than 3% throughout. With a deterministic arrival process, the LES predictor performs better, compared to QL_m, than with a Poisson arrival process. Similarly, Table A.10 shows that, consistent with equation (3.29), the NI predictor is less efficient than QL_m: The RE between the simulation estimates for ASE(NI)/ASE(QL_m) and the numerical value, 2.25, given by (3.29) is less than 4% throughout. The QL predictor is, once more, the least efficient predictor: The ratio ASE(QL)/ASE(QL_m) ranges from about 3 when $s = 100$ to about 15 when $s = 1000$.

Finally, Table A.10 shows that the ASE's of QL_m, LES, and NI are consistent with

the analytical formulas for the expected MSE's given in (3.25), (3.36), and (3.29), respectively. These formulas are quite accurate: The RE's reported are less than 3% throughout, except when the number of servers is large (e.g., with $s = 1000$ in Table 21, RE $\approx 7\%$).

A.2.3 The $H_2/M/s + M$ Model with $\nu = 0.2$ and $\nu = 1.0$

Simulation results for the $H_2/M/s+M$ model with $\nu = 5.0$ were described in §3.8.2. The observations made above for the $M/M/s + M$ model with $\nu = 0.2$ still apply, essentially, to the $H_2/M/s+M$ model with the same value of ν (see Table A.12), so we will not treat this case separately here. In the following, we describe simulation results for the $H_2/M/s + M$ model with $\nu = 1.0$; see Table A.13.

With hyperexponential interarrival times, Table A.13 shows that, consistent with theory, QL_m is the best possible delay predictor, under the MSE criterion. The RRASE for QL_m ranges from about 16% for $s = 100$ to about 5% when $s = 1000$. The QL_r predictor is only slightly outperformed by QL_m.

The ED approximations are less accurate with highly variable interarrival times than with exponential interarrival times. Table A.13 shows that, consistent with equation (3.56), the LES predictor performs worse than QL_m: The RE between the simulation estimates for ASE(LES)/ASE(QL_m) and the numerical value, 4.143, given by (3.56) ranges from about 6% when $s = 100$ to about 2% when $s = 1000$. The LES predictor performs worse, compared to QL_m, with hyperexponential interarrival times, than with exponential interarrival times. Table A.13 shows that, consistent with equation (3.29), the NI predictor is significantly less efficient than QL_m (and LES): The RE between the simulation estimates for ASE(NI)/ASE(QL_m) and the numerical value, 7.25, given by (3.29) ranges from about -9.0% when $s = 100$ to about

-0.5% when $s = 1000$. The QL predictor performs significantly worse than all the other predictors, particularly for large values of s. The ratio ASE(QL)/ASE(QL_m) ranges from about 4 when $s = 100$ to nearly 16 when $s = 1000$.

Finally, Table A.13 shows that the ASE's of QL_m, LES, and NI are consistent with the analytical formulas for the expected MSE's given in (3.25), (3.36), and (3.29), respectively. These formulas are accurate for large values of s, but less so for smaller values of s: e.g., RE $\approx -8\%$ when $s = 100$ and RE $\approx -0.033\%$ when $s = 1000$.

A.3 Additional Simulation Results for the $M(t)/GI/s+GI$ Model

In this section, we present simulation results for the $M(t)/GI/s+GI$ model. For the service-time distribution, we consider D and H_2 distributions. For the abandonment-time distribution, we consider M, H_2, and E_{10}. With sinusoidal arrival rates, we consider a relative frequency $\gamma = 1.571$ which corresponds to a mean service time $E[S] = 6$ hours with daily cycles. The average traffic intensity ρ is fixed at $\rho = \bar{\lambda}/s\mu = 1.4$. For the relative amplitude, we let $\alpha = 0.5$. Our simulation results are based on 10 independent replications of length 1 month each.

A.3.1 D service times

In Tables A.14, A.16, and A.18, we present simulation results for the $M(t)/D/s+M$, $M(t)/D/s + H_2$, and $M(t)/D/s + E_{10}$ models, respectively. Tables A.14 and A.16 show that, with both M and H_2 abandonment, we get simulation results consistent with those reported earlier.

Table A.14 shows that, with M abandonment, QL_m remains the most effective predictor under the MSE criterion. The RRASE of QL_m ranges from about 17% for $s = 100$ to about 14% for $s = 1000$. The second best predictor is the HOL_a predictor. The RRASE of HOL_a ranges from about 20% for $s = 100$ to about 15% for $s = 1000$. The difference in performance between QL_m and HOL_a is not too great: ASE(HOL_a)/ASE(QLm) ranges from about 1.5 for $s = 100$ to 1.05 for $s = 1000$. That is, the performance of QL_m and HOL_a is roughly the same for large values of s. That is to be expected, since both predictors are asymptotically correct. The least effective predictor, among those considered, is the HOL predictor. The RRASE of HOL is close to 30% for all values of s considered. The HOL predictor is not asymptotically correct for this model, as expected.

Table A.16 shows that we get similar results with H_2 abandonment. In this case, HOL_a is the most effective predictor for large s. The RRASE of HOL_a ranges from about 25% for $s = 100$ to about 15% for $s = 1000$. The difference in performance between QL_m and HOL_a is not too great: ASE(QL_m)/ASE(HOL_a) ranges from about 0.95 for $s = 100$ to about 1.34 for $s = 1000$. The HOL predictor is, once more, the least effective predictor: ASE(HOL)/ASE(HOL_a) ranges from about 2 for $s = 100$ to about 3 for $s = 1000$.

With E_{10} abandonment, Table A.18 shows that we get results different from those described above. Indeed, the performance of all predictors is bad, with performance tending to be independent of the number of servers in the system. The performance of QL_m and HOL_a is nearly the same, and they are both somewhat ineffective: RRASE(HOL_a) is close to 20% for all values of s considered. Moreover, plots of $s\times$ ASE for all predictors show that all predictors are not asymptotically correct in this model. The HOL predictor is, once more, the least effective predictor. The ratio ASE(HOL)/ASE(HOL_a) is close to 3 for all values of s considered. There is a

need to consider other predictors in this model. We leave this interesting research direction to future work.

A.3.2 H_2 service times

In Tables A.15, A.17, and A.19, we present simulation results for the $M(t)/H_2/s+M$, $M(t)/H_2/s+H_2$, and $M(t)/H_2/s+E_{10}$ models, respectively.

Table A.15 shows that QL_m remains the most effective predictor in this model. The RRASE of QL_m ranges from about 16% for $s = 100$ to less than 5% for $s = 1000$. The HOL_a predictor is the second best predictor in this model. The ratio $\text{ASE}(\text{HOL}_a)/\text{ASE}(\text{QL}_m)$ is close to 1.5 for all values of s considered. The HOL predictor is, once more, the least effective predictor among those considered. The RRASE of HOL ranges from about 30% for $s = 100$ to about 24% for $s = 1000$. The ratio $\text{ASE}(\text{HOL})/\text{ASE}(\text{QL}_m)$ ranges from about 4 for $s = 100$ to about 24 for $s = 1000$. Once more, we see a significant degradation in the performance of HOL with time-varying arrivals.

Table A.17 shows that, with H_2 abandonment, QL_m is no longer the most effective predictor particularly for a large number of servers. The ratio $\text{ASE}(\text{QL}_m)/\text{ASE}(\text{HOL}_a)$ ranges from about 0.8 for $s = 100$ to about 2.5 for $s = 1000$. The RRASE of HOL_a (which is the best possible in this model) ranges from about 21% for $s = 100$ to about 6% for $s = 1000$. The RRASE of QL_m ranges from about 20% for $s = 100$ to about 11% for $s = 1000$. The HOL predictor is the least effective predictor: $\text{ASE}(\text{HOL})/\text{ASE}(\text{HOL}_a)$ ranges from about 2 for $s = 100$ to about 8 for $s = 1000$. Plots of $s\times$ASE of the predictors show that HOL_a is asymptotically correct in this model, whereas QL_m and HOL are not.

Table A.19 shows that, with E_{10} abandonment, HOL_a is the most effective predictor

under the MSE criterion. The QL_m predictor is the second best predictor. The ratio $ASE(QL_m)/ASE(HOL_a)$ ranges from about 1.4 for $s = 100$ to about 7 for $s = 1000$. The RRASE of HOL_a (which is the best possible in this model) ranges from about 10% for $s = 100$ to less than 5% for $s = 1000$. The RRASE of QL_m is close to 10% for all s considered. The HOL predictor is the least effective predictor: $ASE(HOL)/ASE(HOL_a)$ ranges from about 6 for $s = 100$ to about 34 for $s = 1000$. Plots of $s \times ASE$ of the predictors show that HOL_a is asymptotically correct, whereas QL_m and HOL are not.

A.4 Estimating the Required Additional Information for HOL_m

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

In this section, we describe additional simulation experiments quantifying the effect of additional information on HOL_m. In particular, we would like to assess the level of error x that is allowed for the performance of HOL_m to remain better than that of HOL. In general, we find that the relative of admissible error x is around 5%; see Tables A.20-A.31. Note that the length of the estimation interval needed for each of the service-time distributions depends on the variability of the distribution itself.

In particular, high variability distributions, such as H_2, require longer intervals.

A.5 Additional Simulation Results for the $M(t)/M/s(t)+GI$ Model

In this section, we study the performance of the alternative delay predictors with a general (non-exponential) abandonment-time distribution and an exponential service-time distribution. In particular, we consider the $M(t)/M/s(t)+GI$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28). We let $\gamma_s = \gamma_a = 1.57$, which corresponds to $E[S] = 6$ hours with a 24 hour cycle. We let $\alpha_a = 0.5$ and $\alpha_s = 0.3$. We vary the average number of servers, $\bar{s}$, from 10 to 1000. To consider both higher and lower variability relative to the exponential distribution considered previously, we consider H_2 (hyper-exponential with balanced means and SCV equal to 4), and E_{10} (Erlang, sum of 10 exponentials) abandonment-time distributions. In Tables A.32 and A.33, we present point estimates of the ASE and half width of the 95% confidence intervals for the $M(t)/M/s(t)+H_2$ and $M(t)/M/s(t)+E_{10}$ models, respectively, as a function of $\bar{s}$. Additionally, in Figures A.1-A.4, we plot $\bar{s}\times$ASE (average number of servers times the ASE) for the alternative delay predictors in those two models.

A.5.1 Results for the $M(t)/M/s(t)+H_2$ Model.

A.5.1.1 Less reliable predictions in small systems.

Simulation results with H_2 abandonment times are generally consistent with those obtained with M abandonment times; see §4.9. However, with H_2 abandonment, all predictors are slightly less accurate when the number of servers is small. For one ex-

ample, in the $M(t)/M/s(t)+H_2$ model, RRASE(QL_a^m) is roughly equal to 72% (63% with M abandonment) for $\bar{s} = 10$. For another example, in the $M(t)/M/s(t)+H_2$ model, RRASE(QL_{rt}) is roughly equal to 74% (67% with M abandonment) for $\bar{s} = 10$; see Tables A.32 and A.34. In large systems, all predictors perform nearly the same in both models.

A.5.1.2 Superiority of fluid-based predictors.

As in Figure 5.2, Figure A.1 shows that fluid-based predictors are competitive with H_2 abandonment, even when the number of servers is not too large. For example, Table A.32 shows that ASE(QL_a^m)/ASE(QL_{rt}) is roughly equal to 1.2 for $\bar{s} = 20$. (That is consistent with M abandonment; see Table A.34.) Consistent with Figure 5.3, Figure A.1 shows that $\bar{s}\times$ASE for fluid-based predictors is roughly equal to a constant for $\bar{s} \geq 50$. In contrast, $\bar{s}\times$ASE for QL_a^m and HOL_a^m increases roughly linearly with $\bar{s}$.

As with M abandonment, the accuracy of fluid-based predictors greatly improves as the number of servers increases. The QL_{rt} predictor is the most accurate predictor for $\bar{s} \geq 20$, and RRASE(QL_{rt}) ranges from about 74% (67% with M abandonment) for $\bar{s} = 10$ to less than 9% (8% with M abandonment) for $\bar{s} = 1000$. The difference in performance between QL_{rt} and QL_a^m can be, as with M abandonment, remarkable; e.g., Table A.32 shows that ASE(QL_a^m)/ASE(QL_{rt}) ranges from about 0.9 for $\bar{s} = 20$ (same as with M abandonment) to about 22 for $\bar{s} = 1000$ (26 with M abandonment). The HOL_{rt} predictor is relatively accurate as well: RRASE(HOL_{rt}) ranges from about 83% for $\bar{s} = 10$ to about 11% for $\bar{s} = 1000$.

A.5.1.3 Comparison of QL_{rt} and HOL_{rt}.

Interestingly, the difference in performance between HOL_{rt} and QL_{rt} is roughly independent of the number of servers, for large systems. That is consistent with simulation results for the $M(t)/M/s(t)+M$ model, and with prior theoretical results in chapters 2 and 3. Indeed, Table A.32 shows that $\text{ASE}(\text{HOL}_{rt})/\text{ASE}(\text{QL}_{rt})$ is roughly equal to 1.4, particularly for large $\bar{s}$. That is slightly larger than with M abandonment, where the ratio $\text{ASE}(\text{HOL}_{rt})/\text{ASE}(\text{QL}_{rt})$ is roughly equal to 1.3 for large $\bar{s}$; see Table A.34.

A.5.2 Results for the $M(t)/M/s(t)+E_{10}$ Model.

A.5.2.1 More reliable predictions in small systems.

Simulation results with E_{10} abandonment times are consistent with those obtained with M or H_2 abandonment, so we will be brief. With E_{10} abandonment, Table A.33 shows that all predictors are relatively more accurate than with M or H_2 abandonment, particularly when the number of servers is small ($\bar{s} \leq 20$). For example, $\text{RRASE}(\text{QL}_a^m)$ is roughly equal to 47% for $\bar{s} = 10$ (as opposed to 72% with H_2 abandonment, and 63% with M abandonment). Similarly, $\text{RRASE}(\text{QL}_{rt})$ is roughly equal to 52% for $\bar{s} = 10$ (as opposed to 74% with H_2 abandonment, and 67% with M abandonment). Consistent with §4.9 and §A.5.1, Table A.33 shows that all predictors are more accurate in large systems. Fluid-based predictors are particularly accurate in that case.

A.5.2.2 Superiority of fluid-based predictors.

As with M or H_2 abandonment times, there is no advantage in using the fluid-based predictors over the modified predictors when the number of servers is small. Indeed, QL_a^m is the most accurate predictor for small $\bar{s}$. For example, Table A.33 shows that $\text{ASE}(\text{QL}_a^m)/\text{ASE}(\text{QL}_{rt})$ is roughly equal to 0.8 for $\bar{s} = 10$. As the system size increases, fluid-based predictors gain in accuracy, compared to the remaining predictors. Figure A.3 shows that QL_{rt} and HOL_{rt} are more accurate than the remaining predictors for $\bar{s} \geq 40$. Also, consistent with Figures 5.3 and A.2, Figure A.4 shows that QL_{rt} and HOL_{rt} are asymptotically correct, unlike QL_a^m and HOL_a^m. Finally, as with M or H_2 abandonment times, the QL_{rt} predictor is the most accurate predictor for $\bar{s} \geq 30$. For example, $\text{ASE}(\text{QL}_a^m)/\text{ASE}(\text{QL}_{rt})$ ranges from about 1.2 (1.5 with M abandonment) for $\bar{s} = 20$ to about 17 (26 with M abandonment) for $\bar{s} = 1000$; see Tables A.33 and A.34.

A.5.2.3 Comparison of QL_{rt} and HOL_{rt}.

The difference in performance between QL_{rt} and HOL_{rt} decreases as the system size increases. Indeed, Table A.33 shows that $\text{ASE}(\text{HOL}_{rt})/\text{ASE}(\text{QL}_{rt})$ ranges from roughly 1.3 for $\bar{s} = 10$ (consistent with both M and H_2 abandonment) to roughly 1.1 for $\bar{s} = 1000$ (as opposed to 1.3 with M abandonment and 1.4 with H_2 abandonment). That is, the difference in performance between QL_{rt} and HOL_{rt} is less significant with E_{10} abandonment than with M or H_2 abandonment. We will see in §A.6 that QL_{rt} is even less accurate than HOL_{rt} in the $M(t)/E_{10}/s(t)+E_{10}$ model, with both E_{10} service and abandonment times.

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

In this section, we describe simulation results for the $M(t)/GI/s(t)+GI$ model. Our objective is to study the performance of the alternative delay predictors with both non-exponential service and abandonment-time distributions. We consider $\lambda(t)$ in (5.27) and $s(t)$ in (5.28). We let $\gamma_s = \gamma_a = 1.57$, which corresponds to $E[S] = 6$ hours with a 24 hour cycle. We let $\alpha_a = 0.5$ and $\alpha_s = 0.3$. We vary the average number of servers, $\bar{s}$, from 10 to 1000.

To consider both higher and lower variability relative to the exponential distribution considered previously, we consider H_2 and E_{10} service and abandonment-time distributions. In Tables A.35-A.38, we present point estimates of the ASE and half width of the 95% confidence intervals in the $M(t)/H_2/s(t) + H_2$, $M(t)/E_{10}/s(t) + H_2$, $M(t)/H_2/s(t) + E_{10}$, and $M(t)/E_{10}/s(t) + E_{10}$ models, respectively, as a function of $\bar{s}$. We also consider the case of D service times and present simulation results for the $M(t)/D/s(t) + H_2$ and $M(t)/D/s(t) + E_{10}$ models in Tables A.39 and A.40, respectively. However, we do not discuss these results separately, because they are largely consistent with those corresponding to E_{10} service times. The fluid model proposed in Lui and Whitt (2010) extends to non-exponential service times. Therefore, there remains the possibility to develop new fluid-based predictors based on the more general, and significantly more complicated, fluid model. We leave such extensions to future research. Here, we implement all predictors by approximating the service-time distribution by an exponential distribution with the same mean service time, $E[S]$.

A.6.1 H_2 Service Times

A.6.1.1 Less reliable predictions.

The H_2 distribution with SCV equal to 4 has higher variability relative to the M distribution. Tables A.35 and A.37 (compared with Tables A.32 and A.33, respectively) show that this extra variability makes all delay predictors relatively less accurate. For one example, in the $M(t)/H_2/s(t)+H_2$ model, RRASE(QL_{rt}) ranges from about 94% (74% with M service times) for $\bar{s}=10$ to about 23% (9% with M service times) for $\bar{s}=1000$; see Tables A.32 and A.35. For another example, in the $M(t)/H_2/s(t)+E_{10}$ model, RRASE(QL_a^m) ranges from about 94% (72% with M service times) for $\bar{s}=10$ to about 53% (41% with M service times) for $\bar{s}=1000$; see Tables A.33 and A.37. Similar results also hold for the remaining predictors.

A.6.1.2 Superiority of fluid-based predictors.

Figures 5.3, A.2, and A.4 showed that fluid-based predictors are asymptotically correct with M service times. With the incorrect fluid model, we no longer anticipate that the fluid-based predictors are asymptotically correct with H_2 service times. Indeed, Tables A.35 and A.37 show that the ASE's of fluid-based predictors are not inversely proportional to $\bar{s}$ in the $M(t)/H_2/s(t)+H_2$ and $M(t)/H_2/s(t)+E_{10}$ models, respectively. Nevertheless, fluid-based predictors remain more accurate than both QL_a^m and HOL_a^m in those models, particularly for large $\bar{s}$. For one example, in the $M(t)/H_2/s(t)+H_2$ model, ASE(HOL_a^m)/ASE(HOL_{rt}) ranges from about 1 (0.9 with M service times) for $\bar{s}=10$ to about 5 (16 with M service times) for $\bar{s}=1000$; see Tables A.32 and A.35. For another example, in the $M(t)/H_2/s(t)+E_{10}$ model, ASE(QL_a^m)/ASE(QL_{rt}) ranges from about 1.2 (0.8 with M service times) for $\bar{s}=10$

to about 2.5 (18 with M service times) for $\bar{s} = 1000$; see Tables A.33 and A.37. That is, the difference in performance between fluid-based and modified predictors remains significant with H_2 service times, but it is considerably less than with M service times.

A.6.1.3 Comparison of QL_{rt} and HOL_{rt}.

The QL_{rt} predictor is generally the most accurate predictor with M service times. In the $M(t)/M/s(t) + H_2$ model, Table A.32 showed that QL_{rt} outperforms the remaining predictors for $\bar{s} \geq 20$. In the $M(t)/M/s(t) + E_{10}$ model, Table A.33 showed that QL_{rt} outperforms the remaining predictors for $\bar{s} \geq 30$. The second most accurate predictor in both models is HOL_{rt}. With H_2 service times, QL_{rt} and HOL_{rt} remain the most accurate predictors, but they have nearly identical performance for large $\bar{s}$. For one example, in the $M(t)/H_2/s(t) + H_2$ model, $\mathrm{ASE}(\mathrm{HOL}_{rt})/\mathrm{ASE}(\mathrm{QL}_{rt})$ is roughly equal to 1.1 (1.4 with M service times) for $\bar{s} = 1000$; see Tables A.32 and A.35. For another example, in the $M(t)/H_2/s(t) + E_{10}$ model, $\mathrm{ASE}(\mathrm{HOL}_{rt})/\mathrm{ASE}(\mathrm{QL}_{rt})$ is roughly equal to 0.9 (1.0 with M service times) for $\bar{s} = 1000$; see Tables A.33 and A.37.

A.6.2 E_{10} Service Times

A.6.2.1 More/less reliable predictions.

The E_{10} distribution is less variable than the M distribution. Tables A.36 and A.38 (compared with Tables A.32 and A.33, respectively) show that this lower variability makes QL_a^m and HOL_a^m relatively more accurate and fluid-based predictors relatively less accurate, particularly for large $\bar{s}$. For one example, in the $M(t)/E_{10}/s(t) + H_2$

model, RRASE(HOL_a^m) ranges from about 67% (80% with M service times) for $\bar{s} = 10$ to about 22% (42% with M service times) for $\bar{s} = 1000$; see Tables A.32 and A.36. For another example, in the $M(t)/E_{10}/s(t) + E_{10}$ model, RRASE(QL_{rt}) ranges from about 43% (52% with M service times) for $\bar{s} = 10$ to about 26% (7% with M service times) for $\bar{s} = 1000$; see Tables A.33 and A.38.

A.6.2.2 Inferiority of fluid-based predictors.

With E_{10} service times, Tables A.36 and A.38 show that fluid-based predictors are not competitive with E_{10} service times, and are consistently less accurate than both QL_a^m and HOL_a^m (particularly for large $\bar{s}$). For example, in the $M(t)/E_{10}/s(t) + H_2$ model, ASE(QL_{rt})/ASE(QL_a^m) ranges from roughly 1.5 (1.0 with M service times) for $\bar{s} = 10$ to roughly 1.8 (0.05 with M service times!) for $\bar{s} = 1000$; see Tables A.32 and A.36. Similarly, in the $M(t)/E_{10}/s(t)+E_{10}$ model, ASE(QL_{rt})/ASE(QL_a^m) ranges from roughly 1.6 (1.2 with M service times) for $\bar{s} = 10$ to roughly 2.4 (0.05 with M service times!) for $\bar{s} = 1000$.

A.6.2.3 Comparison of QL_{rt} and HOL_{rt}.

With E_{10} service times, Tables A.36 and A.38 show that QL_{rt} performs slightly worse than HOL_{rt}, for large $\bar{s}$. For one example, in the $M(t)/E_{10}/s(t) + H_2$ model, ASE(QL_{rt})/ASE(HOL_{rt}) ranges from about 0.7 (0.8 with M service times) for $\bar{s} = 10$ to about 1.2 (0.7 with M service times) for $\bar{s} = 1000$; see Table A.32 and A.36. For another example, in the $M(t)/E_{10}/s(t) + E_{10}$ model, ASE(QL_{rt})/ASE(HOL_{rt}) ranges from about 0.7 (0.8 with M service times) for $\bar{s} = 10$ to about 1.1 (0.9 with M service times) for $\bar{s} = 1000$; see Tables A.33 and A.38.

A.6.2.4 Performance of NIF.

It is worthwhile mentioning that in the $M(t)/E_{10}/s(t)+E_{10}$ model, both QL_{rt} and HOL_{rt} are less accurate than NIF for $\bar{s} \geq 500$; see Table A.38. That may seem counterintuitive, at first glance, because both QL_{rt} and HOL_{rt} exploit information about current system state at the time of prediction, unlike NIF. However, these results should not be too surprising: All fluid-based predictors here are based on the incorrect fluid model, assuming an exponential service-time distribution. Therefore, they all make consistent prediction error. Indeed, QL_a^m performs considerably better than all fluid-based predictors in the $M(t)/E_{10}/s(t)+E_{10}$ model: Table A.38 shows that $\text{ASE(NIF)}/\text{ASE}(\text{QL}_a^m)$ is roughly equal to 2 for $\bar{s} = 1000$.

A.7 A Simple Modified QL_a Predictor: QL_a^{sm}

In this section, we propose a simple modified QL_a predictor, QL_a^{sm}. We define the QL_a^{sm} delay prediction as follows: We replace s in (5.7) by $s(t)$, the number of servers seen in the system upon arrival at time t. That is, we let

$$\theta_{QL_a^{sm}} = \sum_{i=0}^{n} \frac{1}{s(t)\mu + \delta_n - \delta_{n-i}} , \tag{A.1}$$

using the same notation as in (5.7); see §5.3.1. The QL_a^{sm} predictor is appealing because it is easier to implement than QL_a^m, defined in (5.10), and should be relatively accurate when the number of servers does not change too rapidly over time.

In this section, we compare the performance of QL_a^{sm}, QL_a, and QL_a^m in the $M(t)/M/s(t)+M$ model. We consider $\lambda(t)$ in (5.27) and $s(t)$ in (5.28). We let $\alpha_a = 0.5$ and $\alpha_s = 0.3$. We let the average number of servers, $\bar{s}$, range from 10 to 1000. In

Figures A.5 and A.6, we plot the ASE of QL_a^{sm}, QL_a, and QL_a^m, as a function of $\bar{s}$, in the $M(t)/M/s(t)+M$ model with $\gamma_a = \gamma_s = 0.022$, which corresponds to $E[S] = 5$ minutes with a 24 hour cycle. In Figures A.7 and A.8, we plot the ASE of QL_a^{sm}, QL_a, and QL_a^m, as a function of $\bar{s}$, in the $M(t)/M/s(t)+M$ model with $\gamma_a = \gamma_s = 1.57$, which corresponds to $E[S] = 6$ hours with a 24 hour cycle.

A.7.1 Performance of QL_a^{sm}, QL_a, and QL_a^m with Short Service Times

For small $E[S]$, as explained in §5.6.2.2, 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 which use $\lambda(t)$ and $s(t)$ corresponding to each point in time, such as QL_a^{sm}, will be accurate for small $E[S]$. Figures A.5 and A.6 confirm that QL_a^{sm} performs nearly as well as QL_a^m in that case (indeed, the two ASE curves roughly coincide). The ratio $ASE(QL_a^{sm})/ASE(QL_a^m)$ is approximately equal to 1.0 for all values of $\bar{s}$ considered. That is, with small $E[S]$, there is no advantage in using QL_a^m over QL_a^{sm}.

The difference in performance between QL_a and QL_a^{sm} (or, alternatively, QL_a^m) is not too great for small $\bar{s}$: Figure A.5 shows that $ASE(QLa)/ASE(QL_a^{sm})$ is roughly equal to 1.1 for $\bar{s} = 10$. However, as the number of servers increases, the difference in performance between those two predictors becomes significant: Figure A.6 shows that $ASE(QLa)/ASE(QL_a^{sm})$ is roughly equal to 16 for $\bar{s} = 1000$.

A.7.2 Performance of QL_a^{sm}, QL_a, and QL_a^m with Long Service Times

With large $E[S]$, the number of servers varies significantly over time. Therefore, we anticipate that QL_a^{sm} will be less effective than QL_a^m, since it assumes that the number of servers is constant over the waiting time of the arriving customer (and equal to the number of servers seen upon arrival). Figures A.7 and A.8 confirm this, but show that the difference in performance between QL_a^{sm} and QL_a^m is not too great. For one example, Figure A.7 shows that $ASE(QL_a^m)/ASE(QL_a^{sm})$ is roughly equal to 1.1 for $\bar{s} = 10$. For another example, Figure A.8 shows that $ASE(QL_a^m)/ASE(QL_a^{sm})$ is roughly equal to 1.3 for $\bar{s} = 1000$.

The QL_a^{sm} predictor is only slightly more effective than QL_a with a large $E[S]$. Indeed, Figures A.7 and A.8 show that $ASE(QL_a)/ASE(QL_a^{sm})$ is less than 1.02 for all values of $\bar{s}$ considered. That is, with large $E[S]$, simulation shows that there is no considerable advantage in using QL_a^m or QL_a^{sm} over QL_a. Recall from §4.9 that fluid-based predictors are remarkably accurate in that case, and that they significantly outperform both QL_a and QL_a^m.

A.8 Tables and Figures

Conditional ASE in the $M/M/s$ model with $s = 100$ for actual delays in $(E[\widehat{W|W > 0}], 2E[\widehat{W|W > 0}])$

ρ	$E[\widehat{W\|W>0}]$	$ASE(QL)$	$ASE(LES)$	$ASE(HOL)$	$ASE(RCS)$	$ASE(RCS{-}\sqrt{s})$	$ASE(LCS)$	$ASE(NI)$
0.99	9.833×10^{-1}	1.380×10^{-2}	2.802×10^{-2}	2.782×10^{-2}	3.052×10^{-2}	3.090×10^{-2}	4.884×10^{-2}	2.504×10^{-2}
	$\pm6.92\times10^{-2}$	$\pm9.71\times10^{-4}$	$\pm2.08\times10^{-3}$	$\pm2.08\times10^{-3}$	$\pm2.09\times10^{-3}$	$\pm2.09\times10^{-3}$	$\pm2.18\times10^{-3}$	$\pm3.89\times10^{-2}$
0.98	5.039×10^{-1}	7.006×10^{-3}	1.442×10^{-2}	1.421×10^{-2}	1.702×10^{-2}	1.744×10^{-2}	3.565×10^{-2}	6.453×10^{-2}
	$\pm2.48\times10^{-2}$	$\pm3.47\times10^{-4}$	$\pm7.41\times10^{-4}$	$\pm7.42\times10^{-4}$	$\pm7.62\times10^{-4}$	$\pm7.62\times10^{-4}$	$\pm9.67\times10^{-4}$	$\pm6.08\times10^{-3}$
0.95	2.028×10^{-1}	2.772×10^{-3}	5.924×10^{-3}	5.695×10^{-3}	8.711×10^{-3}	9.145×10^{-3}	2.339×10^{-2}	1.040×10^{-2}
	$\pm3.02\times10^{-3}$	$\pm5.46\times10^{-5}$	$\pm1.07\times10^{-4}$	$\pm1.06\times10^{-4}$	$\pm1.37\times10^{-4}$	$\pm1.44\times10^{-4}$	$\pm4.23\times10^{-4}$	$\pm2.58\times10^{-4}$
0.93	1.435×10^{-1}	1.925×10^{-3}	4.231×10^{-3}	3.996×10^{-3}	7.031×10^{-3}	7.421×10^{-3}	1.776×10^{-2}	5.263×10^{-3}
	$\pm1.78\times10^{-3}$	$\pm2.72\times10^{-5}$	$\pm6.17\times10^{-5}$	$\pm6.00\times10^{-5}$	$\pm8.37\times10^{-5}$	$\pm8.77\times10^{-5}$	$\pm2.67\times10^{-4}$	$\pm1.20\times10^{-4}$
0.90	9.929×10^{-2}	1.293×10^{-3}	2.982×10^{-3}	2.734×10^{-3}	5.564×10^{-3}	5.850×10^{-3}	1.202×10^{-2}	2.516×10^{-3}
	$\pm2.75\times10^{-3}$	$\pm4.18\times10^{-5}$	$\pm8.61\times10^{-5}$	$\pm8.47\times10^{-5}$	$\pm1.44\times10^{-4}$	$\pm1.55\times10^{-4}$	$\pm4.45\times10^{-4}$	$\pm1.27\times10^{-4}$

Table A.1: A comparison of the efficiency of different real-time delay predictors conditional on the level of delay observed for the $M/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the conditional average squared error (ASE) in the interval $(E[\widehat{W|W > 0}], 2E[\widehat{W|W > 0}])$. Each estimate is shown with the half width of the 95 percent confidence interval. The ASE's are measured in units of mean service time squared per customer.

Conditional ASE in the $M/M/s$ model with $s = 100$ for actual delays in $(2E[\widehat{W|W > 0}], 4E[\widehat{W|W > 0}])$

ρ	$ASE(QL)$	$ASE(LES)$	$ASE(HOL)$	$ASE(RCS)$	$ASE(RCS-\sqrt{s})$	$ASE(LCS)$	$ASE(NI)$
0.99	2.589×10^{-2}	5.128×10^{-2}	5.108×10^{-2}	5.373×10^{-2}	5.417×10^{-2}	7.176×10^{-2}	2.882
	$\pm1.56\times10^{-3}$	$\pm2.77\times10^{-3}$	$\pm2.77\times10^{-3}$	$\pm2.79\times10^{-3}$	$\pm2.79\times10^{-3}$	$\pm2.93\times10^{-3}$	$\pm3.81\times10^{-1}$
0.98	2.713×10^{-3}	5.842×10^{-3}	5.556×10^{-3}	9.434×10^{-3}	1.000×10^{-2}	2.731×10^{-2}	3.058×10^{-2}
	$\pm7.57\times10^{-5}$	$\pm1.57\times10^{-4}$	$\pm1.58\times10^{-4}$	$\pm1.89\times10^{-4}$	$\pm1.92\times10^{-4}$	$\pm6.12\times10^{-4}$	$\pm1.62\times10^{-3}$
0.95	1.359×10^{-2}	2.727×10^{-2}	2.706×10^{-2}	2.986×10^{-2}	3.028×10^{-2}	4.974×10^{-2}	7.834×10^{-1}
	$\pm6.75\times10^{-4}$	$\pm1.35\times10^{-3}$	$\pm1.35\times10^{-3}$	$\pm1.27\times10^{-3}$	$\pm1.26\times10^{-3}$	$\pm1.04\times10^{-3}$	$\pm9.44\times10^{-2}$
0.93	5.536×10^{-3}	1.141×10^{-2}	1.117×10^{-2}	1.445×10^{-2}	1.496×10^{-2}	3.762×10^{-2}	1.300×10^{-1}
	$\pm1.26\times10^{-4}$	$\pm1.47\times10^{-4}$	$\pm1.46\times10^{-4}$	$\pm1.56\times10^{-4}$	$\pm1.79\times10^{-4}$	$\pm6.06\times10^{-4}$	$\pm4.04\times10^{-3}$
0.90	3.886×10^{-2}	8.119×10^{-3}	7.867×10^{-3}	1.133×10^{-2}	1.187×10^{-2}	3.345×10^{-2}	6.387×10^{-2}
	$\pm5.65\times10^{-5}$	$\pm8.98\times10^{-5}$	$\pm8.77\times10^{-5}$	$\pm1.20\times10^{-4}$	1.22×10^{-4}	$\pm3.60\times10^{-4}$	$\pm1.73\times10^{-3}$

Table A.2: A comparison of the efficiency of different real-time delay predictors conditional on the level of delay observed for the $M/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the conditional average squared error (ASE) in the interval $(2E[\widehat{W|W > 0}], 4E[\widehat{W|W > 0}])$. Each estimate is shown with the half width of the 95 percent confidence interval. The ASE's are measured in units of mean service time squared per customer.

Conditional ASE in the $M/M/s$ model with $s = 100$ for actual delays in $(4E[\widehat{W|W > 0}], 6E[\widehat{W|W > 0}])$

ρ	$ASE(QL)$	$ASE(LES)$	$ASE(HOL)$	$ASE(RCS)$	$ASE(RCS\text{-}\sqrt{s})$	$ASE(LCS)$	$ASE(NI)$
0.99	4.937×10^{-2}	8.655×10^{-2}	8.632×10^{-2}	8.943×10^{-2}	9.008×10^{-2}	1.088×10^{-1}	11.586
	$\pm7.03\times10^{-3}$	$\pm6.89\times10^{-3}$	$\pm6.86\times10^{-3}$	$\pm7.24\times10^{-3}$	$\pm7.23\times10^{-3}$	$\pm1.02\times10^{-2}$	±1.25
0.98	2.479×10^{-2}	4.747×10^{-2}	4.725×10^{-2}	5.014×10^{-2}	5.057×10^{-2}	6.964×10^{-2}	3.542
	$\pm1.75\times10^{-3}$	$\pm3.06\times10^{-3}$	$\pm3.04\times10^{-3}$	$\pm3.09\times10^{-3}$	$\pm3.13\times10^{-3}$	$\pm3.67\times10^{-3}$	$\pm4.31\times10^{-1}$
0.95	1.052×10^{-2}	2.037×10^{-2}	2.011×10^{-2}	2.350×10^{-2}	2.403×10^{-2}	5.039×10^{-2}	5.641×10^{-1}
	$\pm2.30\times10^{-4}$	$\pm6.30\times10^{-4}$	$\pm6.16\times10^{-4}$	$\pm8.16\times10^{-4}$	$\pm8.00\times10^{-4}$	$\pm3.34\times10^{-3}$	$\pm2.70\times10^{-2}$
0.93	7.544×10^{-3}	1.515×10^{-2}	1.487×10^{-2}	1.870×10^{-2}	1.930×10^{-2}	5.204×10^{-2}	2.860×10^{-1}
	$\pm1.97\times10^{-3}$	$\pm3.08\times10^{-4}$	$\pm2.90\times10^{-4}$	$\pm4.28\times10^{-4}$	$\pm4.52\times10^{-4}$	$\pm3.18\times10^{-3}$	$\pm8.01\times10^{-3}$
0.90	5.622×10^{-3}	1.105×10^{-2}	1.073×10^{-2}	1.531×10^{-2}	1.609×10^{-2}	5.089×10^{-2}	1.374×10^{-1}
	$\pm2.07\times10^{-4}$	$\pm3.82\times10^{-4}$	$\pm3.75\times10^{-4}$	$\pm6.07\times10^{-4}$	$\pm6.62\times10^{-4}$	$\pm2.52\times10^{-2}$	$\pm6.74\times10^{-3}$

Table A.3: A comparison of the efficiency of different real-time delay predictors conditional on the level of delay observed for the $M/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the conditional average squared error (ASE) in the interval $(4E[\widehat{W|W > 0}], 6E[\widehat{W|W > 0}])$. Each estimate is shown with the half width of the 95 percent confidence interval. The ASE's are measured in units of mean service time squared per customer.

Conditional ASE in the $M/M/s$ model with $s = 100$ for actual delays $> 6E[\widehat{W|W > 0}]$)

ρ	$ASE(QL)$	$ASE(LES)$	$ASE(HOL)$	$ASE(RCS)$	$ASE(RCS\text{-}\sqrt{s})$	$ASE(LCS)$	$ASE(NI)$
0.98	3.361×10^{-2}	7.526×10^{-2}	7.521×10^{-2}	7.735×10^{-2}	7.776×10^{-2}	9.012×10^{-2}	7.902
	$\pm1.21\times10^{-2}$	$\pm2.61\times10^{-2}$	$\pm2.60\times10^{-2}$	$\pm2.63\times10^{-2}$	$\pm2.64\times10^{-2}$	$\pm2.30\times10^{-2}$	±1.29
0.95	1.862×10^{-2}	3.353×10^{-2}	3.319×10^{-2}	3.755×10^{-2}	3.820×10^{-2}	6.823×10^{-2}	1.496
	$\pm2.22\times10^{-3}$	$\pm4.14\times10^{-3}$	$\pm4.13\times10^{-3}$	$\pm4.20\times10^{-3}$	$\pm4.49\times10^{-3}$	$\pm8.10\times10^{-3}$	$\pm1.65\times10^{-1}$
0.93	1.306×10^{-2}	2.306×10^{-2}	2.274×10^{-2}	2.739×10^{-2}	2.803×10^{-2}	6.603×10^{-2}	7.362×10^{-1}
	$\pm1.26\times10^{-3}$	$\pm3.16\times10^{-3}$	$\pm3.06\times10^{-3}$	$\pm3.82\times10^{-3}$	$\pm3.96\times10^{-3}$	$\pm1.07\times10^{-2}$	$\pm6.29\times10^{-2}$
0.90	1.042×10^{-2}	1.872×10^{-2}	1.828×10^{-2}	2.378×10^{-2}	2.438×10^{-2}	7.160×10^{-2}	0.3478
	$\pm9.67\times10^{-4}$	$\pm8.63\times10^{-4}$	$\pm8.65\times10^{-4}$	$\pm1.41\times10^{-3}$	$\pm1.34\times10^{-3}$	$\pm6.04\times10^{-3}$	$\pm3.67\times10^{-2}$

Table A.4: A comparison of the efficiency of different real-time delay predictors conditional on the level of delay observed for the $M/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the conditional average squared error (ASE) when delays are larger than $6E[\widehat{W|W > 0}]$. Each estimate is shown with the half width of the 95 percent confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE in the $M/M/s$ model with $s = 100$

ρ	$ASE(RCS)$	$ASE(RCS\text{–}s)$	$ASE(RCS\text{–}4\sqrt{s})$	$ASE(RCS\text{–}2\sqrt{s})$	$ASE(RCS\text{–}\sqrt{s})$	$ASE(RCS\text{–}\log s)$
0.98	9.439×10^{-3}	9.439×10^{-3}	9.439×10^{-3}	9.452×10^{-3}(**0.138**)	9.779×10^{-3}(**3.60**)	1.548×10^{-2}(**64.0**)
	$\pm3.13\times10^{-4}$	$\pm3.13\times10^{-4}$	$\pm3.13\times10^{-4}$	$\pm3.13\times10^{-4}$	$\pm3.18\times10^{-4}$	$\pm4.50\times10^{-4}$
0.97	8.070×10^{-3}	8.070×10^{-3}	8.070×10^{-3}	8.083×10^{-3}(**0.161**)	8.395×10^{-3}(**4.03**)	1.388×10^{-2}(**72.0**)
	$\pm1.71\times10^{-4}$	$\pm1.71\times10^{-4}$	$\pm1.71\times10^{-4}$	$\pm1.72\times10^{-4}$	$\pm1.80\times10^{-4}$	$\pm3.34\times10^{-4}$
0.95	6.280×10^{-3}	6.280×10^{-3}	6.280×10^{-3}	6.295×10^{-3}(**0.239**)	6.571×10^{-3}(**4.63**)	1.135×10^{-2}(**80.7**)
	$\pm1.78\times10^{-4}$	$\pm1.78\times10^{-4}$	$\pm1.78\times10^{-4}$	$\pm1.79\times10^{-4}$	$\pm1.82\times10^{-4}$	$\pm3.01\times10^{-4}$
0.93	4.908×10^{-3}	4.908×10^{-3}	4.908×10^{-3}	4.918×10^{-3}(**0.204**)	5.161×10^{-3}(**5.15**)	9.017×10^{-3}(**83.7**)
	$\pm1.22\times10^{-4}$	$\pm1.22\times10^{-4}$	$\pm1.22\times10^{-4}$	$\pm1.23\times10^{-4}$	$\pm1.23\times10^{-4}$	$\pm1.91\times10^{-4}$
0.9	3.897×10^{-3}	3.897×10^{-3}	3.897×10^{-3}	3.906×10^{-3}(**0.696**)	4.108×10^{-3}(**5.41**)	6.892×10^{-3}(**76.9**)
	$\pm9.62\times10^{-5}$	$\pm9.62\times10^{-5}$	$\pm9.62\times10^{-5}$	$\pm9.62\times10^{-5}$	$\pm1.01\times10^{-4}$	$\pm1.92\times10^{-4}$

Table A.5: A comparison of the efficiency of the candidate RCS-$f(s)$ delay predictors for the $M/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the average squared error – (ASE). Each estimate is shown with the half width of the 95 percent confidence interval. Also included in parentheses are the values of the relative percent difference – (RPD) compared with ASE(RCS). The ASE's are measured in units of mean service time squared per customer.

ASE in the $D/M/s$ model with $s = 100$

ρ	$ASE(RCS)$	$ASE(RCS\text{–}s)$	$ASE(RCS\text{–}4\sqrt{s})$	$ASE(RCS\text{–}2\sqrt{s})$	$ASE(RCS\text{–}\sqrt{s})$	$ASE(RCS\text{–}\log s)$
0.98	3.617×10^{-3}	3.617×10^{-3}	3.617×10^{-3}	3.624×10^{-3}(**0.194**)	3.789×10^{-3}(**4.76**)	6.678×10^{-3}(**84.6**)
	$\pm9.75\times10^{-5}$	$\pm9.75\times10^{-5}$	$\pm9.75\times10^{-5}$	$\pm9.75\times10^{-5}$	$\pm9.10\times10^{-5}$	$\pm1.27\times10^{-4}$
0.97	2.906×10^{-3}	2.906×10^{-3}	2.906×10^{-3}	2.913×10^{-3}(**0.241**)	3.066×10^{-3}(**5.51**)	5.665×10^{-3}(**94.9**)
	$\pm6.80\times10^{-5}$	$\pm6.80\times10^{-5}$	$\pm6.80\times10^{-5}$	$\pm7.13\times10^{-5}$	$\pm6.81\times10^{-5}$	$\pm1.22\times10^{-4}$
0.95	2.200×10^{-3}	2.200×10^{-3}	2.200×10^{-3}	2.205×10^{-3}(**0.227**)	2.337×10^{-3}(**6.23**)	4.406×10^{-3}(**100**)
	$\pm4.40\times10^{-5}$	$\pm4.40\times10^{-5}$	$\pm4.40\times10^{-5}$	$\pm4.42\times10^{-5}$	$\pm4.38\times10^{-5}$	$\pm1.13\times10^{-4}$
0.93	1.852×10^{-3}	1.852×10^{-3}	1.852×10^{-3}	1.856×10^{-3}(**0.216**)	1.966×10^{-3}(**6.16**)	3.594×10^{-3}(**94.1**)
	$\pm4.14\times10^{-5}$	$\pm4.14\times10^{-5}$	$\pm4.14\times10^{-5}$	$\pm4.22\times10^{-5}$	$\pm4.24\times10^{-5}$	$\pm1.06\times10^{-4}$

Table A.6: A comparison of the efficiency of the candidate RCS-$f(s)$ delay predictors for the $D/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the average squared error – (ASE). Each estimate is shown with the half width of the 95 percent confidence interval. Also included in parentheses are the values of the relative percent difference – (RPD) compared with ASE(RCS). The ASE's are measured in units of mean service time squared per customer.

ASE in the $H_2/M/s$ model with $s = 100$

ρ	$ASE(RCS)$	$ASE(RCS{-}s)$	$ASE(RCS{-}4\sqrt{s})$	$ASE(RCS{-}2\sqrt{s})$	$ASE(RCS{-}\sqrt{s})$	$ASE(RCS{-}\log s)$
0.98	2.439×10^{-2}	2.439×10^{-2}	2.439×10^{-2}	2.442×10^{-2}(**0.123**)	2.511×10^{-2}(**2.95**)	3.724×10^{-2}(**52.7**)
	$\pm4.84\times10^{-4}$	$\pm4.84\times10^{-4}$	$\pm4.84\times10^{-4}$	$\pm4.88\times10^{-4}$	$\pm4.92\times10^{-4}$	$\pm6.81\times10^{-4}$
0.97	2.229×10^{-2}	2.229×10^{-2}	2.229×10^{-2}	2.229×10^{-2}(**0.141**)	2.367×10^{-2}(**3.28**)	3.566×10^{-2}(**55.6**)
	$\pm4.70\times10^{-4}$	$\pm4.70\times10^{-4}$	$\pm4.70\times10^{-4}$	$\pm4.73\times10^{-4}$	$\pm4.73\times10^{-4}$	$\pm5.80\times10^{-4}$
0.95	1.989×10^{-2}	1.989×10^{-2}	1.989×10^{-2}	1.992×10^{-2}(**0.136**)	2.058×10^{-2}(**3.48**)	3.175×10^{-2}(**59.6**)
	$\pm3.67\times10^{-4}$	$\pm3.67\times10^{-4}$	$\pm3.67\times10^{-4}$	$\pm3.67\times10^{-4}$	$\pm3.64\times10^{-4}$	$\pm5.45\times10^{-4}$
0.93	1.715×10^{-2}	1.715×10^{-2}	1.715×10^{-2}	1.718×10^{-2}(**0.150**)	1.780×10^{-2}(**3.78**)	2.800×10^{-2}(**63.2**)
	$\pm3.56\times10^{-4}$	$\pm3.56\times10^{-4}$	$\pm3.56\times10^{-4}$	$\pm3.54\times10^{-4}$	$\pm3.60\times10^{-4}$	$\pm5.89\times10^{-4}$
0.90	1.344×10^{-2}	1.344×10^{-2}	1.344×10^{-2}	1.347×10^{-2}(**0.182**)	1.399×10^{-2}(**4.06**)	2.233×10^{-2}(**66.3**)
	$\pm4.90\times10^{-4}$	$\pm4.90\times10^{-4}$	$\pm4.90\times10^{-4}$	$\pm4.89\times10^{-4}$	$\pm4.99\times10^{-4}$	$\pm8.61\times10^{-4}$

Table A.7: A comparison of the efficiency of the candidate RCS-$f(s)$ delay predictors for the $H_2/M/s$ queue with $s = 100$ and $\mu = 1$ as a function of the traffic intensity ρ. We report point estimates for the average squared error – (ASE). Each estimate is shown with the half width of the 95 percent confidence interval. Also included in parentheses are the values of the relative percent difference – (RPD) compared with ASE(RCS). The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $M/M/s + M$ model with $\rho = 1.4$ and $\nu = 5.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	6.318×10^{-4}	7.172×10^{-4}	1.226×10^{-3}	1.391×10^{-3}	1.809×10^{-3}
	$\pm1.53\times10^{-6}$	$\pm2.42\times10^{-6}$	$\pm5.06\times10^{-6}$	$\pm3.09\times10^{-6}$	$\pm7.22\times10^{-6}$
300	1.935×10^{-4}	2.130×10^{-4}	4.813×10^{-4}	4.035×10^{-4}	6.591×10^{-4}
	$\pm6.54\times10^{-7}$	$\pm8.69\times10^{-7}$	$\pm1.86\times10^{-6}$	$\pm1.15\times10^{-6}$	$\pm3.06\times10^{-6}$
500	1.151×10^{-4}	1.253×10^{-4}	3.467×10^{-4}	2.361×10^{-4}	4.009×10^{-4}
	$\pm5.41\times10^{-7}$	$\pm4.54\times10^{-7}$	$\pm8.45\times10^{-7}$	$\pm8.67\times10^{-7}$	$\pm3.05\times10^{-6}$
700	8.235×10^{-5}	8.965×10^{-5}	2.963×10^{-4}	1.675×10^{-4}	2.872×10^{-4}
	$\pm4.04\times10^{-7}$	$\pm3.51\times10^{-7}$	$\pm9.01\times10^{-7}$	$\pm8.21\times10^{-7}$	$\pm2.58\times10^{-6}$
1000	5.772×10^{-5}	6.261×10^{-5}	2.555×10^{-4}	1.167×10^{-4}	2.022×10^{-4}
	$\pm2.33\times10^{-7}$	$\pm2.66\times10^{-7}$	$\pm5.44\times10^{-7}$	$\pm6.87\times10^{-7}$	$\pm2.15\times10^{-6}$

Table A.8: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/M/s + M$ model with $\nu = 5.0$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $M/M/s+M$ model with $\rho = 1.4$ and $\nu = 0.2$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	1.425×10^{-2}	1.545×10^{-2}	1.238×10^{-1}	2.894×10^{-2}	4.963×10^{-2}
	$\pm1.15\times10^{-4}$	$\pm1.22\times10^{-4}$	$\pm5.16\times10^{-4}$	$\pm3.52\times10^{-4}$	$\pm6.31\times10^{-4}$
300	4.705×10^{-3}	5.099×10^{-3}	1.094×10^{-1}	9.573×10^{-3}	1.657×10^{-2}
	$\pm5.33\times10^{-5}$	$\pm5.95\times10^{-5}$	$\pm5.04\times10^{-4}$	$\pm1.20\times10^{-4}$	$\pm4.98\times10^{-4}$
500	2.879×10^{-3}	3.103×10^{-3}	1.046×10^{-1}	5.832×10^{-3}	9.926×10^{-3}
	$\pm4.27\times10^{-5}$	$\pm3.70\times10^{-5}$	$\pm4.19\times10^{-4}$	$\pm7.88\times10^{-5}$	$\pm5.18\times10^{-4}$
700	2.029×10^{-3}	2.194×10^{-3}	1.0479×10^{-1}	4.150×10^{-3}	7.121×10^{-3}
	$\pm2.62\times10^{-5}$	$\pm3.34\times10^{-5}$	$\pm5.54\times10^{-4}$	$\pm1.09\times10^{-4}$	$\pm2.25\times10^{-4}$
1000	1.444×10^{-3}	1.558×10^{-3}	1.031×10^{-1}	2.995×10^{-3}	4.935×10^{-3}
	$\pm4.43\times10^{-5}$	$\pm4.35\times10^{-5}$	$\pm3.47\times10^{-4}$	$\pm6.03\times10^{-5}$	$\pm3.74\times10^{-4}$

Table A.9: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $M/M/s+M$ model with $\nu = 0.2$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $D/M/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.882×10^{-3}	2.994×10^{-3}	7.705×10^{-3}	6.545×10^{-3}	6.496×10^{-3}
	$\pm7.89\times10^{-6}$	$\pm8.28\times10^{-6}$	$\pm1.22\times10^{-5}$	$\pm1.12\times10^{-5}$	$\pm3.60\times10^{-5}$
300	9.520×10^{-4}	9.903×10^{-4}	5.256×10^{-3}	1.243×10^{-3}	2.188×10^{-3}
	$\pm4.42\times10^{-6}$	$\pm4.73\times10^{-6}$	$\pm8.05\times10^{-6}$	$\pm5.70\times10^{-6}$	$\pm2.50\times10^{-5}$
500	5.753×10^{-4}	5.989×10^{-4}	4.774×10^{-3}	7.537×10^{-4}	1.297×10^{-3}
	$\pm3.51\times10^{-6}$	$\pm3.87\times10^{-6}$	$\pm5.70\times10^{-6}$	$\pm5.44\times10^{-6}$	$\pm1.91\times10^{-5}$
700	4.096×10^{-4}	4.260×10^{-4}	4.548×10^{-3}	9.149×10^{-4}	9.537×10^{-4}
	$\pm3.18\times10^{-6}$	$\pm3.52\times10^{-6}$	$\pm8.71\times10^{-6}$	$\pm4.42\times10^{-6}$	$\pm1.68\times10^{-5}$
1000	2.871×10^{-4}	2.979×10^{-4}	4.392×10^{-3}	3.912×10^{-4}	6.697×10^{-4}
	$\pm3.66\times10^{-6}$	$\pm3.23\times10^{-6}$	$\pm5.34\times10^{-6}$	$\pm5.17\times10^{-6}$	$\pm2.01\times10^{-5}$

Table A.10: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $D/M/s+M$ model with $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $D/M/s+M$ model with $\rho = 1.4$ and $\nu = 5.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	6.0637×10^{-4}	6.336×10^{-4}	9.340×10^{-4}	9.018×10^{-4}	1.285×10^{-3}
	$\pm1.46\times10^{-6}$	$\pm1.28\times10^{-6}$	$\pm9.96\times10^{-7}$	$\pm1.65\times10^{-6}$	$\pm4.81\times10^{-6}$
300	1.929×10^{-4}	2.011×10^{-4}	4.081×10^{-4}	2.625×10^{-4}	4.329×10^{-4}
	$\pm6.27\times10^{-7}$	$\pm6.51\times10^{-7}$	$\pm9.46\times10^{-7}$	$\pm9.16\times10^{-7}$	$\pm1.84\times10^{-6}$
500	1.150×10^{-4}	1.196×10^{-4}	3.084×10^{-4}	1.528×10^{-4}	2.606×10^{-4}
	$\pm3.00\times10^{-7}$	$\pm3.71\times10^{-7}$	$\pm7.29\times10^{-7}$	$\pm4.14\times10^{-7}$	$\pm1.23\times10^{-6}$
700	8.218×10^{-5}	8.545×10^{-5}	2.663×10^{-4}	1.082×10^{-4}	1.858×10^{-4}
	$\pm3.09\times10^{-7}$	$\pm3.22\times10^{-7}$	$\pm3.75\times10^{-7}$	$\pm2.97\times10^{-7}$	$\pm1.37\times10^{-6}$
1000	5.718×10^{-5}	5.950×10^{-5}	2.343×10^{-4}	7.475×10^{-5}	1.274×10^{-4}
	$\pm3.74\times10^{-7}$	$\pm4.16\times10^{-7}$	$\pm4.70\times10^{-7}$	$\pm5.12\times10^{-7}$	$\pm1.05\times10^{-6}$

Table A.11: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $D/M/s+M$ model with $\nu = 5.0$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $H_2/M/s+M$ model with $\rho = 1.4$ and $\nu = 0.2$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	1.429×10^{-2}	1.731×10^{-2}	1.370×10^{-1}	5.866×10^{-2}	1.018×10^{-1}
	$\pm8.51\times10^{-5}$	1.50×10^{-4}	9.49×10^{-4}	$\pm4.87\times10^{-4}$	$\pm1.91\times10^{-3}$
300	4.805×10^{-3}	5.747×10^{-3}	1.141×10^{-1}	2.044×10^{-2}	3.426×10^{-2}
	$\pm1.12\times10^{-4}$	$\pm1.10\times10^{-4}$	$\pm1.01\times10^{-3}$	$\pm3.03\times10^{-4}$	$\pm9.58\times10^{-4}$
500	2.865×10^{-3}	3.419×10^{-3}	1.080×10^{-1}	1.189×10^{-2}	2.043×10^{-2}
	$\pm5.39\times10^{-5}$	$\pm6.99\times10^{-5}$	$\pm1.33\times10^{-3}$	$\pm1.768\times10^{-4}$	$\pm9.72\times10^{-4}$
700	2.046×10^{-3}	2.456×10^{-3}	1.070×10^{-1}	8.471×10^{-3}	1.50×10^{-2}
	$\pm5.45\times10^{-5}$	$\pm6.68\times10^{-5}$	$\pm9.09\times10^{-4}$	$\pm1.72\times10^{-4}$	$\pm8.12\times10^{-4}$
1000	1.422×10^{-3}	1.686×10^{-3}	1.037×10^{-1}	5.962×10^{-3}	9.843×10^{-3}
	$\pm3.61\times10^{-5}$	$\pm3.68\times10^{-5}$	$\pm1.06\times10^{-3}$	$\pm1.24\times10^{-4}$	$\pm4.99\times10^{-4}$

Table A.12: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $H_2/M/s+M$ model with $\nu = 0.2$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of the predictors in the $H_2/M/s+M$ model with $\rho = 1.4$ and $\nu = 1.0$

s	ASE$[\theta_{QL_m}]$	ASE$[\theta_{QL_r}]$	ASE$[\theta_{QL}]$	ASE$[\theta_{LES}]$	ASE$[\theta_{NI}]$
100	2.898×10^{-3}	3.712×10^{-3}	1.190×10^{-2}	1.129×10^{-2}	1.900×10^{-2}
	$\pm1.17\times10^{-5}$	$\pm1.65\times10^{-5}$	$\pm5.26\times10^{-5}$	$\pm5.73\times10^{-5}$	$\pm1.31\times10^{-4}$
300	9.531×10^{-4}	1.173×10^{-3}	6.652×10^{-3}	3.863×10^{-3}	6.829×10^{-3}
	$\pm4.79\times10^{-6}$	$\pm9.07\times10^{-6}$	$\pm5.38\times10^{-5}$	$\pm2.15\times10^{-5}$	$\pm1.09\times10^{-4}$
500	5.701×10^{-4}	6.903×10^{-4}	5.502×10^{-3}	2.346×10^{-3}	4.118×10^{-3}
	$\pm3.01\times10^{-6}$	$\pm2.97\times10^{-6}$	$\pm3.12\times10^{-5}$	$\pm1.78\times10^{-5}$	$\pm4.94\times10^{-5}$
700	4.120×10^{-4}	4.9888×10^{-4}	5.143×10^{-3}	1.694×10^{-3}	2.939×10^{-3}
	$\pm2.38\times10^{-6}$	$\pm3.76\times10^{-6}$	$\pm2.83\times10^{-5}$	$\pm1.28\times10^{-5}$	$\pm5.86\times10^{-5}$
1000	2.870×10^{-4}	3.477×10^{-4}	4.780×10^{-3}	1.211×10^{-3}	2.117×10^{-3}
	$\pm3.33\times10^{-6}$	$\pm2.69\times10^{-6}$	$\pm3.30\times10^{-5}$	$\pm1.70\times10^{-5}$	$\pm5.85\times10^{-5}$

Table A.13: Point and confidence interval estimates of the ASEs - average square errors - of the predictors in the $H_2/M/s+M$ model with $\nu = 1.0$. The ASE's are measured in units of mean service time squared per customer.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/D/s+M$ Model

s	QL_m	HOL_a	HOL
100	4.160×10^{-3}	6.038×10^{-3}	1.740×10^{-2}
	$\pm7.26\times10^{-4}$	$\pm7.40\times10^{-4}$	$\pm9.66\times10^{-4}$
300	2.909×10^{-3}	3.552×10^{-3}	1.378×10^{-2}
	$\pm3.78\times10^{-4}$	$\pm3.97\times10^{-4}$	$\pm5.69\times10^{-4}$
500	2.729×10^{-3}	3.181×10^{-3}	1.319×10^{-2}
	$\pm6.33\times10^{-4}$	$\pm6.53\times10^{-4}$	$\pm8.80\times10^{-4}$
700	2.730×10^{-3}	2.971×10^{-3}	1.277×10^{-2}
	$\pm2.97\times10^{-4}$	$\pm2.91\times10^{-4}$	$\pm4.14\times10^{-4}$
1000	2.963×10^{-3}	3.165×10^{-3}	1.286×10^{-2}
	$\pm4.25\times10^{-4}$	$\pm4.57\times10^{-4}$	$\pm5.83\times10^{-4}$

Table A.14: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours and $\rho = 1.4$. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/H_2/s + M$ Model

s	QL_m	HOL_a	HOL
100	3.927×10^{-3}	5.981×10^{-3}	1.618×10^{-2}
	$\pm5.39\times10^{-4}$	$\pm6.63\times10^{-4}$	$\pm1.22\times10^{-3}$
300	1.214×10^{-3}	1.885×10^{-3}	1.086×10^{-2}
	$\pm9.46\times10^{-5}$	$\pm1.14\times10^{-4}$	$\pm5.24\times10^{-4}$
500	7.404×10^{-4}	1.137×10^{-3}	1.018×10^{-2}
	$\pm5.91\times10^{-5}$	$\pm8.52\times10^{-5}$	$\pm6.05\times10^{-4}$
700	5.542×10^{-4}	7.844×10^{-4}	9.763×10^{-3}
	$\pm3.36\times10^{-5}$	$\pm3.50\times10^{-5}$	$\pm2.65\times10^{-4}$
1000	3.760×10^{-4}	5.655×10^{-4}	9.189×10^{-3}
	$\pm2.56\times10^{-5}$	$\pm2.82\times10^{-5}$	$\pm2.42\times10^{-4}$

Table A.15: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours and $\rho = 1.4$. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/D/s + H_2$ Model

s	QL_m	HOL_a	HOL
100	4.675×10^{-3}	4.959×10^{-3}	9.991×10^{-3}
	$\pm3.20\times10^{-4}$	$\pm3.60\times10^{-4}$	$\pm5.07\times10^{-4}$
300	3.732×10^{-3}	3.158×10^{-3}	7.574×10^{-3}
	$\pm1.50\times10^{-4}$	$\pm1.35\times10^{-4}$	$\pm2.48\times10^{-4}$
500	3.454×10^{-3}	2.723×10^{-3}	7.044×10^{-3}
	$\pm1.84\times10^{-4}$	$\pm1.59\times10^{-4}$	$\pm3.08\times10^{-4}$
700	3.309×10^{-3}	2.552×10^{-3}	6.783×10^{-3}
	$\pm1.19\times10^{-4}$	$\pm1.19\times10^{-4}$	$\pm1.96\times10^{-4}$
1000	3.269×10^{-3}	2.433×10^{-3}	6.585×10^{-3}
	$\pm7.00\times10^{-5}$	$\pm7.07\times10^{-5}$	$\pm1.04\times10^{-4}$

Table A.16: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours and $\rho = 1.4$. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/H_2/s + H_2$ Model

s	QL_m	HOL_a	HOL
100	3.307×10^{-3}	3.972×10^{-3}	7.816×10^{-3}
	$\pm2.10\times10^{-4}$	$\pm3.58\times10^{-4}$	$\pm6.48\times10^{-4}$
300	1.642×10^{-3}	1.285×10^{-3}	4.636×10^{-3}
	$\pm1.38\times10^{-4}$	$\pm9.20\times10^{-5}$	$\pm2.87\times10^{-4}$
500	1.282×10^{-3}	7.739×10^{-4}	3.985×10^{-3}
	$\pm9.09\times10^{-5}$	$\pm3.40\times10^{-5}$	$\pm1.79\times10^{-4}$
700	1.155×10^{-3}	5.510×10^{-4}	3.862×10^{-3}
	$\pm8.24\times10^{-5}$	$\pm3.34\times10^{-5}$	$\pm1.50\times10^{-4}$
1000	1.099×10^{-3}	4.310×10^{-4}	3.557×10^{-3}
	$\pm5.33\times10^{-5}$	$\pm1.80\times10^{-5}$	$\pm8.50\times10^{-5}$

Table A.17: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours and $\rho = 1.4$. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/D/s + E_{10}$ Model

s	QL_m	HOL_a	HOL
100	2.341×10^{-2}	2.182×10^{-2}	6.829×10^{-2}
	$\pm3.82\times10^{-3}$	$\pm4.15\times10^{-3}$	$\pm7.04\times10^{-3}$
300	2.355×10^{-2}	2.101×10^{-2}	6.645×10^{-2}
	$\pm1.60\times10^{-3}$	$\pm1.78\times10^{-3}$	$\pm2.19\times10^{-3}$
500	2.433×10^{-2}	2.119×10^{-2}	6.674×10^{-2}
	$\pm1.97\times10^{-3}$	$\pm2.16\times10^{-3}$	$\pm3.55\times10^{-3}$
700	2.374×10^{-2}	2.033×10^{-2}	6.585×10^{-2}
	$\pm1.10\times10^{-3}$	$\pm1.29\times10^{-3}$	$\pm1.86\times10^{-3}$
1000	2.372×10^{-2}	1.997×10^{-2}	6.492×10^{-2}
	$\pm8.48\times10^{-4}$	$\pm9.19\times10^{-4}$	$\pm1.33\times10^{-3}$

Table A.18: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours and $\rho = 1.4$. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

Efficiency of QL_m, HOL_a, and HOL in the $M(t)/H_2/s + E_{10}$ Model

s	QL_m	HOL_a	HOL
100	9.970×10^{-3}	7.130×10^{-3}	4.190×10^{-2}
	$\pm6.08\times10^{-4}$	$\pm4.09\times10^{-4}$	$\pm1.82\times10^{-3}$
300	7.599×10^{-3}	3.025×10^{-3}	3.746×10^{-2}
	$\pm2.80\times10^{-4}$	$\pm2.18\times10^{-4}$	$\pm9.82\times10^{-4}$
500	7.097×10^{-3}	1.983×10^{-3}	3.589×10^{-2}
	$\pm2.57\times10^{-4}$	$\pm8.23\times10^{-5}$	$\pm3.96\times10^{-4}$
700	6.373×10^{-3}	1.741×10^{-3}	3.554×10^{-2}
	$\pm1.70\times10^{-4}$	$\pm1.01\times10^{-4}$	$\pm3.13\times10^{-4}$
1000	6.548×10^{-3}	1.016×10^{-3}	3.469×10^{-2}
	$\pm1.72\times10^{-4}$	$\pm7.86\times10^{-5}$	$\pm5.58\times10^{-4}$

Table A.19: A comparison of the efficiency of QL_m, HOL_a, and HOL as a function of the number of servers s, for sinusoidal arrival rates with $\bar{\lambda}$ and μ corresponding to a mean service time of 6 hours and $\rho = 1.4$. Point and 95% confidence interval estimates of the ASE's are shown. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/M/100$ model with $\alpha = 0.1$ and $E[S] = 5$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	9.48×10^{-3}	7.18×10^{-3}	6.60×10^{-3}	6.48×10^{-3}	6.55×10^{-3}	6.96×10^{-3}	8.31×10^{-3}	3.37×10^{-3}	6.74×10^{-3}
	$\pm8.2\times10^{-4}$	$\pm5.3\times10^{-4}$	$\pm4.7\times10^{-4}$	$\pm4.8\times10^{-4}$	$\pm5.2\times10^{-4}$	$\pm6.2\times10^{-4}$	$\pm8.9\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm5.0\times10^{-4}$
0.93	4.11×10^{-2}	2.42×10^{-2}	2.01×10^{-2}	1.94×10^{-2}	2.02×10^{-2}	2.37×10^{-2}	3.47×10^{-2}	9.79×10^{-3}	2.15×10^{-2}
	$\pm1.5\times10^{-3}$	$\pm7.5\times10^{-4}$	$\pm5.7\times10^{-4}$	$\pm5.8\times10^{-4}$	$\pm6.7\times10^{-4}$	$\pm9.5\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm2.8\times10^{-4}$	$\pm7.6\times10^{-4}$
0.95	0.102	5.07×10^{-2}	3.82×10^{-2}	3.62×10^{-2}	3.85×10^{-2}	4.91×10^{-2}	8.23×10^{-2}	1.82×10^{-2}	4.67×10^{-2}
	$\pm3.3\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm9.6\times10^{-4}$	$\pm9.8\times10^{-4}$	$\pm1.2\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm3.8\times10^{-3}$	$\pm4.5\times10^{-4}$	$\pm1.9\times10^{-3}$
0.97	0.221	9.45×10^{-2}	6.34×10^{-2}	5.79×10^{-2}	6.30×10^{-2}	8.80×10^{-2}	0.167	2.92×10^{-2}	9.59×10^{-2}
	$\pm4.1\times10^{-3}$	$\pm1.8\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm2.10\times10^{-3}$	$\pm3.6\times10^{-3}$	$\pm7.3\times10^{-3}$	$\pm1.0\times10^{-3}$	$\pm3.1\times10^{-3}$
0.98	0.309	0.125	7.89×10^{-2}	7.07×10^{-2}	7.78×10^{-2}	0.1135	0.228	3.52×10^{-2}	0.135
	$\pm6.4\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm2.2\times10^{-3}$	$\pm3.3\times10^{-3}$	$\pm6.26\times10^{-3}$	$\pm9.4\times10^{-4}$	$\pm3.8\times10^{-3}$
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.20: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/M/100$ queueing model with $\alpha = 0.1$ and $E[S] = 5$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/M/100$ model with $\alpha = 0.5$ and $E[S] = 5$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	4.40	1.24	0.449	0.302	0.417	1.02	2.96	0.148	16.92
	$\pm5.3\times10^{-2}$	$\pm2.53\times10^{-2}$	$\pm1.21\times10^{-2}$	$\pm6.4\times10^{-3}$	$\pm9.3\times10^{-3}$	$\pm2.1\times10^{-2}$	$\pm4.1\times10^{-2}$	$\pm6.8\times10^{-3}$	$\pm1.4\times10^{-1}$
0.93	6.01	1.63	0.548	0.351	0.520	1.37	4.09	0.177	28.0
	$\pm5.0\times10^{-2}$	$\pm2.9\times10^{-2}$	$\pm1.5\times10^{-2}$	$\pm8.8\times10^{-3}$	$\pm1.5\times10^{-2}$	$\pm3.4\times10^{-2}$	$\pm7.2\times10^{-2}$	$\pm6.0\times10^{-3}$	±0.27
0.95	7.29	1.96	0.645	0.410	0.620	1.66	4.98	0.202	38.06
	$\pm9.3\times10^{-2}$	$\pm3.7\times10^{-2}$	$\pm1.7\times10^{-2}$	$\pm1.8\times10^{-2}$	$\pm2.8\times10^{-2}$	$\pm4.5\times10^{-2}$	$\pm7.1\times10^{-2}$	$\pm7.4\times10^{-3}$	±0.32
0.97	8.48	2.21	0.688	0.431	0.702	1.97	5.96	0.216	49.8
	±0.12	$\pm5.5\times10^{-2}$	$\pm2.4\times10^{-2}$	$\pm1.4\times10^{-2}$	$\pm2.7\times10^{-2}$	$\pm5.7\times10^{-2}$	±0.11	$\pm6.6\times10^{-3}$	±0.43
0.98	9.21	2.40	0.741	0.454	0.737	2.09	6.39	0.226	56.3
	$\pm8.2\times10^{-2}$	$\pm3.5\times10^{-2}$	$\pm2.3\times10^{-2}$	$\pm2.3\times10^{-2}$	$\pm3.0\times10^{-2}$	$\pm4.4\times10^{-2}$	$\pm7.4\times10^{-2}$	$\pm6.9\times10^{-3}$	±0.40
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.21: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/M/100$ queueing model with $\alpha = 0.5$ and $E[S] = 5$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/H_2/100$ model with $\alpha = 0.1$ and $E[S] = 5$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	2.39×10^{-2}	1.77×10^{-2}	1.68×10^{-2}	1.72×10^{-2}	1.82×10^{-2}	2.08×10^{-2}	2.75×10^{-2}	1.10×10^{-2}	1.77×10^{-2}
	$\pm1.9\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm3.7\times10^{-3}$	$\pm1.0\times10^{-3}$	$\pm1.7\times10^{-3}$
0.93	7.12×10^{-2}	4.44×10^{-2}	3.87×10^{-2}	3.84×10^{-2}	4.05×10^{-2}	4.77×10^{-2}	6.83×10^{-2}	2.64×10^{-2}	4.31×10^{-2}
	$\pm5.2\times10^{-3}$	$\pm2.91\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm2.6\times10^{-3}$	$\pm3.0\times10^{-3}$	$\pm3.8\times10^{-3}$	$\pm6.0\times10^{-3}$	$\pm1.7\times10^{-3}$	3.3×10^{-3}
0.95	1.54×10^{-1}	8.70×10^{-2}	7.19×10^{-2}	7.05×10^{-2}	7.51×10^{-2}	9.18×10^{-2}	1.41×10^{-1}	5.06×10^{-2}	8.87×10^{-2}
	$\pm8.0\times10^{-3}$	$\pm4.0\times10^{-3}$	$\pm3.6\times10^{-3}$	$\pm3.9\times10^{-3}$	$\pm4.8\times10^{-3}$	$\pm7.0\times10^{-3}$	$\pm1.3\times10^{-2}$	$\pm3.3\times10^{-3}$	6.4×10^{-3}
0.97	2.93×10^{-1}	1.51×10^{-1}	1.17×10^{-1}	1.13×10^{-1}	1.20×10^{-1}	1.51×10^{-1}	2.45×10^{-1}	8.13×10^{-2}	1.62×10^{-1}
	$\pm1.4\times10^{-2}$	$\pm6.1\times10^{-3}$	$\pm4.9\times10^{-3}$	$\pm5.5\times10^{-3}$	$\pm7.0\times10^{-3}$	$\pm1.1\times10^{-2}$	2.06×10^{-2}	$\pm4.2\times10^{-3}$	1.2×10^{-2}
0.98	4.36×10^{-1}	2.11×10^{-1}	1.58×10^{-1}	1.51×10^{-1}	1.64×10^{-1}	2.15×10^{-1}	3.70×10^{-1}	1.13×10^{-1}	2.56×10^{-1}
	$\pm1.7\times10^{-2}$	$\pm8.0\times10^{-3}$	$\pm5.63\times10^{-3}$	$\pm5.4\times10^{-3}$	$\pm6.1\times10^{-3}$	$\pm8.7\times10^{-3}$	1.53×10^{-2}	4.24×10^{-3}	1.21×10^{-2}
$n(x)$	1537	6147	38416		38416	6147	1537		
Interval	(76 min.)	(307 min.)	(1920 min.)		(1920 min.)	(307 min.)	(76 min.)		

Table A.22: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/H_2/100$ queueing model with $\alpha = 0.1$ and a mean service time of 5 minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/H_2/100$ model with $\alpha = 0.5$ and $E[S] = 5$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	4.83	1.62	0.835	0.696	0.825	1.55	3.46	0.550	17.9
	$\pm5.5\times10^{-2}$	$\pm3.4\times10^{-2}$	$\pm2.5\times10^{-2}$	$\pm2.4\times10^{-2}$	$\pm3.1\times10^{-2}$	$\pm5.3\times10^{-2}$	±0.11	$\pm1.7\times10^{-2}$	±0.42
0.93	6.49	2.11	1.04	0.856	1.04	1.92	4.70	0.666	29.0
	$\pm7.3\times10^{-2}$	$\pm4.5\times10^{-2}$	$\pm3.6\times10^{-2}$	$\pm3.6\times10^{-2}$	$\pm4.3\times10^{-2}$	$\pm6.1\times10^{-2}$	±0.10	$\pm2.5\times10^{-2}$	±0.44
0.95	7.77	2.45	1.15	0.919	1.14	2.19	5.53	0.728	38.6
	±0.21	$\pm9.8\times10^{-2}$	$\pm4.5\times10^{-2}$	$\pm2.7\times10^{-2}$	$\pm4.7\times10^{-2}$	$\pm9.9\times10^{-2}$	±0.20	$\pm2.7\times10^{-2}$	±0.81
0.97	9.17	2.85	1.29	1.01	1.26	2.50	6.45	0.789	50.3
	0.19	0.11	$\pm6.3\times10^{-2}$	$\pm4.7\times10^{-2}$	$\pm6.1\times10^{-2}$	±0.12	±0.23	$\pm3.6\times10^{-2}$	0.95
0.98	9.71	2.91	1.29	1.03	1.35	2.77	7.17	0.8196	57.9
	±0.16	±0.66	±0.30	±0.24	±0.31	±0.63	±1.6	±0.19	±13
$n(x)$	1537	6147	38416		38416	6147	1537		
Interval	(76 min.)	(307 min.)	(1920 min.)		(1920 min.)	(307 min.)	(76 min.)		

Table A.23: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/H_2/100$ queueing model with $\alpha = 0.5$ and mean service time of 5 minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/D/100$ model with $\alpha = 0.1$ and $E[S] = 5$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	4.34×10^{-3}	3.29×10^{-3}	2.98×10^{-3}	2.88×10^{-3}	2.85×10^{-3}	2.93×10^{-3}	3.34×10^{-3}	1.01×10^{-3}	3.02×10^{-3}
	$\pm2.3\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.2\times10^{-4}$	$\pm1.45\times10^{-4}$	$\pm2.26\times10^{-4}$	2.97×10^{-5}	$\pm1.1\times10^{-4}$
0.93	2.32×10^{-2}	1.23×10^{-2}	9.50×10^{-3}	8.88×10^{-3}	9.14×10^{-3}	1.10×10^{-2}	1.71×10^{-2}	1.20×10^{-3}	9.79×10^{-3}
	$\pm1.4\times10^{-3}$	$\pm4.8\times10^{-4}$	$\pm3.1\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm4.4\times10^{-4}$	0.000750349	0.001581422	3.74×10^{-5}	5.0×10^{-4}
0.95	7.02×10^{-2}	2.94×10^{-2}	1.90×10^{-2}	1.68×10^{-2}	1.80×10^{-2}	2.51×10^{-2}	4.89×10^{-2}	1.30×10^{-3}	2.32×10^{-2}
	$\pm8.4\times10^{-4}$	$\pm3.8\times10^{-4}$	$\pm2.7\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm5.9\times10^{-4}$	$\pm1.3\times10^{-3}$	$\pm3.7\times10^{-5}$	$\pm4.4\times10^{-4}$
0.97	0.178	6.27×10^{-2}	3.41×10^{-2}	2.86×10^{-2}	3.27×10^{-2}	5.43×10^{-2}	0.124	1.27×10^{-3}	6.02×10^{-2}
	$\pm6.0\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm8.7\times10^{-4}$	$\pm8.1\times10^{-4}$	$\pm1.3\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm5.7\times10^{-3}$	5.6×10^{-5}	$\pm3.2\times10^{-3}$
0.98	0.250	8.53×10^{-2}	4.32×10^{-2}	3.45×10^{-2}	3.92×10^{-2}	6.84×10^{-2}	0.170	1.28×10^{-3}	8.66×10^{-2}
	$\pm3.4\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm4.8\times10^{-4}$	$\pm4.3\times10^{-4}$	$\pm5.7\times10^{-4}$	$\pm9.1\times10^{-4}$	$\pm1.9\times10^{-3}$	$\pm6.6\times10^{-5}$	$\pm1.6\times10^{-3}$
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.24: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/D/100$ queueing model with $\alpha = 0.1$ and $E[S] = 5$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/D/100$ model with $\alpha = 0.5$ and $E[S] = 5$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	4.23	1.08	0.298	0.149	0.260	0.852	2.77	1.37×10^{-3}	16.5
	$\pm4.1\times10^{-2}$	$\pm1.6\times10^{-2}$	$\pm5.2\times10^{-3}$	$\pm3.3\times10^{-3}$	$\pm8.2\times10^{-3}$	$\pm1.6\times10^{-2}$	$\pm2.6\times10^{-2}$	$\pm4.8\times10^{-5}$	$\pm8.6\times10^{-2}$
0.93	5.79	1.44	0.370	0.184	0.363	1.23	3.97	1.32×10^{-3}	27.9
	$\pm9.2\times10^{-2}$	$\pm4.2\times10^{-2}$	$\pm1.3\times10^{-2}$	$\pm6.9\times10^{-3}$	$\pm2.6\times10^{-2}$	$\pm5.6\times10^{-2}$	±0.11	$\pm4.6\times10^{-5}$	±0.40
0.95	7.03	1.75	0.437	0.198	0.3991	1.42	4.69	1.39×10^{-3}	37.2
	$\pm2.1\times10^{-2}$	$\pm9.3\times10^{-3}$	$\pm7.5\times10^{-3}$	$\pm7.9\times10^{-3}$	$\pm8.6\times10^{-3}$	$\pm1.0\times10^{-2}$	$\pm1.4\times10^{-2}$	$\pm6.0\times10^{-5}$	±0.12
0.97	8.40	2.09	0.514	0.215	0.4386	1.62	5.47	1.39×10^{-3}	48.3
	$\pm7.0\times10^{-2}$	$\pm3.2\times10^{-2}$	$\pm1.3\times10^{-2}$	$\pm3.8\times10^{-3}$	$\pm9.2\times10^{-3}$	$\pm2.3\times10^{-2}$	$\pm4.6\times10^{-2}$	$\pm8.6\times10^{-5}$	±0.22
0.98	9.02	2.22	0.531	0.217	0.468	1.76	5.94	1.37×10^{-3}	54.8
	$\pm7.2\times10^{-2}$	$\pm3.4\times10^{-2}$	$\pm1.4\times10^{-2}$	$\pm5.5\times10^{-3}$	$\pm1.2\times10^{-2}$	$\pm2.9\times10^{-2}$	$\pm5.5\times10^{-2}$	$\pm7.9\times10^{-5}$	±0.16
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.25: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/D/100$ queueing model with $\alpha = 0.5$ and $E[S] = 5$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/M/100$ model with $\alpha = 0.1$ and $E[S] = 30$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	5.48×10^{-3}	4.57×10^{-3}	4.34×10^{-3}	4.29×10^{-3}	4.32×10^{-3}	4.49×10^{-3}	5.04×10^{-3}	2.26×10^{-3}	4.61×10^{-3}
	$\pm1.1\times10^{-4}$	$\pm9.0\times10^{-5}$	$\pm8.7\times10^{-5}$	$\pm8.8\times10^{-5}$	$\pm9.1\times10^{-5}$	$\pm9.9\times10^{-5}$	$\pm1.2\times10^{-4}$	$\pm5.1\times10^{-5}$	$\pm9.8\times10^{-5}$
0.93	1.02×10^{-2}	7.95×10^{-3}	7.39×10^{-3}	7.29×10^{-3}	7.39×10^{-3}	7.84×10^{-3}	9.28×10^{-3}	3.77×10^{-3}	8.04×10^{-3}
	$\pm2.9\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm3.6\times10^{-4}$	$\pm1.0\times10^{-4}$	$\pm2.6\times10^{-4}$
0.95	1.54×10^{-2}	1.13×10^{-2}	1.02×10^{-2}	1.01×10^{-2}	1.02×10^{-2}	1.10×10^{-2}	1.37×10^{-2}	5.08×10^{-3}	1.17×10^{-2}
	$\pm1.8\times10^{-4}$	$\pm1.4\times10^{-4}$	$\pm1.4\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm1.9\times10^{-4}$	$\pm2.7\times10^{-4}$	$\pm7.2\times10^{-5}$	$\pm2.0\times10^{-4}$
0.97	2.42×10^{-2}	1.64×10^{-2}	1.44×10^{-2}	1.41×10^{-2}	1.44×10^{-2}	1.60×10^{-2}	2.10×10^{-2}	7.16×10^{-3}	1.75×10^{-2}
	$\pm2.9\times10^{-4}$	$\pm2.0\times10^{-4}$	$\pm1.9\times10^{-4}$	$\pm2.0\times10^{-4}$	$\pm2.2\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm3.9\times10^{-4}$	$\pm9.8\times10^{-5}$	$\pm2.4\times10^{-4}$
0.98	3.41×10^{-2}	2.16×10^{-2}	1.85×10^{-2}	1.80×10^{-2}	1.85×10^{-2}	2.10×10^{-2}	2.88×10^{-2}	9.14×10^{-3}	2.39×10^{-2}
	$\pm1.3\times10^{-3}$	$\pm6.9\times10^{-4}$	$\pm5.7\times10^{-4}$	$\pm5.9\times10^{-4}$	$\pm6.6\times10^{-4}$	$\pm8.8\times10^{-4}$	$\pm1.5\times10^{-3}$	$\pm3.0\times10^{-4}$	$\pm1.0\times10^{-3}$
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.26: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/M/100$ queueing model with $\alpha = 0.1$ and $E[S] = 30$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/H_2/100$ model with $\alpha = 0.1$ and $E[S] = 30$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	9.08×10^{-3}	7.89×10^{-3}	7.85×10^{-3}	8.06×10^{-3}	8.42×10^{-3}	9.22×10^{-3}	1.11×10^{-2}	4.92×10^{-3}	8.65×10^{-3}
	$\pm3.3\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm3.6\times10^{-4}$	$\pm4.0\times10^{-4}$	$\pm4.4\times10^{-4}$	$\pm5.1\times10^{-4}$	$\pm6.7\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm4.5\times10^{-4}$
0.93	1.53×10^{-2}	1.25×10^{-2}	1.22×10^{-2}	1.24×10^{-2}	1.30×10^{-2}	1.45×10^{-2}	1.81×10^{-2}	7.73×10^{-3}	1.41×10^{-2}
	$\pm3.5\times10^{-4}$	$\pm3.0\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm3.6\times10^{-4}$	$\pm3.9\times10^{-4}$	$\pm4.7\times10^{-4}$	$\pm6.1\times10^{-4}$	$\pm2.5\times10^{-4}$	$\pm4.1\times10^{-4}$
0.95	2.52×10^{-2}	1.92×10^{-2}	1.83×10^{-2}	1.87×10^{-2}	1.97×10^{-2}	2.22×10^{-2}	2.87×10^{-2}	1.20×10^{-2}	2.23×10^{-2}
	$\pm9.1\times10^{-4}$	$\pm6.7\times10^{-4}$	$\pm6.8\times10^{-4}$	$\pm7.4\times10^{-4}$	$\pm8.4\times10^{-4}$	$\pm1.0\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm6.5\times10^{-4}$	$\pm9.7\times10^{-4}$
0.97	4.37×10^{-2}	3.02×10^{-2}	2.75×10^{-2}	2.77×10^{-2}	2.91×10^{-2}	3.33×10^{-2}	4.50×10^{-2}	1.79×10^{-2}	3.53×10^{-2}
	$\pm2.9\times10^{-3}$	$\pm1.8\times10^{-3}$	$\pm1.60\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm2.4\times10^{-3}$	$\pm3.62\times10^{-3}$	1.25×10^{-3}	2.28×10^{-3}
0.98	8.33×10^{-2}	5.15×10^{-2}	4.48×10^{-2}	4.46×10^{-2}	4.74×10^{-2}	5.63×10^{-2}	8.13×10^{-2}	3.03×10^{-2}	6.24×10^{-2}
	$\pm7.8\times10^{-3}$	$\pm3.8\times10^{-3}$	$\pm3.1\times10^{-3}$	$\pm3.2\times10^{-3}$	$\pm3.6\times10^{-3}$	$\pm5.0\times10^{-3}$	$\pm8.7\times10^{-3}$	$\pm2.3\times10^{-3}$	$\pm5.8\times10^{-3}$
$n(x)$	1537	6147	38416		38416	6147	1537		
Interval	(76 min.)	(307 min.)	(1920 min.)		(1920 min.)	(307 min.)	(76 min.)		

Table A.27: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/H_2/100$ queueing model with $\alpha = 0.1$ and $E[S] = 30$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/D/100$ model with $\alpha = 0.1$ and $E[S] = 30$ min

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	3.11×10^{-3}	2.56×10^{-3}	2.38×10^{-3}	2.31×10^{-3}	2.28×10^{-3}	2.29×10^{-3}	2.44×10^{-3}	9.72×10^{-4}	2.47×10^{-3}
	$\pm5.1\times10^{-5}$	$\pm3.8\times10^{-5}$	$\pm3.4\times10^{-5}$	$\pm3.4\times10^{-5}$	$\pm3.3\times10^{-5}$	$\pm3.5\times10^{-5}$	$\pm4.1\times10^{-5}$	$\pm2.5\times10^{-5}$	$\pm3.6\times10^{-5}$
0.93	5.94×10^{-3}	4.42×10^{-3}	3.97×10^{-3}	3.84×10^{-3}	3.81×10^{-3}	4.02×10^{-3}	4.60×10^{-3}	1.23×10^{-3}	4.18×10^{-3}
	$\pm1.1\times10^{-4}$	$\pm7.2\times10^{-5}$	$\pm6.3\times10^{-5}$	$\pm6.3\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm7.6\times10^{-5}$	$\pm1.1\times10^{-4}$	$\pm2.4\times10^{-5}$	$\pm7.8\times10^{-5}$
0.95	9.21×10^{-3}	6.23×10^{-3}	5.41×10^{-3}	5.19×10^{-3}	5.20×10^{-3}	5.60×10^{-3}	7.05×10^{-3}	1.31×10^{-3}	6.01×10^{-3}
	$\pm4.6\times10^{-5}$	$\pm3.9\times10^{-5}$	$\pm4.0\times10^{-5}$	$\pm4.1\times10^{-5}$	$\pm4.3\times10^{-5}$	$\pm4.6\times10^{-5}$	$\pm5.4\times10^{-5}$	$\pm2.7\times10^{-5}$	$\pm4.1\times10^{-5}$
0.97	1.51×10^{-2}	9.21×10^{-3}	7.64×10^{-3}	7.26×10^{-3}	7.34×10^{-3}	8.21×10^{-3}	1.13×10^{-2}	1.35×10^{-3}	9.29×10^{-3}
	$\pm9.7\times10^{-5}$	$\pm7.5\times10^{-5}$	$\pm6.8\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm6.9\times10^{-5}$	$\pm9.1\times10^{-5}$	$\pm2.6\times10^{-5}$	$\pm3.8\times10^{-5}$
0.98	1.90×10^{-2}	1.10×10^{-2}	8.84×10^{-3}	8.29×10^{-3}	8.37×10^{-3}	9.45×10^{-3}	1.36×10^{-2}	1.34×10^{-3}	1.13×10^{-2}
	$\pm1.0\times10^{-4}$	$\pm6.6\times10^{-5}$	$\pm5.8\times10^{-5}$	$\pm5.7\times10^{-5}$	$\pm6.3\times10^{-5}$	$\pm8.3\times10^{-5}$	$\pm1.4\times10^{-4}$	$\pm4.2\times10^{-5}$	$\pm6.9\times10^{-5}$
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.28: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/D/100$ queueing model with $\alpha = 0.1$ and $E[S] = 30$ minutes. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/M/100$ model with $\alpha = 0.5$ and $E[S] = 6$ hours

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	5.18×10^{-3}	4.48×10^{-3}	4.30×10^{-3}	4.26×10^{-3}	4.29×10^{-3}	4.41×10^{-3}	4.84×10^{-3}	2.24×10^{-3}	9.00×10^{-3}
	$\pm4.7\times10^{-5}$	$\pm4.8\times10^{-5}$	$\pm5.2\times10^{-5}$	$\pm5.6\times10^{-5}$	$\pm6.0\times10^{-5}$	$\pm6.7\times10^{-5}$	$\pm8.0\times10^{-5}$	$\pm2.4\times10^{-5}$	$\pm1.6\times10^{-4}$
0.93	6.94×10^{-3}	5.84×10^{-3}	5.57×10^{-3}	5.53×10^{-3}	5.57×10^{-3}	5.80×10^{-3}	6.51×10^{-3}	2.85×10^{-3}	1.43×10^{-2}
	$\pm7.6\times10^{-5}$	$\pm7.0\times10^{-5}$	$\pm7.2\times10^{-5}$	$\pm7.5\times10^{-5}$	$\pm7.9\times10^{-5}$	$\pm8.7\times10^{-5}$	$\pm1.0\times10^{-4}$	$\pm3.7\times10^{-5}$	$\pm2.4\times10^{-4}$
0.95	9.07×10^{-3}	7.35×10^{-3}	6.94×10^{-3}	6.87×10^{-3}	6.96×10^{-3}	7.32×10^{-3}	8.46×10^{-3}	3.52×10^{-3}	2.19×10^{-2}
	$\pm1.6\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm1.4\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm1.8\times10^{-4}$	$\pm2.5\times10^{-4}$	$\pm4.8\times10^{-5}$	$\pm6.5\times10^{-4}$
0.97	1.39×10^{-2}	1.04×10^{-2}	9.58×10^{-3}	9.44×10^{-3}	9.59×10^{-3}	1.03×10^{-2}	1.25×10^{-2}	4.80×10^{-3}	3.86×10^{-2}
	$\pm4.9\times10^{-4}$	$\pm2.7\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm3.0\times10^{-4}$	$\pm3.9\times10^{-4}$	$\pm6.4\times10^{-4}$	$\pm1.2\times10^{-4}$	$\pm1.4\times10^{-3}$
0.98	2.04×10^{-2}	1.42×10^{-2}	1.27×10^{-2}	1.24×10^{-2}	1.26×10^{-2}	1.38×10^{-2}	1.76×10^{-2}	6.34×10^{-3}	5.97×10^{-2}
	$\pm8.5\times10^{-4}$	$\pm4.7\times10^{-4}$	$\pm4.0\times10^{-4}$	$\pm4.1\times10^{-4}$	$\pm4.5\times10^{-4}$	$\pm5.8\times10^{-4}$	$\pm9.4\times10^{-4}$	$\pm2.2\times10^{-4}$	$\pm2.8\times10^{-4}$
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.29: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/M/100$ queueing model with $\alpha = 0.5$ and $E[S] = 6$ hours. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/H_2/100$ model with $\alpha = 0.5$ and $E[S] = 6$ hours

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	6.21×10^{-3}	5.50×10^{-3}	5.40×10^{-3}	5.44×10^{-3}	5.56×10^{-3}	5.87×10^{-3}	6.64×10^{-3}	3.19×10^{-3}	1.30×10^{-2}
	$\pm1.2\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm1.8\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.8\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm6.0\times10^{-4}$
0.93	9.72×10^{-3}	8.31×10^{-3}	8.13×10^{-3}	8.24×10^{-3}	8.51×10^{-3}	9.17×10^{-3}	1.08×10^{-2}	4.93×10^{-3}	2.49×10^{-2}
	$\pm2.5\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm2.8\times10^{-4}$	$\pm3.1\times10^{-4}$	$\pm3.6\times10^{-4}$	$\pm4.7\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm9.1\times10^{-4}$
0.95	1.58×10^{-2}	1.27×10^{-2}	1.23×10^{-2}	1.26×10^{-2}	1.31×10^{-2}	1.46×10^{-2}	1.82×10^{-2}	7.67×10^{-3}	4.35×10^{-2}
	$\pm7.6\times10^{-4}$	$\pm5.9\times10^{-4}$	$\pm6.1\times10^{-4}$	$\pm6.6\times10^{-4}$	$\pm7.4\times10^{-4}$	$\pm9.1\times10^{-4}$	$\pm1.3\times10^{-3}$	$\pm4.9\times10^{-4}$	$\pm1.3\times10^{-3}$
0.97	3.15×10^{-2}	2.26×10^{-2}	2.09×10^{-2}	2.10×10^{-2}	2.20×10^{-2}	2.50×10^{-2}	3.27×10^{-2}	1.36×10^{-2}	7.65×10^{-2}
	$\pm2.2\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.4\times10^{-4}$	$\pm1.4\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm2.8\times10^{-3}$	$\pm1.03\times10^{-3}$	$\pm3.2\times10^{-3}$
0.98	6.63×10^{-2}	4.18×10^{-2}	3.68×10^{-2}	3.67×10^{-2}	3.90×10^{-2}	4.61×10^{-2}	6.61×10^{-2}	2.47×10^{-2}	1.13×10^{-1}
	$\pm9.1\times10^{-3}$	$\pm4.2\times10^{-3}$	$\pm3.4\times10^{-3}$	$\pm3.6\times10^{-3}$	$\pm4.2\times10^{-3}$	$\pm6.1\times10^{-3}$	$\pm1.1\times10^{-2}$	$\pm2.8\times10^{-3}$	$\pm4.6\times10^{-3}$
$n(x)$	1537	6147	38416		38416	6147	1537		
Interval	(76 min.)	(307 min.)	(1920 min.)		(1920 min.)	(307 min.)	(76 min.)		

Table A.30: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/H_2/100$ queueing model with $\alpha = 0.5$ and $E[S] = 6$ hours. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

Efficiency in the $M(t)/D/100$ model with $\alpha = 0.5$ and $E[S] = 6$ hours

	$HOL_m(x)$								
ρ	x = 0.1	x = 0.05	x = 0.02	HOL_m	x = -0.02	x = -0.05	x = -0.1	QL	HOL
0.9	7.29×10^{-3}	6.06×10^{-3}	5.65×10^{-3}	5.49×10^{-3}	5.42×10^{-3}	5.42×10^{-3}	5.72×10^{-3}	3.08×10^{-3}	1.06×10^{-2}
	$\pm1.4\times10^{-5}$	$\pm1.3\times10^{-5}$	$\pm1.5\times10^{-5}$	$\pm1.6\times10^{-5}$	$\pm1.7\times10^{-5}$	$\pm1.9\times10^{-5}$	$\pm2.3\times10^{-5}$	$\pm1.3\times10^{-5}$	$\pm4.8\times10^{-5}$
0.93	8.50×10^{-3}	6.93×10^{-3}	6.43×10^{-3}	6.25×10^{-3}	6.17×10^{-3}	6.22×10^{-3}	6.69×10^{-3}	3.377×10^{-3}	1.43×10^{-2}
	$\pm6.0\times10^{-5}$	$\pm5.3\times10^{-5}$	$\pm5.1\times10^{-5}$	$\pm5.0\times10^{-5}$	$\pm4.9\times10^{-5}$	$\pm4.8\times10^{-5}$	$\pm4.9\times10^{-5}$	$\pm2.7\times10^{-5}$	$\pm6.6\times10^{-5}$
0.95	9.40×10^{-3}	7.52×10^{-3}	6.93×10^{-3}	6.72×10^{-3}	6.64×10^{-3}	6.73×10^{-3}	7.34×10^{-3}	3.52×10^{-3}	1.76×10^{-2}
	$\pm4.7\times10^{-5}$	$\pm3.8\times10^{-5}$	$\pm3.5\times10^{-5}$	$\pm3.4\times10^{-5}$	$\pm3.4\times10^{-5}$	$\pm3.5\times10^{-5}$	$\pm4.1\times10^{-5}$	$\pm2.3\times10^{-5}$	$\pm1.1\times10^{-4}$
0.97	1.09×10^{-2}	8.34×10^{-3}	7.57×10^{-3}	7.32×10^{-3}	7.25×10^{-3}	7.45×10^{-3}	8.43×10^{-3}	3.49×10^{-3}	2.49×10^{-2}
	$\pm4.4\times10^{-5}$	$\pm3.0\times10^{-5}$	$\pm2.7\times10^{-5}$	$\pm2.6\times10^{-5}$	$\pm2.8\times10^{-5}$	$\pm3.3\times10^{-5}$	$\pm5.0\times10^{-5}$	$\pm3.6\times10^{-5}$	$\pm2.9\times10^{-4}$
0.98	1.27×10^{-2}	9.20×10^{-3}	8.19×10^{-3}	7.88×10^{-3}	7.83×10^{-3}	8.18×10^{-3}	9.67×10^{-3}	3.36×10^{-3}	3.40×10^{-2}
	$\pm1.4\times10^{-4}$	$\pm8.3\times10^{-5}$	$\pm7.4\times10^{-5}$	$\pm7.7\times10^{-5}$	$\pm8.6\times10^{-5}$	$\pm1.1\times10^{-4}$	$\pm1.8\times10^{-4}$	$\pm4.9\times10^{-5}$	$\pm8.9\times10^{-4}$
$n(x)$	385	1537	9604		9604	1537	385		
Interval	(20 min.)	(77 min.)	(480 min.)		(480 min.)	(77 min.)	(20 min.)		

Table A.31: Performance of $HOL_m(x)$ delay predictors, as a function of the traffic intensity, ρ, and alternative x, in the $M(t)/D/100$ queueing model with $\alpha = 0.5$ and $E[S] = 6$ hours. Sample sizes needed and length of estimation intervals required are also included. The ASE's are measured in units of mean service time squared per customer.

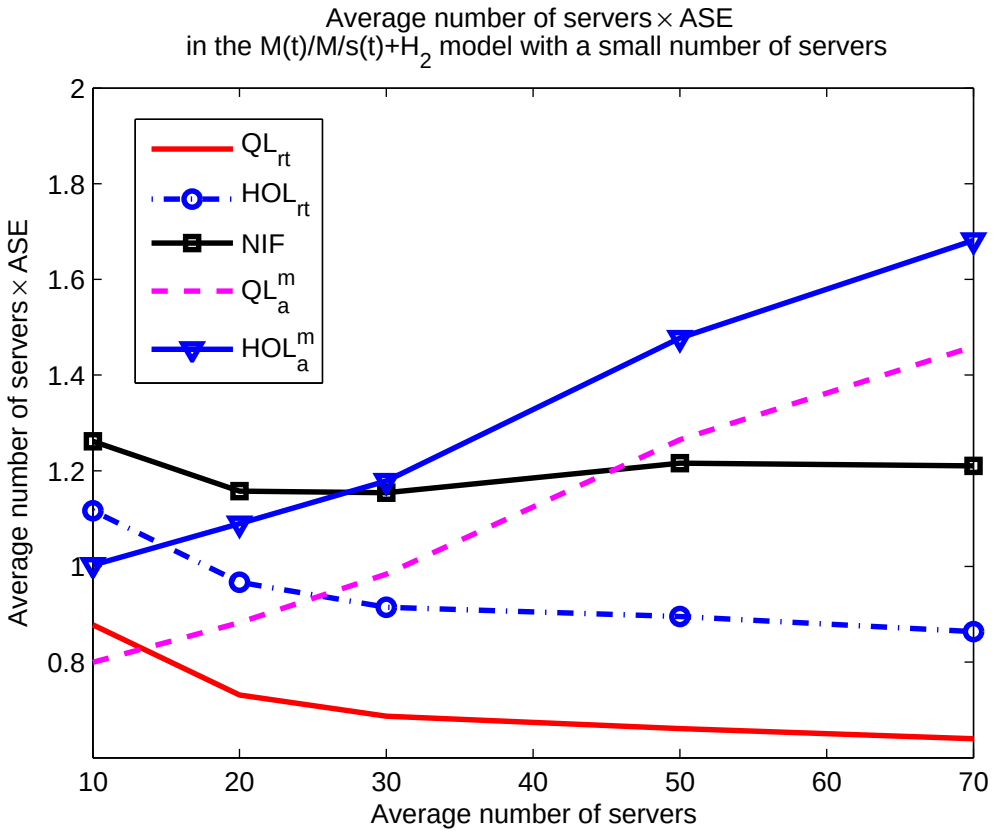

Figure A.1: $\bar{s} \times$ ASE of the alternative predictors in the $M(t)/M/s(t) + H_2$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a small average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S] = 6$ hours with a 24 hour arrival-rate cycle.

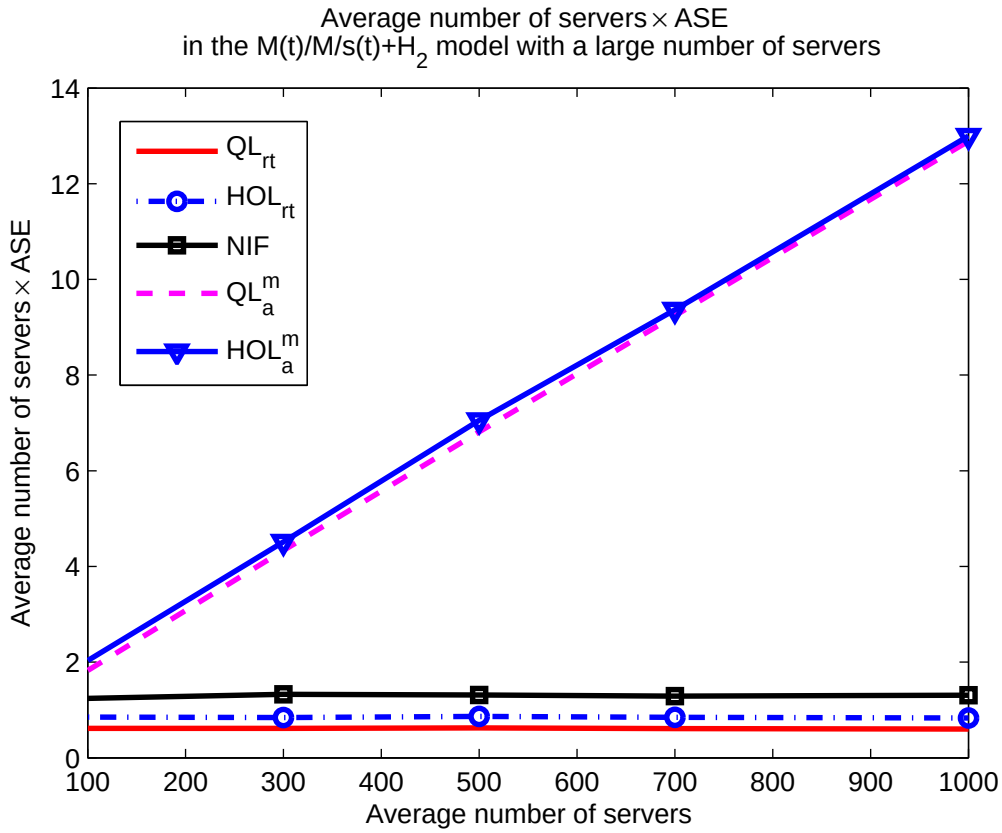

Figure A.2: $\bar{s} \times$ ASE of the alternative predictors in the $M(t)/M/s(t) + H_2$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a large average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S] = 6$ hours with a 24 hour arrival-rate cycle.

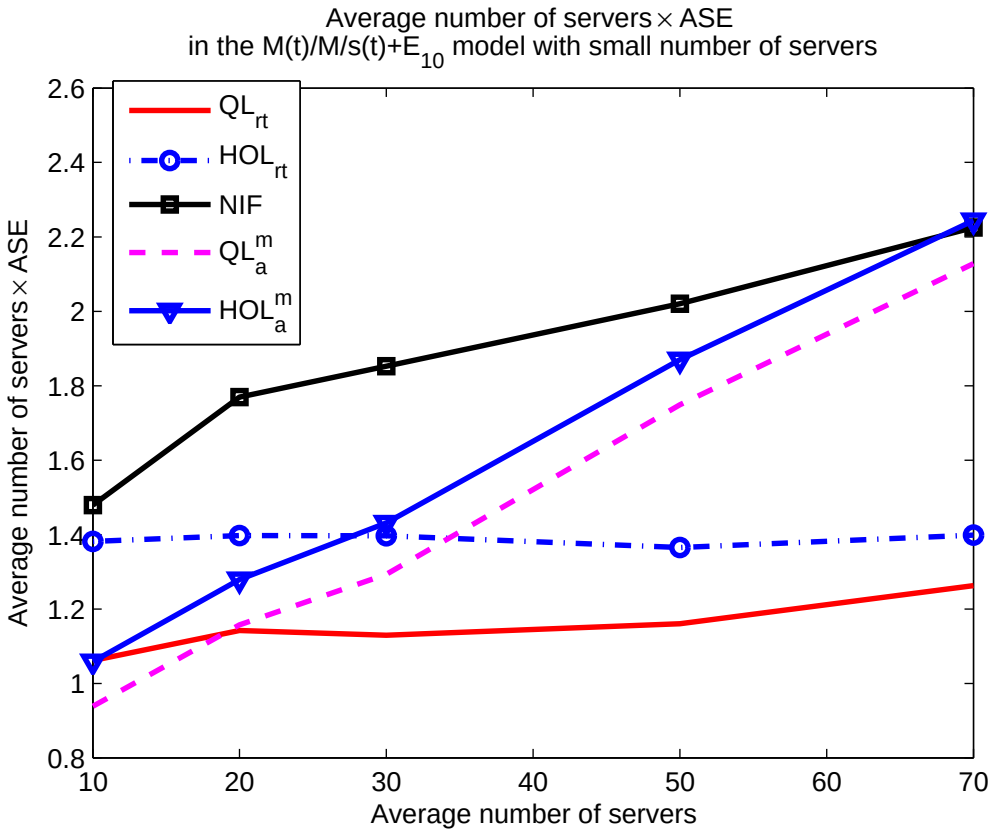

Figure A.3: $\bar{s} \times$ ASE of the alternative predictors in the $M(t)/M/s(t) + E_{10}$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a small average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S] = 6$ hours with a 24 hour arrival-rate cycle.

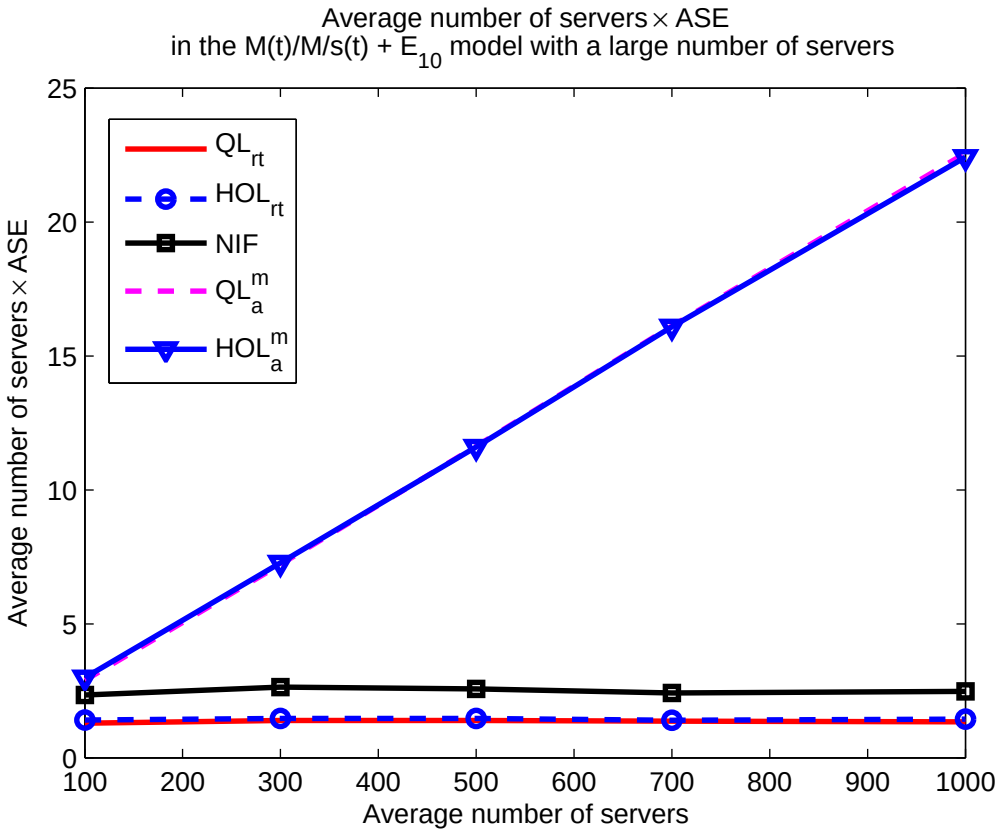

Figure A.4: $\bar{s} \times$ ASE of the alternative predictors in the $M(t)/M/s(t) + E_{10}$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a large average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S] = 6$ hours with a 24 hour arrival-rate cycle.

ASE of the predictors in the $M(t)/M/s(t)+H_2$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	8.78×10^{-2}	1.12×10^{-1}	1.26×10^{-1}	8.00×10^{-2}	1.00×10^{-1}	1.14×10^{-1}	1.34×10^{-1}
	$\pm3.2\times10^{-3}$	$\pm3.2\times10^{-3}$	$\pm5.1\times10^{-3}$	$\pm4.9\times10^{-3}$	$\pm4.3\times10^{-3}$	$\pm3.9\times10^{-3}$	$\pm5.2\times10^{-3}$
20	3.66×10^{-2}	4.83×10^{-2}	5.79×10^{-2}	4.41×10^{-2}	5.45×10^{-2}	5.89×10^{-2}	6.99×10^{-2}
	$\pm1.2\times10^{-3}$	$\pm2.0\times10^{-3}$	$\pm3.1\times10^{-3}$	$\pm2.1\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm2.9\times10^{-3}$
30	2.29×10^{-2}	3.05×10^{-2}	3.85×10^{-2}	3.28×10^{-2}	3.93×10^{-2}	4.27×10^{-2}	4.95×10^{-2}
	$\pm9.3\times10^{-4}$	$\pm1.2\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm1.7\times10^{-3}$
50	1.32×10^{-2}	1.79×10^{-2}	2.43×10^{-2}	2.53×10^{-2}	2.95×10^{-2}	3.24×10^{-2}	3.68×10^{-2}
	$\pm5.3\times10^{-4}$	$\pm4.6\times10^{-4}$	$\pm1.3\times10^{-3}$	$\pm1.0\times10^{-3}$	$\pm1.0\times10^{-3}$	$\pm8.9\times10^{-4}$	$\pm1.1\times10^{-3}$
70	9.14×10^{-3}	1.23×10^{-2}	1.73×10^{-2}	2.09×10^{-2}	2.40×10^{-2}	2.69×10^{-2}	3.02×10^{-2}
	$\pm3.3\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm7.2\times10^{-4}$	$\pm7.4\times10^{-4}$	$\pm6.3\times10^{-4}$	$\pm6.4\times10^{-4}$	$\pm7.2\times10^{-4}$
100	6.15×10^{-3}	8.49×10^{-3}	1.24×10^{-2}	1.83×10^{-2}	2.03×10^{-2}	2.34×10^{-2}	2.54×10^{-2}
	$\pm2.0\times10^{-4}$	$\pm4.0\times10^{-4}$	$\pm6.2\times10^{-4}$	$\pm7.0\times10^{-4}$	$\pm8.1\times10^{-4}$	$\pm6.6\times10^{-4}$	$\pm8.2\times10^{-4}$
300	2.05×10^{-3}	2.80×10^{-3}	4.42×10^{-3}	1.44×10^{-2}	1.51×10^{-2}	1.84×10^{-2}	1.90×10^{-2}
	$\pm5.4\times10^{-5}$	$\pm5.4\times10^{-5}$	$\pm1.9\times10^{-4}$	$\pm2.9\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm3.1\times10^{-4}$
500	1.25×10^{-3}	1.73×10^{-3}	2.63×10^{-3}	1.36×10^{-2}	1.41×10^{-2}	1.74×10^{-2}	1.78×10^{-2}
	$\pm3.2\times10^{-5}$	$\pm4.7\times10^{-5}$	$\pm1.1\times10^{-4}$	$\pm2.0\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm1.8\times10^{-4}$	$\pm2.6\times10^{-4}$
700	8.70×10^{-4}	1.21×10^{-3}	1.84×10^{-3}	1.32×10^{-2}	1.34×10^{-2}	1.68×10^{-2}	1.70×10^{-2}
	$\pm4.0\times10^{-5}$	$\pm4.9\times10^{-5}$	$\pm9.0\times10^{-5}$	$\pm2.3\times10^{-4}$	$\pm2.5\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm2.5\times10^{-4}$
1000	6.02×10^{-4}	8.31×10^{-4}	1.31×10^{-3}	1.29×10^{-2}	1.30×10^{-2}	1.64×10^{-2}	1.65×10^{-2}
	$\pm2.1\times10^{-5}$	$\pm1.5\times10^{-5}$	$\pm5.3\times10^{-5}$	$\pm1.7\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm2.0\times10^{-4}$

Table A.32: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/M/s(t)+H_2$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/M/s(t)+E_{10}$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	1.06×10^{-1}	1.38×10^{-1}	1.48×10^{-1}	9.38×10^{-2}	1.06×10^{-1}	1.19×10^{-1}	1.28×10^{-1}
	$\pm5.7\times10^{-3}$	$\pm6.0\times10^{-3}$	$\pm6.3\times10^{-3}$	$\pm3.1\times10^{-3}$	$\pm3.2\times10^{-3}$	$\pm3.9\times10^{-3}$	$\pm4.3\times10^{-3}$
20	5.71×10^{-2}	6.99×10^{-2}	8.85×10^{-2}	5.79×10^{-2}	6.40×10^{-2}	6.81×10^{-2}	7.34×10^{-2}
	$\pm3.4\times10^{-3}$	$\pm3.9\times10^{-3}$	$\pm4.7\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm2.9\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm2.7\times10^{-3}$
30	3.76×10^{-2}	4.65×10^{-2}	6.17×10^{-2}	4.31×10^{-2}	4.77×10^{-2}	4.95×10^{-2}	5.33×10^{-2}
	$\pm1.5\times10^{-3}$	$\pm2.3\times10^{-3}$	$\pm2.0\times10^{-3}$	$\pm1.8\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm1.7\times10^{-3}$
50	2.32×10^{-2}	2.73×10^{-2}	4.04×10^{-2}	3.50×10^{-2}	3.74×10^{-2}	3.93×10^{-2}	4.14×10^{-2}
	$\pm1.6\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm2.6\times10^{-3}$	$\pm8.9\times10^{-4}$	$\pm9.6\times10^{-4}$	$\pm8.6\times10^{-4}$	$\pm9.6\times10^{-4}$
70	1.80×10^{-2}	2.00×10^{-2}	3.18×10^{-2}	3.04×10^{-2}	3.21×10^{-2}	3.39×10^{-2}	3.51×10^{-2}
	$\pm7.6\times10^{-4}$	$\pm8.0\times10^{-4}$	$\pm1.1\times10^{-3}$	$\pm8.8\times10^{-4}$	$\pm9.1\times10^{-4}$	$\pm7.4\times10^{-4}$	$\pm8.4\times10^{-4}$
100	1.29×10^{-2}	1.41×10^{-2}	2.35×10^{-2}	2.89×10^{-2}	3.00×10^{-2}	3.14×10^{-2}	3.22×10^{-2}
	$\pm5.0\times10^{-4}$	$\pm3.8\times10^{-4}$	$\pm1.3\times10^{-3}$	$\pm5.0\times10^{-4}$	$\pm6.5\times10^{-4}$	$\pm4.9\times10^{-4}$	$\pm5.1\times10^{-4}$
300	4.64×10^{-3}	4.91×10^{-3}	8.81×10^{-3}	2.41×10^{-2}	2.42×10^{-2}	2.56×10^{-2}	2.57×10^{-2}
	$\pm2.3\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm3.9\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.7\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm2.4\times10^{-4}$
500	2.78×10^{-3}	2.93×10^{-3}	5.14×10^{-3}	2.33×10^{-2}	2.32×10^{-2}	2.45×10^{-2}	2.44×10^{-2}
	$\pm1.0\times10^{-4}$	$\pm1.2\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.8\times10^{-4}$	$\pm1.2\times10^{-4}$
700	1.95×10^{-3}	2.00×10^{-3}	3.46×10^{-3}	2.30×10^{-2}	2.30×10^{-2}	2.42×10^{-2}	2.41×10^{-2}
	$\pm6.5\times10^{-5}$	$\pm7.8\times10^{-5}$	$\pm2.1\times10^{-4}$	$\pm2.4\times10^{-4}$	$\pm3.0\times10^{-4}$	$\pm2.5\times10^{-4}$	$\pm2.9\times10^{-4}$
1000	1.34×10^{-3}	1.44×10^{-3}	2.48×10^{-3}	2.26×10^{-2}	2.24×10^{-2}	2.38×10^{-2}	2.36×10^{-2}
	$\pm6.0\times10^{-5}$	$\pm6.1\times10^{-5}$	$\pm1.0\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm1.7\times10^{-4}$

Table A.33: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/M/s(t)+E_{10}$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/M/s(t)+M$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	1.07×10^{-1}	1.30×10^{-1}	1.68×10^{-1}	9.60×10^{-2}	1.24×10^{-1}	1.01×10^{-1}	1.28×10^{-1}
	$\pm5.4\times10^{-3}$	$\pm5.9\times10^{-3}$	$\pm8.9\times10^{-3}$	$\pm4.3\times10^{-3}$	$\pm6.3\times10^{-3}$	$\pm4.1\times10^{-3}$	$\pm6.0\times10^{-3}$
30	2.75×10^{-2}	3.49×10^{-2}	5.13×10^{-2}	4.10×10^{-2}	5.01×10^{-2}	4.63×10^{-2}	5.51×10^{-2}
	$\pm1.6\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm2.6\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm2.3\times10^{-3}$	$\pm2.0\times10^{-3}$	$\pm2.5\times10^{-3}$
50	1.55×10^{-2}	1.97×10^{-2}	3.19×10^{-2}	3.20×10^{-2}	3.72×10^{-2}	3.72×10^{-2}	4.23×10^{-2}
	$\pm5.5\times10^{-4}$	$\pm7.7\times10^{-4}$	$\pm9.2\times10^{-4}$	$\pm1.4\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm2.0\times10^{-3}$
70	1.08×10^{-2}	1.39×10^{-2}	2.30×10^{-2}	2.82×10^{-2}	3.17×10^{-2}	3.36×10^{-2}	3.69×10^{-2}
	$\pm2.5\times10^{-4}$	$\pm5.3\times10^{-4}$	$\pm8.6\times10^{-4}$	$\pm8.0\times10^{-4}$	$\pm1.1\times10^{-3}$	$\pm9.0\times10^{-4}$	$\pm1.1\times10^{-3}$
100	7.16×10^{-3}	9.27×10^{-3}	1.57×10^{-2}	2.46×10^{-2}	2.68×10^{-2}	3.00×10^{-2}	3.22×10^{-2}
	$\pm2.0\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm5.2\times10^{-4}$	$\pm3.8\times10^{-4}$	$\pm4.4\times10^{-4}$	$\pm4.4\times10^{-4}$	$\pm5.0\times10^{-4}$
300	2.50×10^{-3}	3.21×10^{-3}	5.63×10^{-3}	2.13×10^{-2}	2.19×10^{-2}	2.70×10^{-2}	2.75×10^{-2}
	$\pm5.6\times10^{-5}$	$\pm9.7\times10^{-5}$	$\pm2.1\times10^{-4}$	$\pm4.1\times10^{-4}$	$\pm4.1\times10^{-4}$	$\pm4.4\times10^{-4}$	$\pm4.5\times10^{-4}$
500	1.48×10^{-3}	1.91×10^{-3}	3.44×10^{-3}	2.03×10^{-2}	2.08×10^{-2}	2.61×10^{-2}	2.65×10^{-2}
	$\pm3.6\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm1.1\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.5\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.4\times10^{-4}$
700	1.04×10^{-3}	1.38×10^{-3}	2.48×10^{-3}	1.99×10^{-2}	2.01×10^{-2}	2.57×10^{-2}	2.58×10^{-2}
	$\pm2.1\times10^{-5}$	$\pm1.9\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm1.5\times10^{-4}$	$\pm2.2\times10^{-4}$	$\pm1.7\times10^{-4}$	$\pm2.4\times10^{-4}$
1000	7.30×10^{-4}	9.79×10^{-4}	1.77×10^{-3}	1.95×10^{-2}	1.96×10^{-2}	2.53×10^{-2}	2.53×10^{-2}
	$\pm2.0\times10^{-5}$	$\pm2.0\times10^{-5}$	$\pm6.2\times10^{-5}$	$\pm2.1\times10^{-4}$	$\pm2.8\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm2.9\times10^{-4}$

Table A.34: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/M/s(t)+M$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

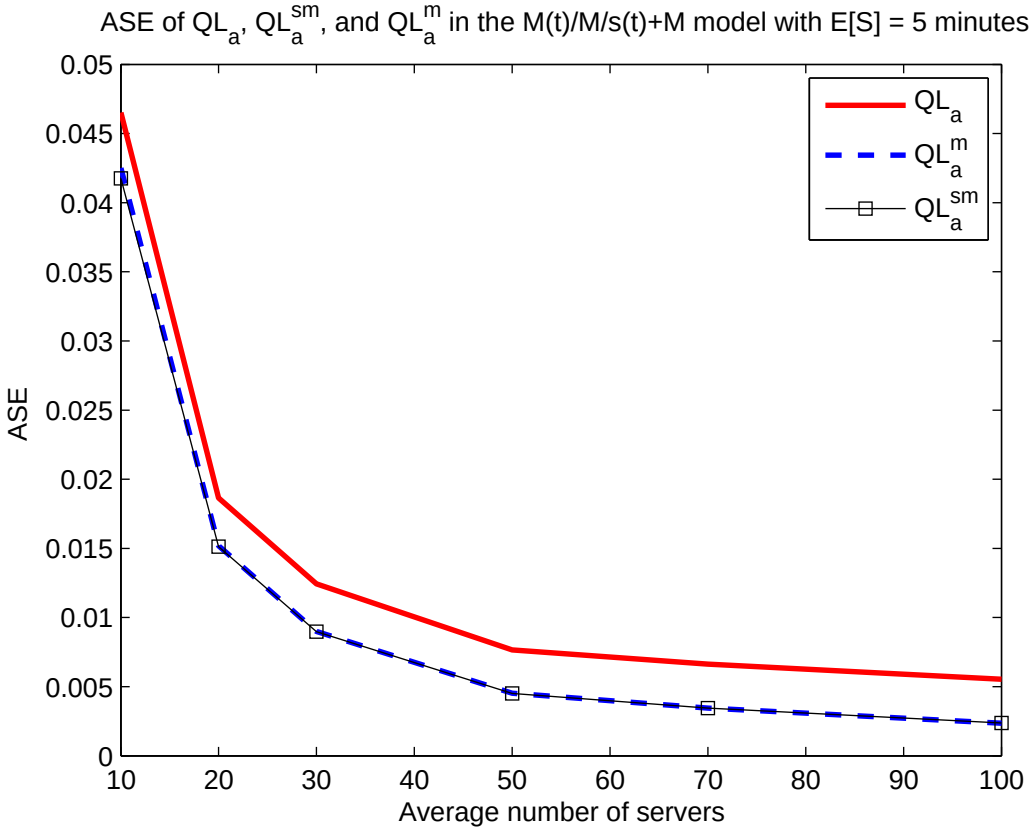

Figure A.5: ASE of QL_a, QL_a^{sm}, and QL_a^{m} in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a small average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 0.022$ which corresponds to $E[S]$ = 5 minutes with a 24 hour arrival-rate cycle.

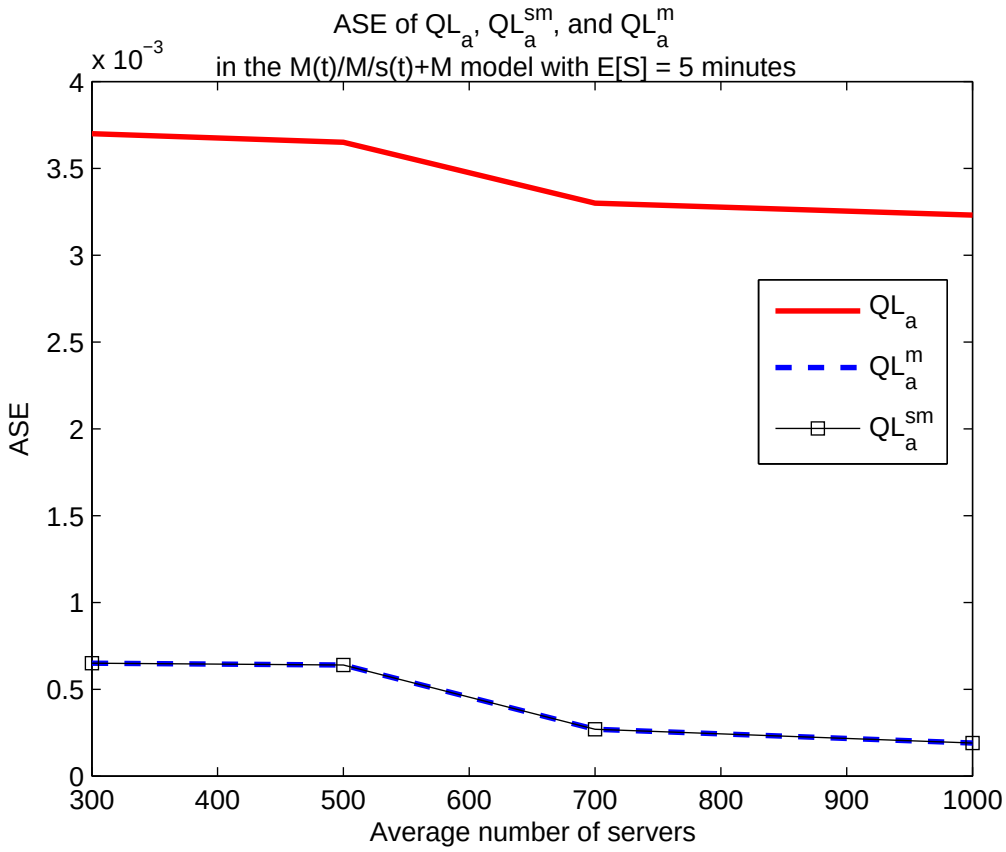

Figure A.6: ASE of QL_a, QL_a^{sm}, and QL_a^m in the $M(t)/M/s(t) + M$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a large average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 0.022$ which corresponds to $E[S] = 5$ minutes with a 24 hour arrival-rate cycle.

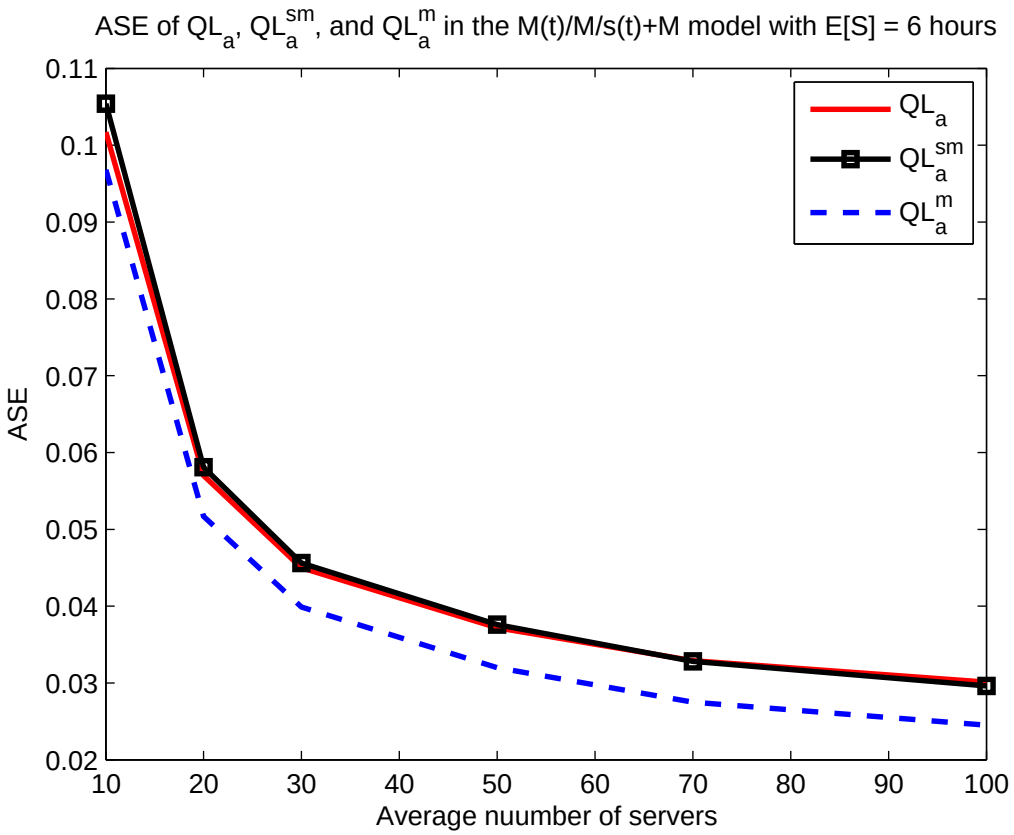

Figure A.7: ASE of QL_a, QL_a^{sm}, and QL_a^m in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a small average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S] = 6$ hours with a 24 hour arrival-rate cycle.

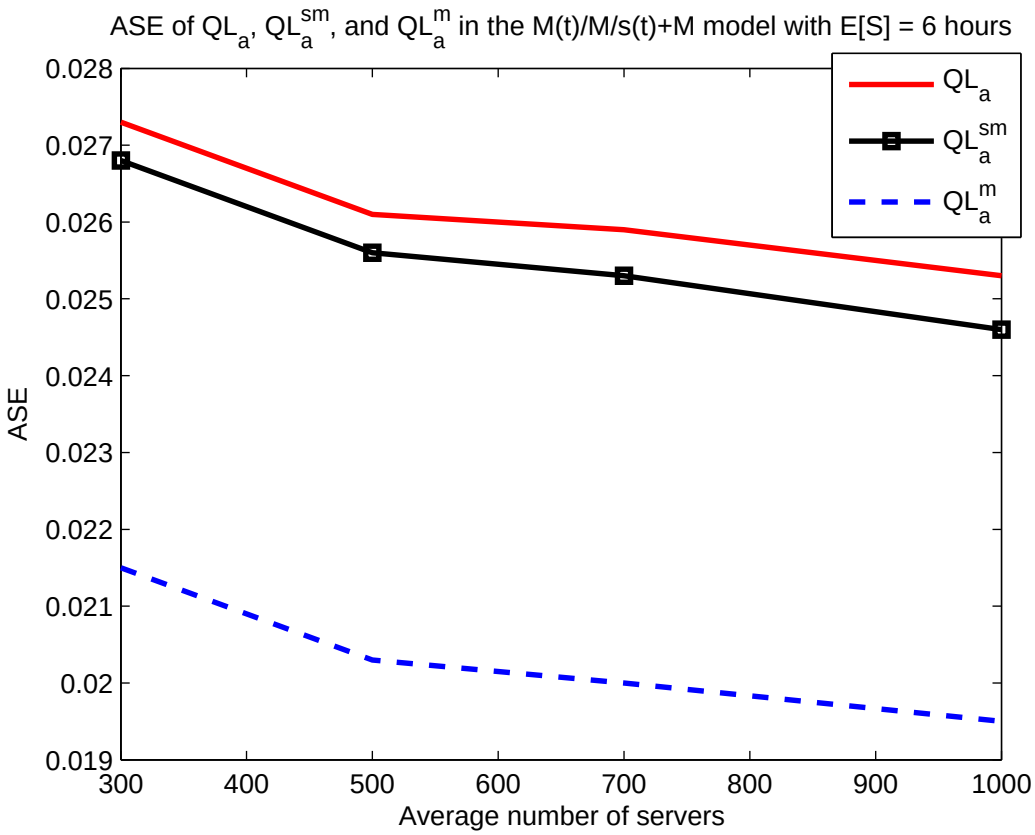

Figure A.8: ASE of QL_a, QL_a^{sm}, and QL_a^m in the $M(t)/M/s(t)+M$ model for $\lambda(t)$ in (5.27) and $s(t)$ in (5.28), and a large average number of servers, $\bar{s}$. We let $\gamma_a = \gamma_s = 1.57$ which corresponds to $E[S]$ = 6 hours with a 24 hour arrival-rate cycle.

ASE of the predictors in the $M(t)/H_2/s(t)+H_2$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	2.07×10^{-1}	2.34×10^{-1}	2.86×10^{-1}	2.01×10^{-1}	2.23×10^{-1}	2.55×10^{-1}	2.76×10^{-1}
	$\pm3.4\times10^{-2}$	$\pm3.0\times10^{-2}$	$\pm5.0\times10^{-2}$	$\pm4.0\times10^{-2}$	$\pm3.4\times10^{-2}$	$\pm3.4\times10^{-2}$	$\pm4.0\times10^{-2}$
20	7.05×10^{-2}	8.39×10^{-2}	1.10×10^{-1}	8.88×10^{-2}	1.02×10^{-1}	1.11×10^{-1}	1.25×10^{-1}
	$\pm1.6\times10^{-2}$	$\pm1.5\times10^{-2}$	$\pm2.9\times10^{-2}$	$\pm2.1\times10^{-2}$	$\pm2.2\times10^{-2}$	$\pm1.9\times10^{-2}$	$\pm2.4\times10^{-2}$
30	3.91×10^{-2}	4.84×10^{-2}	6.85×10^{-2}	5.94×10^{-2}	6.75×10^{-2}	7.53×10^{-2}	8.41×10^{-2}
	$\pm7.1\times10^{-3}$	$\pm7.7\times10^{-3}$	$\pm1.1\times10^{-2}$	$\pm1.3\times10^{-2}$	$\pm1.2\times10^{-2}$	$\pm1.1\times10^{-2}$	$\pm1.4\times10^{-2}$
50	2.49×10^{-2}	3.11×10^{-2}	4.56×10^{-2}	4.49×10^{-2}	4.96×10^{-2}	5.63×10^{-2}	6.09×10^{-2}
	$\pm4.0\times10^{-3}$	$\pm4.7\times10^{-3}$	$\pm7.0\times10^{-3}$	$\pm7.1\times10^{-3}$	$\pm7.1\times10^{-3}$	$\pm6.5\times10^{-3}$	$\pm7.7\times10^{-3}$
70	1.89×10^{-2}	2.15×10^{-2}	3.39×10^{-2}	3.79×10^{-2}	4.01×10^{-2}	4.79×10^{-2}	5.01×10^{-2}
	$\pm3.5\times10^{-3}$	$\pm3.7\times10^{-3}$	$\pm4.9\times10^{-3}$	$\pm7.3\times10^{-3}$	$\pm6.7\times10^{-3}$	$\pm6.5\times10^{-3}$	$\pm7.5\times10^{-3}$
100	1.25×10^{-2}	1.55×10^{-2}	2.51×10^{-2}	3.10×10^{-2}	3.32×10^{-2}	3.99×10^{-2}	4.20×10^{-2}
	$\pm1.1\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm2.0\times10^{-3}$	$\pm3.0\times10^{-3}$	$\pm2.8\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm3.1\times10^{-3}$
300	6.75×10^{-3}	7.80×10^{-3}	1.49×10^{-2}	2.55×10^{-2}	2.59×10^{-2}	3.31×10^{-2}	3.34×10^{-2}
	$\pm4.9\times10^{-4}$	$\pm5.3\times10^{-4}$	$\pm8.8\times10^{-4}$	$\pm1.5\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm1.5\times10^{-3}$
500	5.31×10^{-3}	5.77×10^{-3}	1.12×10^{-2}	2.32×10^{-2}	2.31×10^{-2}	3.04×10^{-2}	3.02×10^{-2}
	$\pm4.4\times10^{-4}$	$\pm3.8\times10^{-4}$	$\pm5.6\times10^{-4}$	$\pm1.4\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm1.4\times10^{-3}$
700	4.67×10^{-3}	5.18×10^{-3}	1.04×10^{-2}	2.26×10^{-2}	2.25×10^{-2}	2.97×10^{-2}	2.95×10^{-2}
	$\pm1.9\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm4.2\times10^{-4}$	$\pm9.3\times10^{-4}$	$\pm7.6\times10^{-4}$	$\pm8.2\times10^{-4}$	$\pm8.7\times10^{-4}$
1000	4.11×10^{-3}	4.52×10^{-3}	9.16×10^{-3}	2.20×10^{-2}	2.19×10^{-2}	2.90×10^{-2}	2.87×10^{-2}
	$\pm2.0\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm2.8\times10^{-4}$	$\pm7.9\times10^{-4}$	$\pm6.8\times10^{-4}$	$\pm6.7\times10^{-4}$	$\pm7.8\times10^{-4}$

Table A.35: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/H_2/s(t)+H_2$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/E_{10}/s(t)+H_2$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	4.95×10^{-2}	6.86×10^{-2}	6.92×10^{-2}	3.17×10^{-2}	4.80×10^{-2}	4.82×10^{-2}	6.39×10^{-2}
	$\pm2.4\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm3.6\times10^{-3}$	$\pm1.8\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm2.4\times10^{-3}$
20	2.26×10^{-2}	2.83×10^{-2}	3.41×10^{-2}	1.34×10^{-2}	2.07×10^{-2}	1.90×10^{-2}	2.66×10^{-2}
	$\pm7.8\times10^{-4}$	$\pm1.1\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm5.6\times10^{-4}$	$\pm8.7\times10^{-4}$	$\pm4.9\times10^{-4}$	$\pm9.5\times10^{-4}$
30	1.56×10^{-2}	1.87×10^{-2}	2.58×10^{-2}	9.19×10^{-3}	1.42×10^{-2}	1.27×10^{-2}	1.80×10^{-2}
	$\pm3.3\times10^{-4}$	$\pm3.0\times10^{-4}$	$\pm6.6\times10^{-4}$	$\pm3.0\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm2.3\times10^{-4}$	$\pm4.3\times10^{-4}$
50	1.05×10^{-2}	1.16×10^{-2}	1.99×10^{-2}	6.09×10^{-3}	8.96×10^{-3}	8.41×10^{-3}	1.13×10^{-2}
	$\pm2.8\times10^{-4}$	$\pm2.2\times10^{-4}$	$\pm5.8\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm1.7\times10^{-4}$	$\pm3.2\times10^{-4}$
70	8.45×10^{-3}	8.89×10^{-3}	1.62×10^{-2}	5.01×10^{-3}	7.23×10^{-3}	6.88×10^{-3}	9.13×10^{-3}
	$\pm2.0\times10^{-4}$	$\pm1.7\times10^{-4}$	$\pm4.0\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.7\times10^{-4}$	$\pm6.7\times10^{-5}$	$\pm2.3\times10^{-4}$
100	6.91×10^{-3}	6.95×10^{-3}	1.46×10^{-2}	3.95×10^{-3}	5.42×10^{-3}	5.48×10^{-3}	6.94×10^{-3}
	$\pm1.9\times10^{-4}$	$\pm2.2\times10^{-4}$	$\pm3.7\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.7\times10^{-4}$
300	4.48×10^{-3}	4.05×10^{-3}	1.17×10^{-2}	2.60×10^{-3}	3.07×10^{-3}	3.71×10^{-3}	4.15×10^{-3}
	$\pm8.6\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm9.7\times10^{-5}$	$\pm4.0\times10^{-5}$	$\pm6.2\times10^{-5}$	$\pm3.1\times10^{-5}$	$\pm8.4\times10^{-5}$
500	4.06×10^{-3}	3.42×10^{-3}	1.10×10^{-2}	2.27×10^{-3}	2.55×10^{-3}	3.29×10^{-3}	3.56×10^{-3}
	$\pm2.5\times10^{-5}$	$\pm4.6\times10^{-5}$	$\pm6.2\times10^{-5}$	$\pm4.5\times10^{-5}$	$\pm4.6\times10^{-5}$	$\pm3.0\times10^{-5}$	$\pm6.0\times10^{-5}$
700	3.84×10^{-3}	3.21×10^{-3}	1.08×10^{-2}	2.17×10^{-3}	2.36×10^{-3}	3.15×10^{-3}	3.34×10^{-3}
	$\pm4.7\times10^{-5}$	$\pm3.9\times10^{-5}$	$\pm9.5\times10^{-5}$	$\pm1.8\times10^{-5}$	$\pm2.5\times10^{-5}$	$\pm1.3\times10^{-5}$	$\pm3.0\times10^{-5}$
1000	3.72×10^{-3}	2.99×10^{-3}	1.05×10^{-2}	2.09×10^{-3}	2.23×10^{-3}	3.05×10^{-3}	3.18×10^{-3}
	$\pm3.9\times10^{-5}$	$\pm2.8\times10^{-5}$	$\pm7.4\times10^{-5}$	$\pm2.5\times10^{-5}$	$\pm3.7\times10^{-5}$	$\pm2.7\times10^{-5}$	$\pm3.2\times10^{-5}$

Table A.36: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/E_{10}/s(t)+H_2$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/H_2/s(t)+E_{10}$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	1.17×10^{-1}	1.58×10^{-1}	2.31×10^{-1}	1.35×10^{-1}	1.35×10^{-1}	1.70×10^{-1}	1.76×10^{-1}
	$\pm2.1\times10^{-2}$	$\pm2.5\times10^{-2}$	$\pm4.2\times10^{-2}$	$\pm3.6\times10^{-2}$	$\pm2.8\times10^{-2}$	$\pm4.2\times10^{-2}$	$\pm3.8\times10^{-2}$
20	7.84×10^{-2}	8.73×10^{-2}	1.52×10^{-1}	1.04×10^{-1}	9.80×10^{-2}	1.24×10^{-1}	1.22×10^{-1}
	$\pm1.3\times10^{-2}$	$\pm1.1\times10^{-2}$	$\pm2.2\times10^{-2}$	$\pm1.8\times10^{-2}$	$\pm1.5\times10^{-2}$	$\pm2.0\times10^{-2}$	$\pm1.8\times10^{-2}$
30	4.98×10^{-2}	5.76×10^{-2}	1.06×10^{-1}	7.09×10^{-2}	7.04×10^{-2}	8.75×10^{-2}	8.80×10^{-2}
	$\pm9.8\times10^{-3}$	$\pm9.9\times10^{-3}$	$\pm2.0\times10^{-2}$	$\pm1.8\times10^{-2}$	$\pm1.5\times10^{-2}$	$\pm1.8\times10^{-2}$	$\pm1.5\times10^{-2}$
50	3.02×10^{-2}	3.44×10^{-2}	6.77×10^{-2}	4.74×10^{-2}	4.99×10^{-2}	5.93×10^{-2}	6.19×10^{-2}
	$\pm5.4\times10^{-3}$	$\pm6.9\times10^{-3}$	$\pm1.5\times10^{-2}$	$\pm7.2\times10^{-3}$	$\pm8.1\times10^{-3}$	$\pm7.2\times10^{-3}$	$\pm8.0\times10^{-3}$
70	2.61×10^{-2}	2.82×10^{-2}	5.94×10^{-2}	4.22×10^{-2}	4.27×10^{-2}	5.38×10^{-2}	5.42×10^{-2}
	$\pm1.8\times10^{-3}$	$\pm2.1\times10^{-3}$	$\pm8.8\times10^{-3}$	$\pm4.2\times10^{-3}$	$\pm3.8\times10^{-3}$	$\pm5.0\times10^{-3}$	$\pm4.8\times10^{-3}$
100	2.14×10^{-2}	2.25×10^{-2}	4.90×10^{-2}	4.24×10^{-2}	4.19×10^{-2}	5.38×10^{-2}	5.33×10^{-2}
	$\pm3.2\times10^{-3}$	$\pm3.0\times10^{-3}$	$\pm6.1\times10^{-3}$	$\pm5.1\times10^{-3}$	$\pm5.1\times10^{-3}$	$\pm5.2\times10^{-3}$	$\pm5.3\times10^{-3}$
300	1.30×10^{-2}	1.33×10^{-2}	3.13×10^{-2}	3.25×10^{-2}	3.27×10^{-2}	4.20×10^{-2}	4.20×10^{-2}
	$\pm2.1\times10^{-3}$	$\pm1.8\times10^{-3}$	$\pm3.8\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.8\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.7\times10^{-3}$
500	1.32×10^{-2}	1.26×10^{-2}	3.14×10^{-2}	3.12×10^{-2}	3.10×10^{-2}	4.00×10^{-2}	3.96×10^{-2}
	$\pm1.4\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm3.9\times10^{-3}$	$\pm9.2\times10^{-4}$	$\pm1.3\times10^{-3}$	$\pm9.8\times10^{-4}$	$\pm1.2\times10^{-3}$
700	1.37×10^{-2}	1.24×10^{-2}	2.87×10^{-2}	3.10×10^{-2}	3.01×10^{-2}	3.94×10^{-2}	3.84×10^{-2}
	$\pm1.1\times10^{-3}$	$\pm7.6\times10^{-4}$	$\pm2.8\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm1.3\times10^{-3}$
1000	1.23×10^{-2}	1.14×10^{-2}	2.47×10^{-2}	3.14×10^{-2}	3.08×10^{-2}	4.02×10^{-2}	3.94×10^{-2}
	$\pm9.5\times10^{-4}$	$\pm8.7\times10^{-4}$	$\pm2.1\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm1.3\times10^{-3}$

Table A.37: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/H_2/s(t)+E_{10}$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/E_{10}/s(t)+E_{10}$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	7.19×10^{-2}	1.10×10^{-1}	1.13×10^{-1}	4.36×10^{-2}	6.86×10^{-2}	5.73×10^{-2}	6.91×10^{-2}
	$\pm1.2\times10^{-2}$	$\pm1.1\times10^{-2}$	$\pm1.5\times10^{-2}$	$\pm2.6\times10^{-3}$	$\pm4.3\times10^{-3}$	$\pm4.5\times10^{-3}$	$\pm4.4\times10^{-3}$
20	6.60×10^{-2}	6.62×10^{-2}	7.88×10^{-2}	2.75×10^{-2}	3.45×10^{-2}	3.36×10^{-2}	3.75×10^{-2}
	$\pm9.7\times10^{-3}$	$\pm6.7\times10^{-3}$	$\pm7.8\times10^{-3}$	$\pm2.3\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm3.0\times10^{-3}$	$\pm3.1\times10^{-3}$
30	4.33×10^{-2}	4.61×10^{-2}	5.01×10^{-2}	1.93×10^{-2}	2.52×10^{-2}	2.42×10^{-2}	2.79×10^{-2}
	$\pm6.9\times10^{-3}$	$\pm4.4\times10^{-3}$	$\pm4.5\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm2.8\times10^{-3}$
50	3.60×10^{-2}	3.44×10^{-2}	3.60×10^{-2}	1.56×10^{-2}	1.87×10^{-2}	1.79×10^{-2}	2.02×10^{-2}
	$\pm5.1\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm4.3\times10^{-3}$	$\pm4.8\times10^{-4}$	$\pm7.8\times10^{-4}$	$\pm1.1\times10^{-3}$	$\pm1.2\times10^{-3}$
70	3.46×10^{-2}	3.26×10^{-2}	3.67×10^{-2}	1.44×10^{-2}	1.67×10^{-2}	1.56×10^{-2}	1.71×10^{-2}
	$\pm5.0\times10^{-3}$	$\pm4.5\times10^{-3}$	$\pm5.5\times10^{-3}$	$\pm5.7\times10^{-4}$	$\pm6.2\times10^{-4}$	$\pm9.1\times10^{-4}$	$\pm1.1\times10^{-3}$
100	3.00×10^{-2}	2.90×10^{-2}	2.90×10^{-2}	1.29×10^{-2}	1.41×10^{-2}	1.40×10^{-2}	1.49×10^{-2}
	$\pm2.6\times10^{-3}$	$\pm2.4\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm5.5\times10^{-4}$	$\pm6.2\times10^{-4}$	$\pm5.7\times10^{-4}$	$\pm4.7\times10^{-4}$
300	2.54×10^{-2}	2.26×10^{-2}	2.16×10^{-2}	1.01×10^{-2}	1.05×10^{-2}	1.12×10^{-2}	1.15×10^{-2}
	$\pm2.0\times10^{-3}$	$\pm1.3\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm2.8\times10^{-4}$	$\pm4.8\times10^{-4}$	$\pm2.5\times10^{-4}$	$\pm3.1\times10^{-4}$
500	2.19×10^{-2}	2.08×10^{-2}	1.86×10^{-2}	9.65×10^{-3}	9.81×10^{-3}	1.01×10^{-2}	1.03×10^{-2}
	$\pm1.5\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.5\times10^{-3}$	$\pm1.7\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm2.6\times10^{-4}$	$\pm2.3\times10^{-4}$
700	2.23×10^{-2}	2.10×10^{-2}	1.89×10^{-2}	9.42×10^{-3}	9.56×10^{-3}	9.90×10^{-3}	1.00×10^{-2}
	$\pm8.9\times10^{-4}$	$\pm7.6\times10^{-4}$	$\pm7.2\times10^{-4}$	$\pm1.4\times10^{-4}$	$\pm1.5\times10^{-4}$	$\pm1.9\times10^{-4}$	$\pm2.2\times10^{-4}$
1000	2.24×10^{-2}	2.09×10^{-2}	1.91×10^{-2}	9.27×10^{-3}	9.36×10^{-3}	9.91×10^{-3}	1.01×10^{-2}
	$\pm1.1\times10^{-3}$	$\pm8.5\times10^{-4}$	$\pm1.0\times10^{-3}$	$\pm1.4\times10^{-4}$	$\pm2.9\times10^{-4}$	$\pm1.4\times10^{-4}$	$\pm2.2\times10^{-4}$

Table A.38: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/E_{10}/s(t)+E_{10}$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a=\gamma_s=1.57$ (corresponding to $E[S]=6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/D/s(t)+H_2$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	4.80×10^{-2}	6.38×10^{-2}	6.58×10^{-2}	2.73×10^{-2}	4.23×10^{-2}	4.19×10^{-2}	5.61×10^{-2}
	$\pm1.4\times10^{-3}$	$\pm2.1\times10^{-3}$	$\pm3.0\times10^{-3}$	$\pm1.9\times10^{-3}$	$\pm1.7\times10^{-3}$	$\pm1.6\times10^{-3}$	$\pm2.1\times10^{-3}$
20	2.22×10^{-2}	2.78×10^{-2}	3.29×10^{-2}	1.19×10^{-2}	1.90×10^{-2}	1.69×10^{-2}	2.41×10^{-2}
	$\pm7.9\times10^{-4}$	$\pm8.6\times10^{-4}$	$\pm4.7\times10^{-4}$	$\pm5.0\times10^{-4}$	$\pm5.1\times10^{-4}$	$\pm3.9\times10^{-4}$	$\pm6.6\times10^{-4}$
30	1.60×10^{-2}	1.86×10^{-2}	2.56×10^{-2}	8.29×10^{-3}	1.29×10^{-2}	1.15×10^{-2}	1.62×10^{-2}
	$\pm4.3\times10^{-4}$	$\pm4.9\times10^{-4}$	$\pm5.1\times10^{-4}$	$\pm3.6\times10^{-4}$	$\pm4.4\times10^{-4}$	$\pm2.9\times10^{-4}$	$\pm5.1\times10^{-4}$
50	1.16×10^{-2}	1.23×10^{-2}	2.04×10^{-2}	5.74×10^{-3}	8.44×10^{-3}	7.69×10^{-3}	1.04×10^{-2}
	$\pm5.0\times10^{-4}$	$\pm5.1\times10^{-4}$	$\pm4.2\times10^{-4}$	$\pm1.9\times10^{-4}$	$\pm2.7\times10^{-4}$	$\pm1.9\times10^{-4}$	$\pm2.7\times10^{-4}$
70	1.01×10^{-2}	1.00×10^{-2}	1.80×10^{-2}	4.76×10^{-3}	6.74×10^{-3}	6.37×10^{-3}	8.35×10^{-3}
	$\pm3.1\times10^{-4}$	$\pm3.3\times10^{-4}$	$\pm4.0\times10^{-4}$	$\pm1.6\times10^{-4}$	$\pm2.1\times10^{-4}$	$\pm1.3\times10^{-4}$	$\pm2.5\times10^{-4}$
100	8.64×10^{-3}	8.12×10^{-3}	1.66×10^{-2}	3.88×10^{-3}	5.30×10^{-3}	5.23×10^{-3}	6.63×10^{-3}
	$\pm2.8\times10^{-4}$	$\pm1.7\times10^{-4}$	$\pm3.0\times10^{-4}$	$\pm1.2\times10^{-4}$	$\pm1.7\times10^{-4}$	$\pm1.1\times10^{-4}$	$\pm1.8\times10^{-4}$
300	6.73×10^{-3}	5.86×10^{-3}	1.43×10^{-2}	2.70×10^{-3}	3.18×10^{-3}	3.64×10^{-3}	4.11×10^{-3}
	$\pm8.1\times10^{-5}$	$\pm7.2\times10^{-5}$	$\pm8.1\times10^{-5}$	$\pm4.4\times10^{-5}$	$\pm4.0\times10^{-5}$	$\pm4.1\times10^{-5}$	$\pm4.5\times10^{-5}$
500	6.25×10^{-3}	5.18×10^{-3}	1.36×10^{-2}	2.41×10^{-3}	2.67×10^{-3}	3.29×10^{-3}	3.56×10^{-3}
	$\pm6.0\times10^{-5}$	$\pm6.7\times10^{-5}$	$\pm1.0\times10^{-4}$	$\pm3.9\times10^{-5}$	$\pm6.5\times10^{-5}$	$\pm4.2\times10^{-5}$	$\pm6.3\times10^{-5}$
700	6.11×10^{-3}	5.06×10^{-3}	1.35×10^{-2}	2.33×10^{-3}	2.53×10^{-3}	3.18×10^{-3}	3.36×10^{-3}
	$\pm1.0\times10^{-4}$	$\pm5.6\times10^{-5}$	$\pm1.1\times10^{-4}$	$\pm3.1\times10^{-5}$	$\pm4.0\times10^{-5}$	$\pm2.7\times10^{-5}$	$\pm4.8\times10^{-5}$
1000	5.96×10^{-3}	4.83×10^{-3}	1.34×10^{-2}	2.20×10^{-3}	2.32×10^{-3}	3.02×10^{-3}	3.13×10^{-3}
	$\pm5.8\times10^{-5}$	$\pm5.3\times10^{-5}$	$\pm9.1\times10^{-5}$	$\pm2.9\times10^{-5}$	$\pm4.6\times10^{-5}$	$\pm3.4\times10^{-5}$	$\pm3.3\times10^{-5}$

Table A.39: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/D/s(t)+H_2$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

ASE of the predictors in the $M(t)/D/s(t)+E_{10}$ model as a function of $\bar{s}$

$\bar{s}$	QL_{rt}	HOL_{rt}	NIF	QI_a^m	HOL_a^m	QL_a	HOL_a
10	8.72×10^{-2}	1.09×10^{-1}	1.15×10^{-1}	4.30×10^{-2}	6.20×10^{-2}	5.95×10^{-2}	6.64×10^{-2}
	$\pm2.1\times10^{-2}$	$\pm1.3\times10^{-2}$	$\pm1.2\times10^{-2}$	$\pm4.0\times10^{-3}$	$\pm4.9\times10^{-3}$	$\pm6.9\times10^{-3}$	$\pm4.2\times10^{-3}$
20	6.00×10^{-2}	6.31×10^{-2}	7.09×10^{-2}	2.42×10^{-2}	3.28×10^{-2}	3.09×10^{-2}	3.57×10^{-2}
	$\pm1.1\times10^{-2}$	$\pm5.3\times10^{-3}$	$\pm7.3\times10^{-3}$	$\pm3.5\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm4.0\times10^{-3}$	$\pm4.0\times10^{-3}$
30	5.49×10^{-2}	5.16×10^{-2}	5.68×10^{-2}	2.02×10^{-2}	2.63×10^{-2}	2.40×10^{-2}	2.73×10^{-2}
	$\pm7.3\times10^{-3}$	$\pm7.9\times10^{-3}$	$\pm7.3\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm2.3\times10^{-3}$	$\pm3.0\times10^{-3}$	$\pm2.8\times10^{-3}$
50	3.84×10^{-2}	3.84×10^{-2}	4.03×10^{-2}	1.64×10^{-2}	1.94×10^{-2}	1.93×10^{-2}	2.15×10^{-2}
	$\pm4.2\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm3.4\times10^{-3}$	$\pm2.2\times10^{-3}$	$\pm2.1\times10^{-3}$	$\pm2.9\times10^{-3}$	$\pm2.9\times10^{-3}$
70	3.82×10^{-2}	3.69×10^{-2}	3.68×10^{-2}	1.56×10^{-2}	1.73×10^{-2}	1.87×10^{-2}	1.98×10^{-2}
	$\pm5.5\times10^{-3}$	$\pm4.1\times10^{-3}$	$\pm4.6\times10^{-3}$	$\pm2.1\times10^{-3}$	$\pm2.2\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm2.7\times10^{-3}$
100	3.69×10^{-2}	3.70×10^{-2}	3.56×10^{-2}	1.53×10^{-2}	1.65×10^{-2}	1.79×10^{-2}	1.86×10^{-2}
	$\pm4.4\times10^{-3}$	$\pm2.6\times10^{-3}$	$\pm2.7\times10^{-3}$	$\pm2.0\times10^{-3}$	$\pm2.2\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm2.6\times10^{-3}$
300	3.21×10^{-2}	3.07×10^{-2}	2.68×10^{-2}	1.32×10^{-2}	1.36×10^{-2}	1.49×10^{-2}	1.52×10^{-2}
	$\pm2.1\times10^{-3}$	$\pm2.6\times10^{-3}$	$\pm2.5\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.5\times10^{-3}$
500	3.14×10^{-2}	2.98×10^{-2}	2.54×10^{-2}	1.32×10^{-2}	1.34×10^{-2}	1.52×10^{-2}	1.53×10^{-2}
	$\pm2.3\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.0\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm1.1\times10^{-3}$
700	3.15×10^{-2}	3.00×10^{-2}	2.51×10^{-2}	1.33×10^{-2}	1.33×10^{-2}	1.51×10^{-2}	1.52×10^{-2}
	$\pm1.6\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.4\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm1.2\times10^{-3}$	$\pm1.3\times10^{-3}$
1000	3.03×10^{-2}	2.85×10^{-2}	2.39×10^{-2}	1.29×10^{-2}	1.29×10^{-2}	1.46×10^{-2}	1.46×10^{-2}
	$\pm1.2\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm1.1\times10^{-3}$	$\pm8.8\times10^{-4}$	$\pm8.8\times10^{-4}$	$\pm1.2\times10^{-3}$	$\pm1.2\times10^{-3}$

Table A.40: Performance of the alternative predictors, as a function of $\bar{s}$, in the $M(t)/D/s(t)+E_{10}$ model with $\lambda(t)$ in (5.27), $s(t)$ in (5.28), and $\gamma_a = \gamma_s = 1.57$ (corresponding to $E[S] = 6$ hours with a 24 hour cycle). Estimates of the ASE are shown together with the half width of the 95% confidence interval. The ASE's are measured in units of mean service time squared per customer.

Bibliography

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

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

Aksin, O.Z., Armony, M. and Mehrotra, V. 2007. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research, *Production and Operations Management*, 16:6, 665 – 688.

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

Allon, G, Bassambo, A. and I. Gurvich. 2010a. We will be right with you: managing customer with vague promises, *Working Paper*, Northwestern Univ., Evanston, IL.

Allon, G, and Bassambo, A. 2010b. The impact of delaying the delay announcements. *Working Paper*, Northwestern Univ., Evanston, IL.

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

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

Asmussen, S. 2003. *Applied Probability and Queues*, second edition, Springer, New York.

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

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

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

Borovkov, A. A. 1967. On limit laws for service processes in multi-channel systems. *Siberian Math.* 8: 746–763.

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.

Choudhury, G. L. and W. Whitt. 1994. Heavy-traffic asymptotic expansions for the asymptotic decay rates in the BMAP/G/1 Queue. *Stochastic Models* 10, 453–498.

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

Cooper, R. B. 1981. *Introduction to Queueing Theory*, second edition, North-Holland, New York.

Dobson, G, and J. Pinker. 2006. The value of sharing lead time information. *IIE Transactions* 38: 171-183.

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

Dube-Rioux, L., Bernd, S., and F. Leclerc. 1988. Consumer reactions to waiting: when delays affect the perception of service quality. *Advances in Consumer Research,* 59-63.

Duenysas, L. and W. Hopp. 1995. Quoting Customer Lead Times. *Management Sci.* 41: 43-57.

Eick, S., W.A. Massey, W. Whitt. 1993a. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* 41: 731–742.

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

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*, vol. II, second ed., Wiley, New York.

Feldman, Z., Mandelbaum, A., Massey, W., and W. Whitt. 2008. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, vol. 54, No.2, February 2008, pp. 324-338.

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

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

Glynn, P. W. and W. Whitt. 1989. Indirect estimation via $L = \lambda W$. *Operations Research* 37: 82-103.

Glynn, P. W. and W. Whitt. 1994. Logarithmic Asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* 31A, 131–156 (also called *Studies in Applied Probability, Papers in Honour of Lajos Takcs*, J. Galambos and J. Gani (eds.), Applied Probability Trust, Sheffield, England).

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

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

Halfin, S. and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29: 567-588.

Hassin, R. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica*. 54: 1185-1195.

Hassin, R. and M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer.

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

Hui, M. and D. Tse. 1996. What to tell customers in waits of different lengths: an integrative model of service evaluation. *Journal of Marketing*. 60: 81–90.

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. 2010a. Real-time Delay Estimation Based on Delay History in Many-Server Queues 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.

Iglehart, D. L. and W. Whitt. 1970. Multiple channel queues in heavy traffic II: sequences, networks, and batches. *Advances in Applied Probability* 2: 355-369.

Jelenkovic P., A. Mandelbaum A. and P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* 47: 53-69.

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

Jouini, O. Y. Dallery and Z. Aksin. 2010. Modeling call centers with delay information. *Working Paper*. Ecole Centrale Paris. Paris, France.

Katz, K., Larson, B, and R. Larson. 1999. Prescription for the waiting-in-line blues: Entertain, enlighten, and engage. *Sloan Management Review*. 44-54.

Larson, R. C. 1987. Perspectives on queues: social justice and the psychology of queueing. *Operations Research*. 35(6): 895-905.

Larson, R. C. 1990. The queue inference engine: deducing queue

Larson, R. C. 1987. Perspectives on queues: social justice and the psychology of queueing. *Operations Research.* 35(6): 895-905. statistics from transactional data. *Management Sci.* 36: 586-601.

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.

Maister, D. 1984. Psychology of waiting lines. *Harvard Business School Cases*. 71–78.

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

Mandelbaum A. and Zeltyn S. 2004. The Impact of Customers' Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue. OR Spectrum, 26 (3), 377-411. Special Issue on Call Centers.

Mandelbaum A. and Zeltyn S. 2007. Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers. Advances in Services Innovations, pp. 17-48, Spath D., Fhnrich, K.-P. (Eds.), Springer-Verlag

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

Morton, T. and A. Vepsalainen. 1987. Priority Rules and Leadtime Estimation for Job Shop Scheduling with Weighted Tardiness Costs. *Management Sci.*, 33: 1036-1047.

Munichor, N., A. Rafaeli. 2007. Number of apologies? Customer reactions to tele-waiting time fillers. *J. Applied Psychology,* 92(2):511-8

Naor, P. 1969. The Regulation of Queue Size by Levying Tolls. *Econometrica*. 37:15-24.

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

Neuts, M. F. 1986. The caudal characteristic curve of queues. *Adv. Appl. Probab.* 18, 221–254.

Ornek, M. and P. Collier. 1988. The Determination of In-Process Inventory and Manufacturing Lead Time in Multi-Stage Production Sys- tems. *International J. Oper. and Production Management*, 8: 74-80.

Pang, G., R. Talreja and W.Whitt. 2007. Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues. *Probability Surveys*, Vol. 4 (2007), pp. 193-267.

Press Ganey Pulse Report. 2009. Emergency departments: Patient perspectives on American healthcare. Available online at http://www.pressganey.com/.

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

Reiman, M. I. 1982. The heavy traffic diffusion approximation for sojourn times in Jackson networks. In *Applied Probability – Computer Science, the Interface, II*, R. L. Disney and T. J. Ott (eds.), Birhauser, Boston, 409-422.

Ross, S. M. 1996. *Stochastic Processes*, second edition, Wiley, New York.

Shanthikumar, J. and U. Sumita. 1988. Approximations for the Time Spent in a Dynamic Job Shop with Applications to Due Date Assign- ment. *International J. Production Research*, 26: 1329- 1352.

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

Shen, H. and J. Huang. 2008b. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Ann. of applied stat.* 2: 601–623.

Smith, W. L. 1953. On the distribution of queueing times. *Proc. Camb. Phil. Soc.* 49, 449–461.

Spearman, M. and R. Zhang. 1999. Optimal lead time policies. *Management Sci.* 45: 290-295.

Talreja R. and W. Whitt. 2009. Heavy-Traffic Limits for Waiting Times in Many-Server Queues with Abandonment. *Annals of Applied Probability*, Vol. 19, No. 6 , pp. 2137-2175

Taylor, S. 1994. Waiting for service: the relationship between delays and evaluations of service. *Journal of Marketing,* 58:56-69.

US Economy in Brief Report. 2008. Available online at http://www.america.gov.

Vocalabs National Customer Service Survey for Computer Tech Support. 2010. Available online at http://www.vocalabs.com.

Ward, A. W. and W. Whitt. 2000. Predicting response times in processor-sharing queues. In *Analysis of Communication Networks: Call Centres, Traffic and*

Performance, D. R. McDonald and S. R. E. Turner (eds.), Fields Institute Communications 28, American Math. Society, Providence, RI, 1-29.

Whitt, W. 1982a. Approximating a point process by a renewal process: two basic methods. *Operations Research* 30: 125-147.

Whitt, W. 1982b. On the Heavy-Traffic Limit Theorem For $GI/G/\inf$ Queues. *Advances in Applied Probability*. 14: 171–190.

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

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

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

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

Whitt, W. 2004a. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50, 1449–1461.

Whitt, W. 2004b. A diffusion approximation for the G/GI/n/m queue. *Operations Research* 52, 922–941.

Whitt, W. 2005a. Heavy-traffic limits for the G/H2*/n/m queue. *Math. Oper. Res.* 30, 1–27.

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

Whitt, W. 2006. Fluid Models for Multiserver Queues with Abandonments. *Operations Research* 54: 37-54.

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

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.

Zeltyn, S. and A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems* 51(3-4): 361-402