A General Framework to Compare Announcement Accuracy: Static vs LES-based Announcement

Achal Bassamboo
Kellogg School of Management, Northwestern University, IL 60208 a-bassamboo@northwestern.edu

Rouba Ibrahim
School of Management, University College London, London E14 5AB rouba.ibrahim@ucl.ac.uk

Service providers often share delay information, in the form of delay announcements, with their customers. In practice, simple delay announcements, such as average waiting times or a weighted average of previously delayed customers, are often used. Our goal in this paper is to gain insight into when such announcements perform well. Specifically, we compare the accuracies of two announcements: (i) *a static announcement* which does not exploit real-time information about the state of the system, and (ii) *a dynamic announcement*, specifically the last-to-enter-service (LES) announcement, which equals the delay of the last customer to have entered service at the time of the announcement. So far, the literature on delay announcements has focused, for the most part, on studying the accuracy of various types of announcements in specific queueing models. As such, the validity of various announcements remains intimately tied to the appropriateness of the technical conditions under which these models were studied. In other words, in order to decide whether a type of announcement is appropriate for her system, a manager would need to assess whether a certain queueing model is a good fit for her data. Because this is often difficult to do, we propose a new approach in this paper. In particular, we propose a novel correlation-based approach which is theoretically appealing because it allows for a comparison of the performances of announcements across different queueing models, including multi-class models with a priority service discipline. It is also practically useful because estimating correlations is much easier than fitting an entire queueing model. Using a combination of queuing-theoretic analysis, real-life data analysis, and simulation, we analyze the performance of static and dynamic announcements, and derive an appropriate weighted average of the two which we demonstrate has a superior performance using both simulation and data from a call center.

Key words: delay announcements; many-server queues; correlation; accuracy.

1. Introduction

In this paper, we consider service systems where information about upcoming waiting times is shared with customers, in the form of delay announcements. In practice, simple, easy-to-implement, *static* announcements, such as average waiting times over a long period, are predominately used, e.g., in hospitals[1], retail stores[2], and immigration and border controls[3]. The real-life prevalence of average-wait-time announcements is the main motivation behind this paper. We refer to these announcements as static, as they change relatively slowly over time, and do not use real-time system-state information.

Since static announcements are so common, it is natural to investigate when they would be accurate in predicting customer delay, in real time, relative to dynamic announcements. To assess accuracy, it is useful to have a frame of reference in mind. In this paper, we compare the performances of *static* and *delay-history-based dynamic* announcements, where the latter is based on information about the recent history of delays in the system at the time of the announcement. Throughout this paper, we assume that a delay announcement is only made at the arrival epoch of a delayed customer, and that the announcement is not updated thereafter. Further, we assume that customer behavior is not altered by the delay announcement (for exceptions, see Armony et al. (2009) and Ibrahim et al. (2016)).

1.1. Comparing Static and Delay-History-Based Announcements: Why is this Hard?

We would like to emphasize that such a comparison is not trivial. One predominant insight in the literature is that, under steady-state conditions, exploiting real-time information is valuable because it mitigates the effect of uncertainty (Whitt 1999). Specifically, let a delay-history-based announcement be equal to the conditional expected value of the steady-state waiting time (assuming that it is well defined) given some recent delay information, and let the static announcement be equal to that unconditional expected value. Also, let us consider the mean-squared-error (MSE) as a measure of accuracy. Then, one can show that the conditional-expected-value announcement, as defined above, is at least as accurate as the static announcement i.e., it leads to a smaller or equal MSE. Of course, it is usually prohibitively difficult to calculate closed-form expressions for conditional expectations of waiting times. Further, such conditional expectations require knowledge of system parameters, which typically change over time; keeping track of such information may not be practical. Thus, alternative simplified delay-history-based announcements are typically used instead, e.g., specific waiting times of recently delayed customers (Armony et al. (2009) and Senderovich et al. (2015)).

It is not clear, a priori, whether and when such simplified announcements would be more accurate than a static announcement. On one hand, a static announcement may be inaccurate because it

[1] http://www.bayareahospital.org/Emergency-Department-Wait-Times.aspx

[2] http://appreviewtimes.com/

[3] http://www.cbsa-asfc.gc.ca/bwt-taf/

does not respond to changes in the underlying state of the system. That is, it may not be useful when there is large variation in delays. On the other hand, while simplified delay-history-based announcements do capture changes in system state, they may be inaccurate if that state changes too rapidly, e.g., during the waiting time of a delayed customer. Thus, it is conceivable that the state of the system is so noisy that it would be preferable to use a static prediction instead. Ultimately, a comparison between those two types of announcements involves studying the time scale of variation of the state of the system, relative to the magnitude of waiting times. To wit, the snapshot principle (Reiman 1984) explains why it may be better to use recent delay information than to eliminate noise by averaging delays, specifically in queueing contexts under heavy-traffic conditions; e.g., see Ibrahim and Whitt (2009). More generally however, characterizing the time scale of variation of the state of the system, in realistic systems with complex features, is usually a prohibitively difficult task.

We consider the Expected waiting time of delayed customers to be the static Announcement (EA), and the delay of the Last-customer-to-have-Entered-Service (LES), at the time of arrival of the current (delayed) customer, to be the simplified delay-history-based announcement, which is a dynamic announcement. The LES announcement is appealing: It is simple, easy to calculate, and robust, i.e., its implementation does not require knowledge of system parameters; see Ibrahim et al. (2016) for background.

1.2. The Role of Correlations

In this paper, we investigate the following questions: When does the LES announcement outperform the simple static announcement, EA? How do the LES announcement and EA perform in realistic settings, such as systems with multiple customer types and a priority service discipline? Importantly, how do the LES announcement and EA perform with real-life data, and can we design a new prediction that consistently outperforms both? By providing answers to those questions, we can shed light on when using a static announcement, as is done in some applications, would be justified.

To help develop a service science, it is necessary to systematically study various delay predictors in controlled environments, i.e., in structured mathematical models. Indeed, the standard approach, in the extant literature, has been to consider specific queueing models, which capture several realistic phenomena, and to study various ways of predicting delays in those models (Whitt (1999), Jouini et al. (2009) and Ibrahim (2010)). As such, the practical validity of various wait-time predictions remains intimately tied to the appropriateness of the technical conditions, e.g., specific distributional assumptions on underlying processes, under which they were studied. However, testing whether such technical modeling assumptions hold with real-life data is not a trivial task. Thus, there remains a need to develop a systematic and practically useful way to assess performances of delay predictors *across different queueing models*. Herein lies one of the main contributions of our work: In order to study the performances of the LES announcement and EA, we propose a new **"correlation-based" approach** which enables, in a simple way, a cross-model assessment.

1.3. Specific Contributions

Our overriding goal, in this paper, is to contrast the accuracy of static versus LES announcements. To do so, we rely on a three-fold methodology: (i) an empirical study of real-life data for an understanding of practical performance; (ii) a queueing-theoretic mode of analysis for the derivation of structural results and related insights; and (iii) a detailed simulation study for additional support.

Both our start and end points are empirical. We begin, in §3, by presenting statistical evidence, based on the analysis of a call-by-call data set from an Israeli bank's call center, that EA may indeed outperform the LES announcement in practice. In other words, exploiting real-time information about the current state of the system, via the delay of the last customer-to-have-entered-service, may not always be the right thing to do. Thus, there is a need for a deeper investigation. We develop a new framework to compare the accuracies of the LES announcement, or any other state-dependent (dynamic) announcement, and EA. Specifically, we show that the correlation between LES announcements and corresponding customer delays plays a key role. (Recall that any dynamic announcement is, itself, a random variable.) In particular, we derive the following result: *Under the MSE criterion for accuracy, an unbiased dynamic prediction, such as the LES announcement, is less accurate than the EA prediction if, and only if, the correlation between the dynamic announcements and corresponding customer waiting times is less than 1/2.* This result applies broadly, irrespective of the specific queueing model at hand, provided that some mild conditions are satisfied. For the LES announcement, we demonstrate that those conditions are indeed satisfied in several queueing models, provided that the system is sufficiently large. Because estimating correlations is much easier than assessing the goodness-of-fit of a particular queueing model, our approach is a practically useful way to determine when to rely on static versus dynamic announcements for accurate wait-time prediction.

For emphasis, here is our key message. We propose a simple way to compare the performances of alternative delay predictors *across different queueing contexts*. This is important because it is well known that the performances of alternative delay predictors are intimately tied to the specific queueing model considered; see Ibrahim (2017). Indeed, our theoretical queueing analysis, in this paper, is grounded in specific queueing contexts. We derive analytical expressions for correlations (between announcements and delays) in different queueing models, and rely on our general result above to gain insight into performance in each setting. The validation of our theoretical results with real-life data reduces to estimating correlations, which is easy to do.

Since doing direct analysis is prohibitively difficult, we rely on a many-server queueing-theoretic mode of analysis to derive asymptotic expressions for correlations, which yield useful approximations. Specifically, we consider Markovian queueing models, with or without abandonment, under both the first-come-first-served (FCFS) and non-preemptive priority service disciplines. To the best of our knowledge, there do not exist theoretical results quantifying the performance of the LES announcement in systems with a priority service discipline, despite the prevalence of such systems in practice. Indeed, the extant literature focuses solely on single-class systems under FCFS (Armony et al. 2009, Ibrahim and Whitt 2009, Ibrahim et al. 2016). State-dependent delay prediction with a priority service discipline is hard primarily because the waiting times of lower-priority customers depend on future higher-priority arrivals to the system. It is well known that analyzing systems which exhibit such dependencies is typically a difficult task; see Ward and Whitt (2000), Bitran and Caldentey (2002), and Armony and Maglaras (2004) for discussion.

The correlation expressions that we obtain have a remarkably simple form: They depend solely on the traffic intensities (ratio between arrival rate and total service capacity) of the alternative customer classes. By relying on those simple expressions, we are able to characterize how the performance of the LES announcement depends, in a non-trivial way, on the interplay between three factors: (i) the congestion level in the system, (ii) customer impatience, and (iii) the service discipline at hand. As such, we provide further theoretical support, and a different interpretation, to existing results for single-class Markovian models, and extend the scope of those results to multi-class systems as well.

We develop our analysis further, and propose a new Weighted Average (WA) prediction, which provably outperforms both the LES announcement and EA. We also characterise, analytically, the improvement in prediction accuracy, in using WA, over both the LES and EA announcements. To check the robustness of our theoretical results, we conduct a detailed simulation study which investigates the performance of the LES and EA announcements, and WA predictions, under general distributional assumptions and with time-varying arrival rates. In all cases considered, WA has a superior performance over both the LES announcement and EA. Finally, we apply our results to the data from two call centers, one which is of medium size and one which is large. In both settings, we show that WA does, indeed, consistently outperform both the LES and EA announcements.

While we mainly focus on EA and the LES announcement in this paper, we do not claim that these are the only announcements that are worth considering. For example, announcements exploiting the queue-length seen upon arrival by delayed customers are natural candidates that have been studied extensively in the literature, e.g., see Ibrahim (2010). The main takeaway from that line of research is that, while queue-length-based predictions are generally superior when the queue length is known and the rate at which customers enter service can be estimated reliably (so that the expected waiting time conditional on the queue-length can be approximated), those predictions may also fail spectacularly when incorrect assumptions are made about the underlying model. Thus, delay-history-based predictions, such as LES, may be preferred and should be studied as well. In this paper, we focus on those announcements. For completeness, we also consider a data-based-queue-length announcement in the e-companion (§EC.4.4), where we estimate the rate at which customers enter service directly from data; we show that this queue-length-based announcement is competitive but does not outperform WA, defined above.

1.4. Organization

Here is how the rest of this paper is organized. In §2, we explain how our work fits into the literature. In §3, we present preliminary results from our empirical study. In §4, we introduce our correlation-based framework. In §5, we derive asymptotic results for correlations in many-server queueing models. In §6, we describe results from a simulation study for robustness checks. Then, we return to our real-life data, and test the performance of the WA, LES, and EA announcements empirically. Finally, we analyze results from an additional real-life data set. We draw conclusions in §7. We present proofs to some technical lemmas in the e-companion, and additional numerical results in an online supplement to the main paper.

2. How Does This Paper Fit into the Literature?

This paper is most closely related to the literature on delay announcements in queueing systems. In broad terms, this literature can be classified into four main categories depending on how customers are assumed to treat the delay information: (i) customers are psychologically affected by the lengths of their waiting times and the delay announcements that they receive; (ii) customers are strategic, forward-looking, utility-maximizing decision makers who respond to delay announcements and decide, accordingly, whether to balk or join the system; (iii) customers are not decision makers, yet their behaviors are exogenously affected by the announcements; and (iv) customers are queued entities who do not respond to the announcements. For a recent survey of the papers in each category, we refer the reader to Ibrahim (2017).

This paper falls into that fourth stream of literature. Specifically, we focus on the question of accuracy of the announcements and assume that customers do not respond to the delay information. The asymptotic accuracy of the LES prediction in large queueing systems is justified by supporting heavy-traffic limit theorems, see Armony et al. (2009), Ibrahim and Whitt (2009), and Ibrahim et al. (2016), albeit in the context of heavily-congested Markovian queueing systems, in steady state, with a single customer class and a FCFS service discipline. The LES prediction has also been shown to be systematically biased, leading to inaccurate announcements, with time-varying arrival rates which cause systematic variations in delays over time (Ibrahim and Whitt 2011). The study of delay prediction in multi-class settings (but not of the LES announcement) has been considered in both Nakibly (2002) and Jouini et al. (2009). However, our asymptotic mode of analysis and our focus on the performance of delay-history-based predictions, such as the LES prediction, in multiclass settings, are different.

For a numerical study of the performance of LES and several LES-based predictions in multiclass settings, see Thiongane et al. (2015). We present theoretical support to the numerical observations in that paper. For an empirical investigation of the (superior) performance of the LES announcement with real-life data, see Senderovich et al. (2014 and 2015). In contrast, our data-based study focuses here on settings where the LES announcement performs poorly relative to EA. Thus, the conclusions that we reach are different. The inaccuracy of delay-history-based predictions, such as the LES announcement, in certain settings has also been acknowledged in Yu et al. (2017), albeit in the context of a field experiment to assess whether customers are loss averse in time, and how the delay announcement made may impact such reference-dependent behavior.

Our proposed WA predictor, which is a weighted sum of the LES announcement and EA, is in the *same spirit* as delay announcements which combine recent delay-history-based information along with average wait-time predictions, as in Gal et al. (2015) and Ang et al. (2015). However, our focus here is on analytically characterising the performance of the announcements, in addition to empirically investigating that performance, whereas the focus in those papers is on combining existing queueing-theoretic results and data-mining techniques to achieve superior predictive power in more complicated settings (urban transportation and hospital emergency departments, respectively). There are very few papers which consider an approach similar to ours, i.e., one which combines mathematical analysis to investigate the accuracy of delay announcements with an empirical validation of those analytical results. For example, Jouini et al. (2015) considers delay prediction in a multi-class queue with a priority service discipline and time-varying arrival rates. However, the focus of our paper, our queueing-asymptotic methodology, our consideration of the LES predictor, and our main results, are all different from theirs. There are also recent papers which adopt a structural estimation approach to modeled customer decisions with delay announcements, e.g., see Akşin et al. (2016) and Yu et al. (2017).

There are several papers which study the impact of using real-time versus static delay information, in problem settings which are different from ours. Armony and Maglaras (2004) studies joint routing and delay-announcement decisions in the context of a call center which offers a call-back option to delayed customers. There, the authors show that state-dependent information increases resource utilization while improving the quality of service for real-time service. Singh et al. (2017)

considers a competitive environment with two service providers. They investigate, from the viewpoint of a service provider who is competing for market share, the question of whether or not to announce real-time delay information. Dong et al. (2017) studies the impact of delay announcements on coordination in a network of hospitals. While the main focus of that paper is empirical, the authors also describe simulation results which illustrate hat using average-wait predictions may lead to asynchronous behavior in the system; this numerical observation is subsequently investigated analytically in Pender et al. (2017), using an approximating fluid model.

In this paper, we supplement the literature by proposing a new correlation-based framework which allows for a broader understanding of performance across different queueing models. Through our framework, we are able to provide a theoretical justification to earlier numerical and empirical observations, e.g., in Thiongane et al. (2015) and Senderovich et al. (2014 and 2015). Importantly, our framework can be useful in practice because estimating correlations is easier than fitting queueing models to data.

3. Preliminary Empirical Results

From a practical and empirical standpoint, there is some empirical evidence substantiating the good performance of the LES announcement with real-life data in some cases; e.g., see Senderovich et al. (2014), Senderovich et al. (2015), and Gal et al. (2017). In contrast, we begin here by presenting conflicting empirical evidence which illustrates the *poor* accuracy of the LES announcement in other cases, relative to the static announcement EA; we will return to this empirical evidence in §6.2.1. For now, we illustrate that the LES announcement may indeed perform worse than EA. We will formally define EA as well as LES in §4.1.

3.1. Definitions: EA and LES Announcements

We denote by W_∞ a random variable with the distribution of the steady-state virtual waiting time, which is the waiting time experienced by an infinitely patient customer. The static announcement, EA, is then given by $\mathbb{E}[W_\infty | W_\infty > 0]$. In our simulations, we use a point estimate of $\mathbb{E}[W_\infty | W_\infty > 0]$ for the EA announcement. In particular, we begin by discarding an initial transient period from our simulation to ensure steady-state conditions. Then, we use the running average of virtual waiting times (for delayed customers) as a point estimate of $\mathbb{E}[W_\infty | W_\infty > 0]$. In our empirical study, we approximate the EA announcement by the average waiting time until abandonment or service, as described in §3.3.

For the LES announcement, we use the delay of the last customer to have entered service; if there are multiple classes, then we use the delay of the LES customer from the same class as the delayed customer to whom the announcement is made. We now formally define the LES announcement. We let t denote the arrival time of a delayed customer, to whom a delay announcement is made, in steady state. We denote the virtual waiting time of this new customer by $W(t)$, and note that $W(t)$ has the same distribution as $W_D \equiv [W_\infty | W_\infty > 0]$. Let τ_t denote the arrival time of the corresponding LES customer, whose delay was used in the announcement, i.e.,

$$\tau_t = \sup\{s \le t : \text{ There is an arrival at time } s \text{ and } s + W(s) \le t\}. \tag{1}$$

Then, $W(\tau_t)$ is equal to the LES prediction.

To quantify the accuracy of the LES announcement and EA in our simulation experiments, we use the *average-squared-error* (ASE):

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

where $a_j > 0$ is the actual virtual delay of customer j, p_j is his predicted delay, and k is the number of customers in our sample. The ASE is a point estimate of the MSE which is defined as the expected value of the square of the difference between the delay prediction and the virtual delay.

For a relative measure of performance, we calculate the percent *relative-root-average-squared-error* (RASE), which is defined as the ratio between the square root of the ASE and the average waiting time in the queue, in percent. The RASE is useful because it relates the errors in the LES and EA announcements to the magnitude of experienced waiting times in the system.

3.2. Description of the Data

General setting. We consider the Israeli call-center data set which was analyzed in Brown et al. (2005)[4]. The call center is small, consisting of at most 13 regular agents, and the call-by-call data spans all 12 months of 1999. In each month, roughly 100,000-120,000 calls are made and 65,000-85,000 of those calls terminated in the voice response unit (VRU). The remaining 30,000-40,000 calls per month were either served by an agent (80%) or abandoned before service (20%).

There are four main call types: Regular (PS), stock transaction (NE), new/potential customer (NW), and internet assistance (IN)[5]. In what follows, we focus on the IN and PS call types; we describe results for the remaining types in §EC.4.3 of the e-companion.

In the months of January to July, all calls were served by the same group of agents. However, in the months of August to December, IN customers were served by a separate pool of agents. Thus, in those later months, the call center effectively consists of two separate systems: A single-class, single-priority, system for IN customers, and a multi-class two-priority system for all remaining types. In particular, for PS callers, there are two priority levels, high and low. High-priority customers are

[4] This data set is publicly available at: http://ie.technion.ac.il/serveng/callcenterdata/index.html.

[5] For the different call types, we use the acronyms that were considered in Brown et al. (2005).

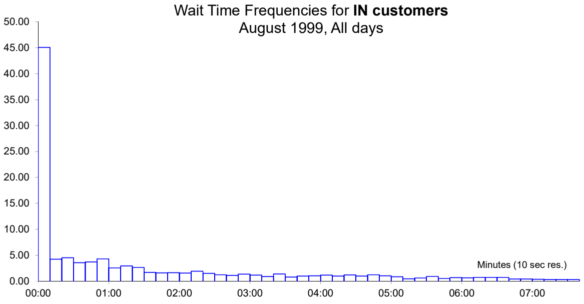

Figure 1 **Frequencies of waiting times for single-class IN customers for August 1999.**

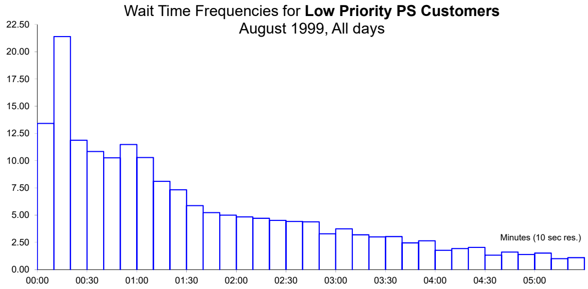

Figure 2 **Frequencies of waiting times for low-priority PS customers for August 1999.**

moved up the queue by subtracting 1.5 minutes from their arrival times. For both the IN and PS types, we restrict attention to the months of August to December, because we are interested in studying the performances of our announcements both with and without priorities. For PS calls, which have a much greater volume than other classes, we focus on low-priority callers. There are roughly 5,200 delayed IN callers and 16,000 delayed low-priority PS callers in our sample.

Customers receive delay announcements every 60 seconds, and they react to those announcements by adjusting their patience levels accordingly; see Figure 5 in Brown et al. (2005). We focus on regular weekdays (Sunday - Thursday)[6] and regular working hours (7am - midnight). Arrival rates vary with time, and the lognormal distribution, with a time-dependent and a class-dependent mean, is a good fit for the service-time distribution; see Figures 1 and 2 in Brown et al. (2005).

Waiting times. In Figures 1 and 2, we plot the wait-time frequencies (until either abandonment or service) for IN and PS callers during the month of August 1999, which has the most delayed callers (1,822 delayed IN callers, and 4,582 delayed low-priority PS callers). Figure 1 shows that most IN callers experience a short waiting time: 45% of the callers wait less than 10 seconds. In Table 1, we present summary statistics for the waiting times of IN and PS callers during that same month, August 1999. Table 1 and Figure 1 show that, while many IN customers do not wait a long time, the average waiting time remains large because some callers experience very long waits.

3.3. Comparing Static EA and the Dynamic LES Announcement

For each delayed caller, we identify the corresponding LES delay from data, and calculate the square of the corresponding prediction error, i.e., the square of the difference between the LES

[6] Sunday to Thursday are regular work days in Israel.

Performance measures	**IN**	**PS**
Expected wait time $\mathbb{E}[W]$	135s	47s
Expected wait time conditional on positive wait $\mathbb{E}[W \mid W > 0]$	185s	113s
Probability of delay $\mathbb{P}(W > 0)$	73%	41%

Table 1 Point estimates of summary statistics for the waiting times (in seconds), of single-class IN and low-priority PS callers (August 1999).

delay and the actual waiting time of the caller, until either service or abandonment. For the EA announcement, we calculate an out-of-sample estimate of the average waiting time based on the first 2,000 delayed callers, beginning August 1, which are then discarded from the sample. We note that our estimate for EA, in the main paper, does not account for seasonality effects, e.g., day-of-week effects. Including such effects, i.e., making different EA announcements depending on the day of week, should generally lead to more accurate predictions; we consider such day-of-week-adjusted announcements in the e-companion (§EC.4.2). For each day in our sample, we average the errors over all (served and abandoning) delayed customers during that day.

In Figures 3 and 4, we plot running root-mean squared errors corresponding to the EA (solid) and LES (dashed) announcements, across successive days. Each running average point estimate is based on averaging squared errors, between actual delays and corresponding announcements, in a centered window of length 10 days. Clearly, Figures 3 and 4 illustrate that the LES announcement performs generally worse than EA in this setting. In other words, existing theoretical results which substantiate the good performance of the LES announcement in large heavily-loaded stationary systems do not describe the given system well. Based on these observations, we see that exploiting recent information about the state of the system, as for the LES announcement, may lead, under certain conditions, to worse performance than EA. Thus, there is a need to investigate further; we devote the remainder of this paper to this investigation.

4. A Correlation-Based Assessment

In this section, we bring out the role of correlations in understanding the accuracy of the announcements. We begin this section with a numerical study. Specifically, we describe results from simulation experiments which quantify the respective performances of the LES announcement and EA in various queueing models. Our numerical results paint a complex picture, with seemingly contradictory results. In particular, they illustrate that the LES announcement may be more or

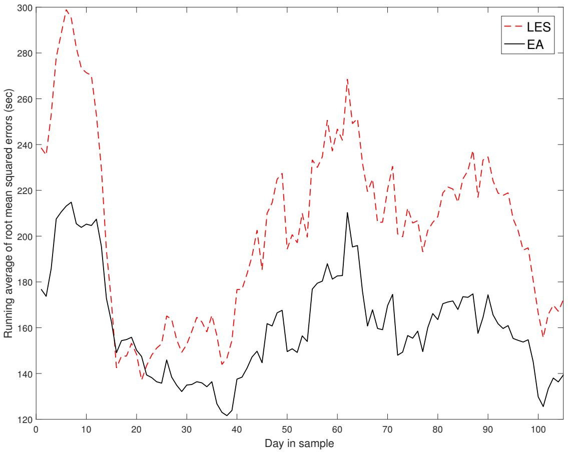

Figure 3 **IN customers (August-December): System is a single-class single-priority queue.**

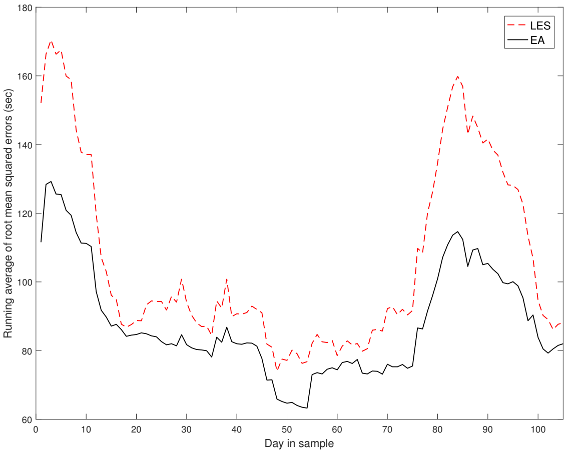

Figure 4 **Low-priority PS customers (August-December).**

less accurate than EA, depending, in a non-trivial manner, on the interplay between a host of factors such as: (i) the congestion level, (ii) customer impatience, and (iii) the service discipline in the system. In the remainder of this section, we develop a correlation-based framework for the performances of the LES announcement and EA. We do so in order to derive general structural insights which hold broadly, based on our numerical results. Then, we introduce our new Weighted Announcement, WA, and quantify its performance.

4.1. Numerical Study

We consider a variety of queueing systems. 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. We want to quantify performance in steady state. To this end, we exclude from each simulation run the first 5,000 events so as to remove the effect of the initial transient period. Although we will later support our numerical findings by using an asymptotic, many-server, mode of analysis, we deliberately consider a relatively small number of servers, $n = 30$, in our numerical study. This will allow us to show that our asymptotic results are also useful in describing performance in small to medium systems. Our simulation results, throughout this paper, are based on 10 independent replications of 2 million arrival events each.

In what follows, we consider single-class and two-class queueing systems, both with patient customers (i.e., systems without abandonment), and with impatient customers (i.e., systems with abandonment). We do so in order to gain a broad understanding of the performance of our alternative announcements across different queueing models. Later, our correlation-based framework will allow us to unify our observations across different models. We begin by describing our numerical results for single-class systems. Then, we turn to two-class queueing systems.

Single-class queueing systems. We assume that customers arrive according to a Poisson process with rate λ. We assume that there are n homogeneous servers working in parallel. We let service times be independent and identically distributed (i.i.d.) exponential random variables with rate μ; without loss of generality, we let $\mu = 1$. We consider systems both with patient and with impatient customers. In systems with impatient customers, we let the times to abandon be i.i.d. exponential random variables with rate θ. The abandonment, service, and arrival processes are mutually independent. The traffic intensity, ρ, is given by $\rho \equiv \lambda/n\mu$. Without abandonment, we assume that $\rho < 1$ so that a proper limiting steady state exists; otherwise, abandonment ensures the stability of the system, irrespective of the value of ρ. We also assume that there is unlimited waiting space. In terms of service discipline, we consider FCFS.

We let $\mu = 1$ and $\theta = 0.5$, and vary the value of λ to vary ρ. In Figures 5 and 6, we present plots of the RASE's of EA and the LES announcement, as a function of ρ, in a single-class $M/M/30$ (Figure 5) and $M/M/30 + M$ (Figure 6) models, under FCFS. Figure 5 shows that, in the absence of abandonment, the LES prediction is less accurate than the EA prediction for small values of ρ, but not otherwise. For the $M/M/n$ model, it is well known that the steady-state waiting time, conditional on the wait being positive, has an exponential distribution with mean $1/n(1-\rho)$. Thus, the RASE of the EA predictor, which is an estimate of the coefficient of variation of an exponential distribution, should be close to 1; this is consistent with Figure 5. In contrast, Figure 6 shows that incorporating customer abandonment *reverses* the result: While the LES announcement is more accurate than EA for small values of ρ, it is marginally less accurate otherwise. In §5, we will present a simple explanation for this puzzling change in performance.

Two-class queueing systems. We consider a non-preemptive priority service discipline. We use subscript H and L to denote the high and low priority classes, respectively. With two classes, we assume identical service and abandonment rates for H and L: $\mu_H = \mu_L = 1$ and $\theta_H = \theta_L = 0.5$. We fix $\rho_H \equiv \lambda_H/n\mu_H = 0.5$, and vary $\rho_L \equiv \lambda_L/n\mu_L$ by varying λ_L; increasing λ_L amounts to increasing the overall congestion level in the system. We consider stable systems, e.g., we assume that $\rho_H + \rho_L < 1$ where there is no abandonment. Both classes are served by the same pool of agents.

In Figures 7 and 8, we present corresponding results for low-priority customers in a two-class queueing model with non-preemptive priority, with and without customer abandonment, respectively. We focus on low-priority customers only, because high-priority customers are approximately unaffected by low-priority customers, particularly when the number of servers is not too small. In other words, high-priority customers roughly "see" a single-class FCFS system, as in Figures 5 and 6. Our simulation results are also in line with this observation. In Figures 7 and 8, we plot the RASE's for low-priority customers, with patient and with impatient customers, as a function of ρ_L/ρ_H, which measures the relative congestion due to L arrivals; recalling that $\rho_H = 0.5$ is held

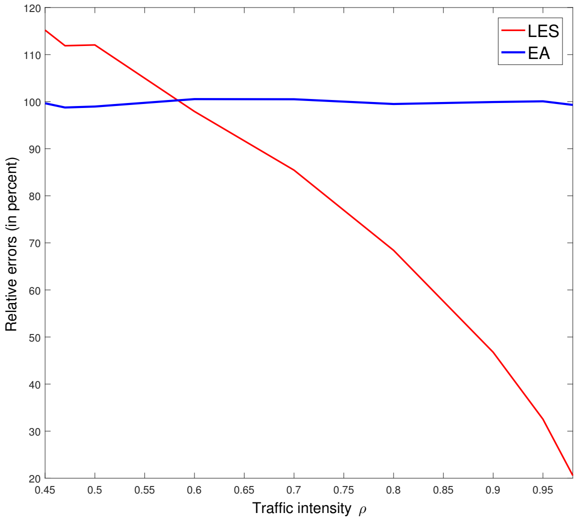

Figure 5 **RASE of LES and EA in the $M/M/30$ model, without abandonment.**

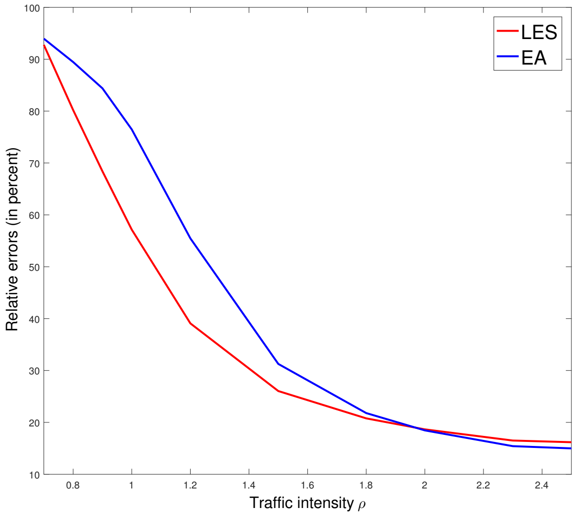

Figure 6 **RASE of LES and EA in the $M/M/30+M$ model, with abandonment.**

constant, the higher ρ_L/ρ_H, the more congested the system. Figure 7 shows that, in the absence of abandonment, the relative performance of the LES announcement and EA is similar to Figure 5. However, Figure 8 illustrates that, due to incorporating abandonment, the LES announcement is marginally less accurate than EA for low and high values of ρ_L/ρ_H, but is marginally more accurate than EA for intermediate values of ρ_L/ρ_H.

In short, Figures 5-8 paint a complex picture: The performances of the LES announcement and EA are strongly tied to the model at hand, and may vary considerably depending on several characteristics of the system. Next, we propose a framework to “unify” our seemingly contradictory numerical observations.

4.2. The Role of Correlations

For ease of exposition, we let $W_D \equiv [W_\infty | W_\infty > 0]$ represent a random variable with the distribution of the steady-state virtual waiting time, conditional on the wait being positive. Thus, the virtual waiting time of a new delayed customer, to whom a delay prediction is made, is distributed as W_D. Let P represent a random variable with the distribution of a given delay prediction in steady state. To illustrate, for the LES prediction, P has the distribution of the steady-state waiting time of a served customer, conditional on the event that the next arrival, after entry of the LES customer to service, is delayed and no other customers have entered service before the new arrival epoch. The MSE corresponding to delay predictor P is given by:

$$\mathrm{MSE}(P) = \mathbb{E}[(W_D - P)^2]. \tag{3}$$

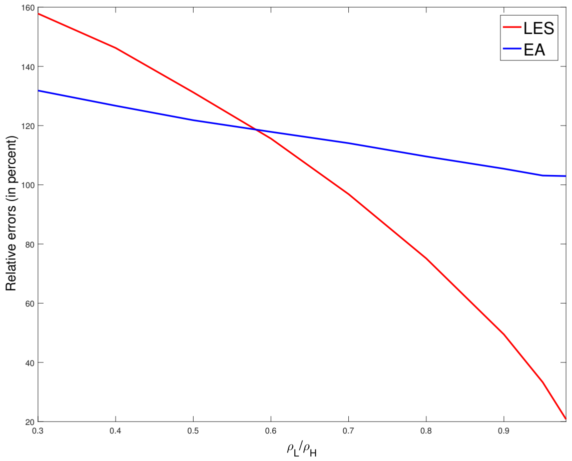

Figure 7 **RASE of LES and EA for low-priority customers in the two-class $M/M/30$ model, with a non-preemptive priority discipline.**

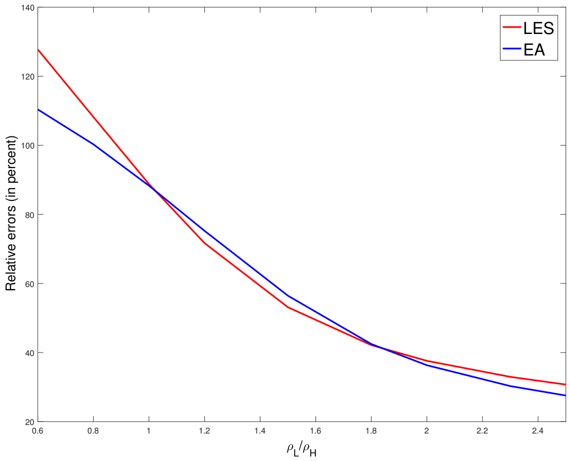

Figure 8 **RASE of LES and EA for low-priority customers in the two-class $M/M/30+M$ model, with a non-preemptive priority discipline.**

Recall that the EA prediction is equal to $\mathbb{E}[W_D]$, so that $\text{MSE(EA)} = \mathbb{E}[(W_D - \mathbb{E}[W_D])^2] = \text{Var}[W_D]$. We let $C_{W_D} \equiv \sqrt{\text{Var}[W_D]}/\mathbb{E}[W_D]^2$ denote the coefficient of variation of W_D. The following proposition, which is straightforward to establish, holds.

Proposition 1. *Let $r[P, W_D]$ denote the correlation between P and W_D, under steady-state conditions. If $\mathbb{E}[P] = \gamma\mathbb{E}[W_D]$ and $Var[P] = \beta\, Var[W_D]$, where $\gamma \geq 0$ and $\beta \geq 0$, then:*

$$MSE(P) = Var[W_D]\left(\beta + 1 + \left(\frac{\gamma - 1}{C_{W_D}}\right)^2 - 2 \cdot r[P, W_D]\sqrt{\beta}\right). \tag{4}$$

Proof. We have:

$$\begin{aligned} MSE(P) \equiv \mathbb{E}[(P - W_D)^2] &= \mathbb{E}[(P - \mathbb{E}[P] + \mathbb{E}[P] - \mathbb{E}[W_D] + \mathbb{E}[W_D] - W_D)^2] \\ &= \text{Var}[P] + \text{Var}[W_D] + (\gamma - 1)^2(\mathbb{E}[W_D])^2 - 2 \cdot r[P, W_D]\sqrt{\beta} \cdot \text{Var}[W_D] \\ &= \text{Var}[W_D]\left(\beta + 1 + \left(\frac{\gamma - 1}{C_{W_D}}\right)^2 - 2 \cdot r[P, W_D]\sqrt{\beta}\right). \end{aligned}$$

■

Using Proposition 1, and the fact that $\mathbb{E}[EA] = \mathbb{E}[W_D]$ and $\text{Var}[EA] = 0$, we get the following corollary.

Corollary 1. *If $\mathbb{E}[P] = \mathbb{E}[W_D]$ and $Var[P] = Var[W_D]$, i.e., $\beta = \gamma = 1$, then:*

$$MSE(EA) \leq MSE(P) \quad \textit{if, and only if,} \quad r[P, W_D] \leq 1/2. \tag{5}$$

Despite its simplicity, Corollary 1 is powerful because it allows for a simple check of the relative performance of any dynamic prediction and EA: Assessing whether or not such a prediction is more accurate than EA reduces to calculating correlations in the system at hand. We prove in §5 that the condition in the corollary is satisfied, asymptotically in large systems, for the LES announcement.

4.3. A New Weighted Average Predictor: WA

To go further, we now propose a new predictor which provably outperforms both P and EA. Let a new Weighted Average predictor, WA(P), be defined as a weighted average between P and EA. Specifically, for a prediction P as defined in Corollary 1 and a scalar α, let:

$$\mathrm{WA}(P) \equiv \alpha \mathbb{E}[W_D] + (1-\alpha)P. \tag{6}$$

Recalling our assumption that $\mathbb{E}[P] = \mathbb{E}[W_D]$, we choose the form in (6) to guarantee that WA(P) is unbiased as well, i.e., $\mathbb{E}[\mathrm{WA}(P)] = \mathbb{E}[W_D] = \mathbb{E}[P]$. We also ignore, for now, the non-negativity restriction on WA(P) in (6). We can then calculate the MSE-minimizing α:

$$\begin{aligned}
\mathrm{MSE}(\mathrm{WA}(P)) &\equiv \mathbb{E}[(\alpha\mathbb{E}[W_D] + (1-\alpha)P - W_D)^2],\\
&= \mathbb{E}[(\alpha\mathbb{E}[W_D] + (1-\alpha)P + (1-\alpha)\mathbb{E}[P] - (1-\alpha)\mathbb{E}[P] + (1-\alpha)P - W_D)^2],\\
&= \mathbb{E}[((\mathbb{E}[W_D] - W_D) + (1-\alpha)(P - \mathbb{E}[P]))^2],\\
&= [\alpha^2 - 2(1 - r[P,W_D])\cdot\alpha + 2(1 - r[P,W_D])]\cdot \mathrm{Var}[W_D],
\end{aligned}$$

which is minimized at $\alpha^* = 1 - r[P,W_D]$. We define the MSE-minimizing, WA*(P), prediction, which corresponds to the MSE-minimizing α^*, as follows:

$$\mathrm{WA}^*(P) \equiv (1-\alpha^*)\cdot\mathbb{E}[W_D] + \alpha^*\cdot P = (1 - r[P,W_D])\cdot\mathbb{E}[W_D] + r[P,W_D]\cdot P. \tag{7}$$

We note that if $0 \le r[P,W_D] \le 1$, which should hold for any reasonable predictor P, then the WA*(P) prediction, as defined in (7), will be nonnegative. For the LES announcement, we prove in §5 that this is indeed the case for large systems. We directly deduce that:

$$\frac{\mathrm{MSE}(\mathrm{WA}^*(P))}{\mathrm{MSE}(\mathrm{EA})} = 1 - r[P,W_D]^2 \quad \text{and} \quad \frac{\mathrm{MSE}(\mathrm{WA}^*(P))}{\mathrm{MSE}(P)} = \frac{1 + r[P,W_D]}{2}. \tag{8}$$

Based on (8), we see that, relative to EA, WA*(P) is increasingly more accurate as $r[P,W_D]$ increases (more weight is assigned to P in (6)). Conversely, relative to P, WA*(P) is increasingly more accurate as $r[P,W_D]$ decreases (more weight is assigned to EA in (6)). Hereafter, for ease of notation, we use WA to denote the MSE-minimizing weighted-average announcement where P coincides with the LES prediction. To be able to use the correlation-based framework of this section, we need to compute $r[\mathrm{LES},W_D]$. Because direct analysis is prohibitively difficult, we focus, in the next section, on establishing many-server heavy-traffic limits for $r[\mathrm{LES},W_D]$ instead.

5. Asymptotic Analysis

In this section, we derive many-server heavy-traffic limits for $r[\text{LES}, W_D]$ in single-class Markovian systems first with patient customers (§5.1), and then with impatient customers (§5.2). Then, we consider two-class systems, under the non-preemptive priority service discipline, also first without (§5.3), and then with (§5.4) abandonment. Following each proposition and corresponding proof, we provide, in turn, explanations to our simulation-based observations of §4.1. We relegate the proofs of several technical lemmas to the e-companion.

5.1. Single Class with Patient Customers

We consider a sequence of queueing models indexed by the number of servers, n. We let λ^n denote the arrival rate in system n, and let $\lambda^n \to \infty$ as $n \uparrow \infty$. The service rate in the n^{th} system is held fixed. We let $\rho^n \equiv \frac{\lambda^n}{n\mu} < 1$ for every n, and assume that $\rho^n \to \rho$ as $n \to \infty$, where $\rho \le 1$. For $\rho = 1$, we consider the Quality-and-Efficiency Driven (QED) or Halfin-Whitt regime (Halfin and Whitt 1981), which is defined by letting the sequence of arrival rates $\{\lambda^n, n \ge 1\}$ satisfy

$$\lim_{n\to\infty} \sqrt{n}\left(1 - \frac{\lambda^n}{n\mu}\right) = \beta \quad \text{where} \quad \beta \ge 0\ . \tag{9}$$

We begin by recalling the formal definition of the LES announcement. Let t^n denote the arrival time of a delayed customer, to whom a delay announcement is made, in steady state. The virtual waiting time of this new customer is $W^n(t^n)$, which has the same distribution as $W_D^n \equiv [W_\infty^n | W_\infty^n > 0]$. Let τ_t^n denote the arrival time of the corresponding LES customer, whose delay was used in the announcement, i.e.,

$$\tau_t^n = \sup\{s \le t^n : \text{ There is an arrival at time } s \text{ and } s + W^n(s) \le t^n\}. \tag{10}$$

Then, $W^n(\tau_t^n)$ is equal to the LES prediction. We let $\{N(s) : s \ge 0\}$ denote the Poisson arrival counting process, i.e., $N(s)$ is the number of arrivals before or at s. Then, the following holds.

Proposition 2. *In the $M/M/n$ model as either (i) $\rho^n = \frac{\lambda^n}{n\mu} \to \rho < 1$, or (ii) $\rho^n \to 1$ in the QED many-server heavy-traffic regime, as given by (9), we have:*

$$r[W^n(\tau_t^n), W^n(t^n)] \to \rho \quad \textit{as} \quad n \to \infty. \tag{11}$$

Proof. For the proof of the proposition, we need the following Lemma 1 which we prove in the e-companion.

Lemma 1. *For all $x \ge 0$,*

$$\lim_{n\to\infty} \mathbb{P}(W^n(\tau_t^n) \ge x) = \lim_{n\to\infty} \mathbb{P}(W^n(t^n) \ge x).$$

Moreover, $\lim_{n\to\infty} \mathit{Var}[W^n(\tau_t^n)] = \lim_{n\to\infty} \mathit{Var}[W^n(t^n)]$.

We let γ_t^n denote the time of entry of the LES customer to service, and let $\xi_t^n \equiv t^n - \gamma_t^n$ denote the time between the entry to service of the LES customer and the new arrival time. We can write

$$W^n(t^n) = \sum_{i=1}^{N(t^n-\tau_t^n)} X_i = \sum_{i=1}^{N(W^n(\tau_t^n)+\xi_t^n)} X_i = \sum_{i=1}^{N(W^n(\tau_t^n))} X_i + \sum_{i=1}^{N(\xi_t^n)-1} X_i = \sum_{i=1}^{N(W^n(\tau_t^n))} X_i + Y^n,$$

where $X_i \sim \text{Exp}(n\mu)$ and Y^n is independent of $W^n(\tau_t^n)$. This is so because γ_t^n is a stopping time for the Poisson arrival process; by the strong Markov property and the memoryless property for the exponential service times, we have independence. Thus, letting $\text{Cov}[X,Y]$ denote the covariance between two random variables X and Y:

$$\begin{aligned}\text{Cov}\left[W^n(\tau_t^n), W^n(t^n)\right] &= \text{Cov}\left[\sum_{i=1}^{N(W^n(\tau_t^n))} X_i, W^n(\tau_t^n)\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{N(W^n(\tau_t^n))} X_i\right)\cdot W^n(\tau_t^n)\right] - \mathbb{E}\left[\left(\sum_{i=1}^{N(W^n(\tau_t^n))} X_i\right)\right]\cdot \mathbb{E}\left[W^n(\tau_t^n)\right] \\ &= \rho^n\mathbb{E}[(W^n(\tau_t^n))^2] - \rho^n(\mathbb{E}[W^n(\tau_t^n)])^2 = \rho^n\text{Var}[W^n(\tau_t^n)].\end{aligned}$$

Therefore, we deduce that:

$$r[W^n(\tau_t^n), W^n(t^n)] = \frac{\rho^n\text{Var}[W^n(\tau_t^n)]}{\sqrt{\text{Var}[W^n(\tau_t^n)]\text{Var}[W^n(t^n)]}} \to \rho \text{ as } n\to\infty.$$

■

Combining the results in Propositions 1 and 2 allows for a simple explanation of our numerical observations in Figure 5. In large systems, based on (11), if $\rho \le 0.5$, then the LES announcement is less accurate than EA, and if $\rho > 0.5$, then the LES announcement is more accurate than EA. In other words, the LES announcement is more accurate than EA when the system is under moderate to heavy congestion, but not otherwise. It is important to note that although the number of servers in Figure 5 is relatively small, our results remain useful in roughly describing performance.

To check the robustness of our theoretical results in Proposition 2, we present in Table 2 simulation point estimates of correlations in the $M/G/100$ model, where service times are allowed to follow three distributions: exponential (M), lognormal (LN with mean and variance both equal to 1), and Erlang (E_2, with mean 1) distributions for service times. We consider the LN distribution because there is empirical evidence suggesting a good fit to this distribution in practice (Brown et al. 2005). Table 2 illustrates that our asymptotic results in Proposition 2 are useful in describing performance in systems with a variety of service-time distributions, particularly when the traffic intensity is not too small (e.g., larger than 0.85), and the service-time distribution has moderate to high variability; indeed, the approximation in Proposition 2 is more accurate for LN than E_2 service times.

ρ	$M/M/100$	$M/LN(1,1)/100$	$M/E_2/100$
0.7	0.712 $\pm 1.30 \times 10^{-3}$	0.687 $\pm 4.8 \times 10^{-4}$	0.704 ± 0.015
0.75	0.726 $\pm 2.92 \times 10^{-4}$	0.700 $\pm 9.0 \times 10^{-4}$	0.654 $\pm 4.0 \times 10^{-4}$
0.8	0.779 $\pm 4.6 \times 10^{-4}$	0.782 $\pm 3.7 \times 10^{-4}$	0.728 $\pm 2.4 \times 10^{-4}$
0.85	0.837 $\pm 7.7 \times 10^{-4}$	0.831 $\pm 8.4 \times 10^{-4}$	0.799 $\pm 9.9 \times 10^{-4}$
0.9	0.894 $\pm 1.7 \times 10^{-4}$	0.896 $\pm 3.3 \times 10^{-5}$	0.869 $\pm 5.6 \times 10^{-5}$
0.95	0.949 $\pm 2.6 \times 10^{-4}$	0.955 $\pm 1.3 \times 10^{-4}$	0.938 $\pm 5.0 \times 10^{-4}$
0.98	0.981 $\pm 1.5 \times 10^{-4}$	0.981 $\pm 1.2 \times 10^{-5}$	0.973 $\pm 5.6 \times 10^{-5}$

Table 2 Point estimates of correlations in the $M/G/100$ queueing model, for alternative values of the traffic intensity ρ.

5.2. Single Class with Impatient Customers

We now derive asymptotic results for the $M/M/n+M$ model. Before getting to those results, we present some intuition on how including impatient customers may change the results of §5.1. At a high level, abandonment on the part of customers when facing long delays adds "noise" to the system. Indeed, when a customer waits for a long time, there will be considerable abandonment, and the next customer is likely to have a shorter waiting time. Similarly, when a customer waits for a short time, there will be little abandonment, and the next customer is likely to have a long waiting time. This seems to suggest that, all else held constant, increasing the congestion level in the system should lead to a decrease in the correlation. Indeed, this is confirmed by the asymptotic results of this section. In §5.2.1, we consider the efficiency-driven (ED) regime, considered in Whitt (2004), where the arrival rate λ^n increases without bound while the traffic intensity $\rho^n = \lambda^n / n\mu = \rho > 1$ is held equal to a constant value. In §5.2.2, we consider the Quality-and-Efficiency Driven (QED) regime; see Garnett et al. (2002). In particular, we let:

$$\lim_{n\to\infty} \sqrt{n}\left(1 - \frac{\lambda^n}{n\mu}\right) = \beta \quad \text{where} \quad \beta \geq 0\ . \tag{12}$$

where $\beta < 0$ in (9) is now allowed because customer abandonment keeps the system stable in this case. We assume that $\theta^n \equiv \theta$ for all n, i.e., the abandonment rate remains constant as n increases.

5.2.1. The ED Regime. Here is our main result for the ED regime.

PROPOSITION 3. *For the $M/M/n+M$ model in the ED many-server heavy-traffic limiting regime, we have that:*

$$r[W^n(\tau_t^n), W^n(t^n)] \to \frac{1}{\rho} \quad as \quad n\to\infty. \tag{13}$$

Proof. For the proof of the proposition, we need the following lemma where $\mathrm{Nor}(\mu,\sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

LEMMA 2. *In the ED regime,*

(a) For any $w\geq 0$,

$$\lim_{n\to\infty}\mathbb{P}(W^n(\tau_t^n)\geq w)=\lim_{n\to\infty}\mathbb{P}(W_S^n\geq w),$$

where W_S^n is a random variable with the distribution of the virtual waiting time of a customer conditional on the customer being both delayed and served.

(b) As $n\to\infty$,

$$\sqrt{n}(W_S^n-\bar{w})\Rightarrow \mathrm{Nor}\left(0,\frac{1}{\theta}\right),$$

where $\bar{w}\equiv\frac{1}{\theta}\ln(\rho)$.

We are now ready to derive an asymptotic expression for the correlation. To this end, we write:

$$r[W^n(t^n), W^n(\tau_t^n)] = \frac{\mathrm{Cov}[W^n(t), W^n(\tau_t^n)]}{\sqrt{\mathrm{Var}[W^n(t)]\mathrm{Var}[W^n(\tau_t^n)]}}.$$

We have,

$$\begin{aligned}\mathrm{Cov}[W^n(t^n), W^n(\tau_t^n)] &= \frac{1}{2}\left(\mathrm{Var}[W^n(t)]+\mathrm{Var}[W^n(\tau_t^n)]-\mathrm{Var}[W^n(t)-W^n(\tau_t^n)]\right)\\ &= \frac{1}{2}\left(\mathrm{Var}[W^n(t)]+\mathrm{Var}[W^n(\tau_t^n)]\right)-\frac{1}{2n}\mathrm{Var}[\sqrt{n}(W^n(t)-W^n(\tau_t^n))].\end{aligned}$$

By Theorem 4 of Ibrahim and Whitt (2009), we have that

$$\mathrm{Var}[\sqrt{n}(W^n(t^n)-W^n(\tau_t^n))]\to\frac{2(\rho-1)}{\rho\theta} \text{ as } n\to\infty.$$

Also, we have that that $n\mathrm{Var}[W^n(t^n)]\to\frac{1}{\theta}$, where we use Theorem 6.4 of Talreja and Whitt (2009)[7]. Lemma 2 implies that $n\mathrm{Var}[W^n(\tau_t^n)]\to\frac{1}{\theta}$. Thus,

$$\begin{aligned} r[W^n(t^n), W^n(\tau_t^n)] &= \frac{\frac{n}{2}\mathrm{Var}[W^n(t^n)]+\frac{n}{2}\mathrm{Var}[W^n(\tau_t^n)]-\frac{1}{2}\mathrm{Var}[\sqrt{n}(W^n(t^n)-W^n(\tau_t^n))]}{\sqrt{n\mathrm{Var}[W^n(t^n)]\cdot n\mathrm{Var}[W^n(\tau_t^n)]}}\\ &\to 1-\frac{\frac{\rho-1}{\rho\theta}}{\frac{1}{\theta}}.\end{aligned}$$

[7] Talreja and Whitt (2009) do not condition on $W^n(t)>0$. However, $\sqrt{n}(W^n(t)-\bar{w})$ and $\sqrt{n}(W^n(t)-\bar{w})$ have the same distribution asymptotically in the ED regime, since $\mathbb{P}(W^n(t)>0)\to 1$.

That is,

$$r[W^n(t), W^n(\tau_t^n)] \to \frac{1}{\rho} \text{ in the ED regime.}$$

■

It is interesting to note that the correlation is independent of the individual abandonment rate, θ, i.e., of the speed at which customers abandon from the system. Moreover, the correlation is decreasing in the traffic intensity, ρ. We now turn to our results for the QED regime.

5.2.2. The QED Regime.

Here is our main result for the QED regime.

Proposition 4. *For the $M/M/n+M$ model in the many-server heavy-traffic QED regime:*

$$r[W^n(t^n), W^n(\tau_t^n)] \to 1 \quad as \quad n \to \infty. \tag{14}$$

Proof. For the proof, we need the following lemma.

Lemma 3. *In the QED regime,*

(a) For any $w \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}W^n(\tau_t^n) \geq w) = \lim_{n\to\infty} \mathbb{P}(\sqrt{n}W_S^n \geq w),$$

where W_S^n is a random variable with the distribution of a served customer who is delayed.

(b) For any $w \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}W^n(t^n) \geq w) = \lim_{n\to\infty} \mathbb{P}(\sqrt{n}W_S^n \geq w).$$

To prove the proposition, we can write:

$$r(W^n(t^n), W^n(\tau_t^n)) = \frac{\mathrm{Cov}[\sqrt{n}W^n(t^n), \sqrt{n}W^n(\tau_t^n)]}{\sqrt{\mathrm{Var}[\sqrt{n}W^n(t)]\mathrm{Var}[\sqrt{n}W^n(\tau_t^n)]}}.$$

We also have that,

$$\mathrm{Cov}[\sqrt{n}W^n(t^n), \sqrt{n}W^n(\tau_t^n)] = \frac{1}{2}\left(\mathrm{Var}[\sqrt{n}W^n(t^n)] + \mathrm{Var}[\sqrt{n}W^n(\tau_t^n)] - \mathrm{Var}[\sqrt{n}(W^n(t^n) - W^n(\tau_t^n))]\right).$$

We know that $\sqrt{n}(W^n(\tau_t^n) - W^n(t^n)) \Rightarrow 0$ as $n \to \infty$, which follows from Theorem 1 of Ibrahim et al. (2016), so that $\mathrm{Var}[\sqrt{n}(W^n(\tau_t^n) - W^n(t^n))] \to 0$. [8] By Lemma 3, and assuming uniform integrability, we have:

$$\lim_{n\to\infty} \mathrm{Var}[\sqrt{n}W^n(t^n)] = \lim_{n\to\infty} \mathrm{Var}[\sqrt{n}W^n(\tau_t^n)].$$

By Garnett et al. (2002), we have that $\sqrt{n}W^n(t^n)$ converges weakly to a finite random variable so that its variance converges as well to a positive constant. Thus,

$$r[W^n(t^n), W^n(\tau_t^n)] \to 1 \text{ in the QED regime.}$$

[8] Assuming uniform integrability for $\{W^n(\tau_t^n), n \geq 1\}$, which we conjecture, but have not established.

■

Combining the results in Propositions 1 and 3 allows for a simple explanation of our numerical observations in Figure 6: Based on (13), the LES announcement is less accurate than EA, in large systems, if $1/\rho \leq 0.5$, i.e., $\rho \geq 2$. Further, if $\rho < 2$, then the LES announcement is more accurate than EA. As before, although the number of servers in Figure 6 is relatively small, our results remain useful in roughly describing performance. Moreover, comparing the results of Propositions 2 and 3 allows for an explanation of the effect of abandonment on the system: Because the correlation increases (decreases) with ρ in the absence (presence) of abandonment, the relative performances of the LES announcement and EA are "reversed" when customers have a finite patience.

For robustness checks, we present in Table 3 simulation point estimates of correlations in the $M/M/100+G$ model, where times to abandon are allowed to follow a general, not necessarily exponential, distribution. In particular, we consider exponential, hyperexponential (H_2 with balanced means, squared coefficient of variation equal to 4, and mean equal to 2), and Erlang (E_2, with mean 2) abandonment distributions. We consider hyperexponential times to abandon because there is empirical evidence suggesting a good fit to this distribution in practice (Roubos and Jouini 2013). Table 3 illustrates that our asymptotic results in Proposition 3 are useful in describing performance in systems with a general time to abandon distribution as well, particularly with hyperexponential abandonment times and when the system is heavily congested. As in Table 2, the asymptotic results of Proposition 3 are less accurate with distributions that have low variability, such as Erlang (which has coefficient of variation equal to 1/2).

5.3. Priority Queue with Patient Customers

We now turn to the system with priority queues. We will first consider the system with patient customers. We index the arrival rate, λ, by either L or H to denote the low and high classes, respectively. As before, we consider a sequence of queueing systems indexed by n. We assume that λ_L^n and λ_H^n increase without bound, and the traffic intensities $\lambda_L^n/n\mu \equiv \rho_L$ and $\lambda_H^n/n\mu \equiv \rho_H$ are held fixed such that $\rho_L + \rho_H < 1$, to ensure stability. Here is our main result which characterizes the correlation between the LES delay and actual waiting time experienced by the low-priority customer. For high-priority customers, the same insights from single-class single-priority queues continue to hold, since high-priority customers do not "see" low-priority customers (if the service discipline is preemptive, this is exactly true; if the discipline is non-preemptive, this is asymptotically true when the system is sufficiently large).

PROPOSITION 5. *For low-priority customers in an $M/M/n$ two-class queueing system with a non-preemptive priority discipline, where $\lambda_L^n/n\mu \equiv \rho_L$, $\lambda_H^n/n\mu \equiv \rho_H$, and $\rho_H + \rho_L < 1$:*

$$r[W^n(\tau_t^n), W^n(t^n)] \to \frac{\rho_L}{1-\rho_H} \quad as \quad n \to \infty. \tag{15}$$

ρ	$1/\rho$	$M/M/100+M$	$M/M/100+H_2$	$M/M/100+E_2$
1.1	0.91	0.862 $\pm 8.2\times 10^{-4}$	0.767 $\pm 4.3\times 10^{-4}$	0.854 $\pm 8.4\times 10^{-4}$
1.3	0.77	0.766 $\pm 9.5\times 10^{-4}$	0.743 $\pm 9.4\times 10^{-4}$	0.653 $\pm 2.1\times 10^{-3}$
1.5	0.67	0.667 $\pm 2.1\times 10^{-3}$	0.665 $\pm 9.2\times 10^{-4}$	0.526 $\pm 2.3\times 10^{-3}$
1.7	0.588	0.588 $\pm 3.1\times 10^{-3}$	0.593 $\pm 1.4\times 10^{-3}$	0.439 $\pm 1.7\times 10^{-3}$
2	0.5	0.496 $\pm 3.9\times 10^{-3}$	0.511 $\pm 1.6\times 10^{-3}$	0.347 $\pm 2.2\times 10^{-3}$
2.2	0.45	0.455 $\pm 3.0\times 10^{-3}$	0.472 $\pm 2.1\times 10^{-3}$	0.304 $\pm 3.7\times 10^{-3}$
2.5	0.4	0.399 $\pm 1.9\times 10^{-3}$	0.424 $\pm 2.4\times 10^{-3}$	0.259 $\pm 4.1\times 10^{-3}$

Table 3 **Point estimates of correlations in the heavily-loaded $M/M/100+G$ queue, for alternative values of the traffic intensity ρ.**

Proof. For the proof of the proposition, we need the following lemma.

Lemma 4. *For all $x \geq 0$,*

$$\lim_{n\to\infty} \mathbb{P}(W^n(\tau_t^n) \geq x) = \lim_{n\to\infty} \mathbb{P}(W^n(t^n) \geq x).$$

Now, proceeding as in the proof for Proposition 2:

$$W^n(t^n) = \sum_{i=1}^{N_L(W^n(\tau_t^n))} B_i^n + Z^n,$$

where B_i^n is the length of a busy period in an $M/M/1$ queue with arrival rate λ_H^n and service rate $n\mu$, and Z^n is independent of $W^n(\tau_t^n)$. This is so because when the LES customer with low priority entered service, there must have been no H customers in queue. It is well known that $\mathbb{E}[B_i^n] = 1/(n\mu - \lambda_H^n)$; e.g., see Kleinrock (1975). Thus,

$$\begin{aligned}
\mathrm{Cov}[W^n(t^n), W^n(\tau_t^n)] &= \mathrm{Cov}\left[\sum_{i=1}^{N_L(W^n(\tau_t^n))} B_i^n, W^n(\tau_t^n)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N_L(W^n(\tau_t^n))} B_i^n\right) W^n(\tau_t^n)\right] - \mathbb{E}\left[\sum_{i=1}^{N_L(W^n(\tau_t^n))} B_i^n\right] \mathbb{E}[W^n(\tau_t^n)] \\
&= \frac{\rho_L^n}{1-\rho_H^n} \mathrm{Var}\left[W^n(\tau_t^n)\right].
\end{aligned}$$

By Lemma 4, assuming that appropriate uniform integrability holds, we obtain that $\lim_{n\to\infty}\text{Var}[W^n(\tau_t^n)] = \lim_{n\to\infty}\text{Var}[W^n(t^n)]$. This implies that:

$$r[W^n(\tau_t^n), W^n(t^n)] \to \frac{\rho_L}{1-\rho_H} \quad \text{as} \quad n \to \infty.$$

■

Combining the results of Propositions 1 and 5 allows for a simple explanation of our numerical observations in Figure 7. Based on (15), if $\rho_L/(1-\rho_H) \leq 0.5$, then the LES announcement is less accurate than EA, in large systems, and otherwise it is more accurate. Recall our assumption in Figure 7 that $\rho_H = 0.5$; thus, $\rho_L/(1-\rho_H) = \rho_L/\rho_H$ in this case, which explains our results in the figure. For robustness checks, we present in Table 4 simulation point estimates of correlations in the two-class $M/G/100$ model with non-preemptive priority, where we increase λ_L to increase ρ_L, and keep $\rho_H = 0.5$ constant. Table 4 illustrates that our asymptotic results in Proposition 4 are useful in describing performance in systems with a general service-time distribution as well.

ρ_L	$\frac{\rho_L}{1-\rho_H}$	$M/M/100$	$M/LN(1,1)/100$	$M/E_2/100$
0.35	0.7	0.659 $\pm 7.1\times10^{-3}$	0.664 $\pm 2.5\times10^{-3}$	0.601 $\pm 5.8\times10^{-3}$
0.375	0.75	0.707 $\pm 4.1\times10^{-3}$	0.729 $\pm 3.5\times10^{-3}$	0.666 $\pm 7.4\times10^{-3}$
0.4	0.8	0.771 $\pm 2.6\times10^{-3}$	0.785 $\pm 2.2\times10^{-3}$	0.734 $\pm 3.6\times10^{-3}$
0.425	0.85	0.8324 $\pm 3.2\times10^{-3}$	0.846 $\pm 2.1\times10^{-3}$	0.805 $\pm 2.5\times10^{-3}$
0.45	0.9	0.892 $\pm 1.9\times10^{-3}$	0.903 $\pm 3.0\times10^{-3}$	0.873 $\pm 3.7\times10^{-3}$
0.475	0.95	0.947 $\pm 1.8\times10^{-3}$	0.951 $\pm 2.0\times10^{-3}$	0.942 $\pm 1.8\times10^{-3}$
0.49	0.98	0.976 $\pm 1.3\times10^{-3}$	0.980 $\pm 8.7\times10^{-4}$	0.979 $\pm 3.2\times10^{-3}$

Table 4 Estimates of correlations for low-priority customers (with corresponding 95% confidence intervals) in the two-class $M/G/100$ queue with $\rho_H = 0.5$ and varying traffic intensity ρ_L.

5.4. Priority Queue with Impatient Customers

We now turn to the case of a priority queue with abandonment. We use the same notation as in the previous section. As before, we consider a sequence of queueing systems indexed by n. We consider an exponential abandonment-time distribution, and assume that the abandonment rate $\theta^n \equiv \theta$ is constant as n increases. Here is our main result, which we prove in the appendix.

Proposition 6. *For low-priority customers in an $M/M/n+M$ two-class queueing system with a non-preemptive priority discipline, where $\lambda_L^n/n\mu \equiv \rho_L$, $\lambda_H^n/n\mu \equiv \rho_H$, and $\rho_L + \rho_H > 1$:*

$$r[W^n(\tau_t^n), W^n(t^n)] \to \frac{1-\rho_H}{\rho_L}. \quad \textit{as} \quad n \to \infty. \tag{16}$$

Proposition 6 allows for a simple explanation of our numerical observations in Figure 8. Based on (16), the LES announcement is less (more) accurate than EA, in large systems, if $(1-\rho_H)/\rho_L \leq$ $(>)$ 0.5. Recall our assumption in Figure 8 that $\rho_H = 0.5$; thus, $\rho_L/(1-\rho_H) = \rho_L/\rho_H$ in this case, which explains our results in the figure. For robustness checks, we present in Table 5 simulation point estimates of correlations in the two-class $M/M/100+G$ model with non-preemptive priority, where we increase λ_L to increase ρ_L, and keep $\rho_H = 0.5$ constant. Table 5 illustrates that our asymptotic results in the proposition remain useful in describing performance in systems with a general abandonment-time distribution as well.

6. Numerical Study

In this section, we begin in §6.1 by describing results of simulation experiments for additional numerical support. Our objective is two-fold: (i) to substantiate and extend our theoretical results; and (ii) to test the performance of the new WA predictor, which combines the LES announcement and EA, under more general modelling assumptions. We also describe the results of additional simulation experiments in the appendix (§EC.3) and in an online supplement to this main paper. In all models considered below, the WA predictor consistently outperforms both the LES announcement and EA. Then, we revisit the empirical results of §3 in §6.2. Finally, we consider an alternative data set in §6.3.

6.1. Performance of the WA Prediction

We begin by considering generally-distributed, i.e., non-exponential, service and abandonment times. In Table 6, we consider both exponential and lognormal ($LN(1,1)$) service times for a single-class model. We vary the value of the traffic intensity, ρ, and present a point estimate of the ASE for each delay announcement, for that value of ρ. (We report corresponding estimates of 95% confidence intervals in the supplement, and key insights continue to hold.) For the WA prediction, we consider two alternatives: (i) we use the theoretical asymptotic expression for the correlation,

ρ_L	$\frac{1-\rho_H}{\rho_L}$	$M/M/100+M$	$M/M/100+H_2$	$M/M/100+E_2$
0.55	0.91	0.728 $\pm 9.0\times 10^{-4}$	0.547 $\pm 1.0\times 10^{-3}$	0.767 $\pm 6.8\times 10^{-4}$
0.65	0.77	0.721 $\pm 7.7\times 10^{-4}$	0.583 $\pm 8.3\times 10^{-4}$	0.647 $\pm 2.0\times 10^{-3}$
0.75	0.67	0.648 $\pm 1.2\times 10^{-3}$	0.580 $\pm 8.3\times 10^{-4}$	0.516 $\pm 2.1\times 10^{-3}$
0.85	0.59	0.574 $\pm 2.4\times 10^{-3}$	0.549 $\pm 9.8\times 10^{-4}$	0.427 $\pm 1.9\times 10^{-3}$
1	0.5	0.487 $\pm 1.5\times 10^{-3}$	0.486 $\pm 1.4\times 10^{-3}$	0.337 $\pm 2.8\times 10^{-3}$
1.1	0.45	0.444 $\pm 2.7\times 10^{-3}$	0.448 $\pm 1.1\times 10^{-3}$	0.294 $\pm 2.4\times 10^{-3}$
1.25	0.4	0.388 $\pm 2.4\times 10^{-3}$	0.402 $\pm 2.1\times 10^{-3}$	0.248 $\pm 4.0\times 10^{-3}$

Table 5 Estimates of correlations for low-priority customers (with corresponding 95% confidence intervals) in the two-class $M/M/100+G$ queue with $\rho_H = 0.5$ and alternative values of traffic intensity ρ_L.

as given by our analysis in §5, depending on the model, and (ii) we consider a running-average simulation-based estimate for the correlation; the corresponding predictor is denoted WA-run. We also consider a delay prediction which is equal to an exponentially smoothed average over previous LES delays (Holt 2004), where we estimate the value of the smoothing factor in the simulation by using a gradient-descent method to minimize the errors between the smoothed averages and actual delays, in a training set consisting of 100 data points (after steady state is reached). We denote this predictor by EXP.

Table 6 shows that both WA and WA-run consistently outperform the remaining predictors, for all values of ρ considered. These two predictors also have a very similar performance, which further substantiates our earlier asymptotic results. We also note that ASE(EXP) is almost indistinguishable from ASE(LES), particularly when ρ is large enough: This illustrates that there is no advantage in averaging over previous LES delays in this case. In Table 7, we present results for generally-distributed times to abandon in a two-class model. In particular, we consider H_2 times to abandon with mean equal to 1 and variance equal to 4; we report results for low-priority customers only. Table 7 shows that similar observations continue to hold: WA and WA-run perform almost the same, consistently outperforming the rest of the predictors. The exponentially-smoothed prediction performs almost the same as the LES announcement, i.e., there is no advantage in averaging over previous LES delays.

$M/M/100$

ρ	LES	EA	WA	WA-run	EXP	$\mathbb{E}[W \vert W>0]$
0.7	0.0293	0.0369	0.0271	0.0274	0.0292	0.0370
0.8	0.0345	0.0502	0.0325	0.0325	0.0344	0.0507
0.9	0.0466	0.0992	0.0453	0.0453	0.0466	0.0993
0.98	0.101	0.508	0.100	0.101	0.101	0.491

$M/LN(1,1)/100$

ρ	LES	EA	WA	WA-run	EXP	$\mathbb{E}[W \vert W>0]$
0.7	0.0299	0.0368	0.0280	0.0282	0.0299	0.0348
0.8	0.0321	0.0467	0.0302	0.0302	0.0320	0.0458
0.9	0.0430	0.0920	0.0418	0.0418	0.0430	0.0920
0.98	0.0970	0.500	0.0967	0.0967	0.0970	0.497

Table 6 Comparison of the square-root ASE's for the different predictions in the $M/G/100$ model for alternative values of the traffic intensity, ρ.

6.2. Revisiting the Data

Recall that the LES prediction may be significantly less accurate than the EA prediction in our data set; see Figures 3 and 4. Primarily, our goal in this section is to test the performance of the WA predictor with data. We will show that WA usually yields superior performance to both the LES announcement and EA, thus further substantiating the usefulness of that predictor in practice.

We set $\text{WA} \equiv (1-\hat{r}) \cdot \text{EA} + \hat{r} \cdot \text{LES}$, where $\hat{r}$ is a point estimate for the correlation, which we calculate based on data. For both EA and $\hat{r}$, we calculate out-of-sample point estimates which are based on the first 2,000 delayed callers, beginning August 1, which are later discarded from the sample. For IN callers, our point estimate for $\hat{r}_{IN} = 0.24$ and, for low-priority PS callers, it is given by $\hat{r}_{PS} = 0.17$. Our point estimates for the out-of-sample average waiting times, for each call type, are $\text{EA}_{IN} = 173$ seconds and $\text{EA}_{PS} = 121$ seconds. We use these estimates as static announcements in our data set. In the e-companion, we take a closer look at performance by classifying delayed callers into different groups, depending on their waiting times, and present results for other call types as well. Here, we summarize our key results, which hold broadly across all call types.

6.2.1. Accuracy of the WA Prediction. Figures 9 and 10 parallel Figures 3 and 4, with an additional curve corresponding to the errors for WA. In particular, for the same days in the sample set, we plot the relative average squared errors corresponding to the new WA predictor as well. Figures 9 and 10 show that the new predictor outperforms both the LES announcement

$M/M/100+M$ with two classes

ρ_L	ρ_H	LES	EA	WA	WA-run	EXP	$\mathbb{E}[W\|W>0]$
0.55	0.5	0.163	0.219	0.157	0.152	0.163	0.301
0.65	0.5	0.203	0.269	0.189	0.189	0.203	0.537
0.85	0.5	0.260	0.281	0.231	0.231	0.260	1.062
1	0.5	0.284	0.281	0.246	0.246	0.284	1.385
1.25	0.5	0.310	0.280	0.259	0.259	0.310	1.832

$M/M/100+H_2$ with two classes

ρ_L	ρ_H	LES	EA	WA	WA-run	EXP	$\mathbb{E}[W\|W>0]$
0.55	0.5	0.114	0.118	0.109	0.101	0.113	0.145
0.65	0.5	0.127	0.137	0.116	0.114	0.126	0.201
0.85	0.5	0.151	0.158	0.134	0.134	0.151	0.346
1	0.5	0.165	0.163	0.143	0.143	0.165	0.454
1.25	0.5	0.181	0.166	0.153	0.153	0.181	0.610

Table 7 Comparison of the square-root ASE's of the different predictions for low-priority customers, in the two-class $M/M/100+G$ model, with alternative values of the traffic intensity, ρ.

and EA. We take a closer look at performance in Table 8, where we present data estimates for the ratios of ASE's of our three predictors. The first sub-table corresponds to IN callers, whereas the second corresponds to low-priority PS callers. Each column in the table corresponds to days where one of the predictors yields the smallest ASE. For example, for the first column, we restrict attention to those days where the LES announcement yielded the smallest ASE: For IN calls, the LES announcement yielded the smallest ASE on 5 days out of the 85 days in our sample, i.e., on 6% of the days. On those days, we present estimates of ASE(EA)/ASE(LES) (first row) and ASE(WA)/ASE(LES) (second row). Based on Table 8, we can make the following observations. First, it is clear that WA outperforms both the LES announcement and EA: The proportions of days over which ASE(WA) is smallest is considerably greater for both call types. For an aggregate measure of performance with IN callers, we note that, averaging the ASE's across all days in our sample: ASE(LES)/ASE(WA) $=1.75$ and ASE(EA)/ASE(WA) $=1.05$. For an aggregate measure of performance with PS callers, we note that, averaging the ASE's across all days in our sample: ASE(LES)$/ASE$(WA) $=1.43$ and ASE(EA)/ASE(WA) $=1.11$. In other words, it is clear that WA usually performs best on average.

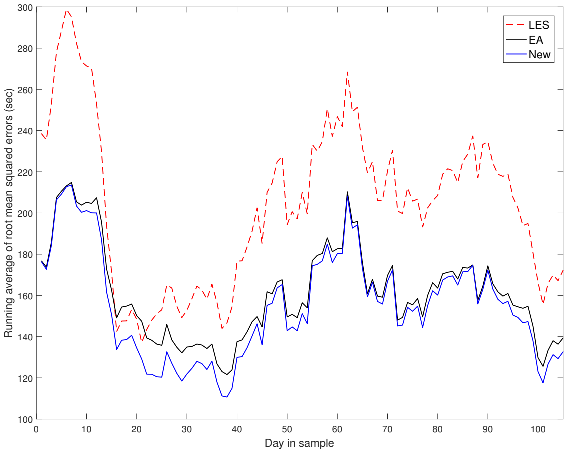

Figure 9 IN customers (August-December): System is a single-class single-priority queue.

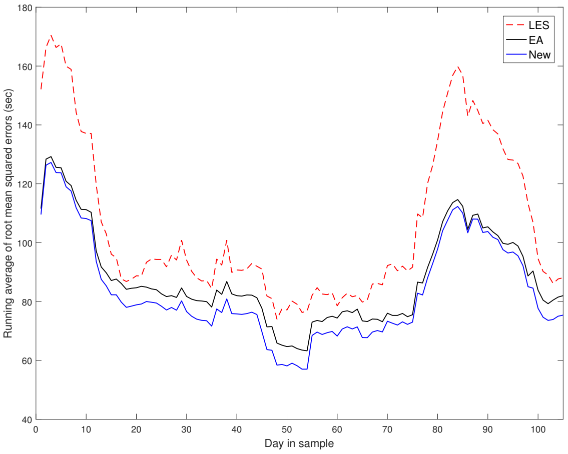

Figure 10 Low-priority PS customers (August-December).

Second, our results indicate that even on days where either the LES announcement or EA yield the smallest ASE's (first and second columns in the tables), WA does not perform too poorly. For example, on days where the LES announcement yields the smallest ASE, we have that ASE(WA)/ASE(LES) = 1.14 (these numbers are computed on average across the sub-sample of days). In contrast, ASE(EA)/ASE(LES) = 2.38 for those same days. This means that even though the LES announcement yields the smallest ASE, it performs much better than EA, but that this is not the case for WA. The same observations hold when EA yields the smallest ASE (second column in the table) as well. In other words, WA usually performs better than both the LES announcement and EA, and when it is outperformed by either predictor, it remains a reasonably accurate prediction, i.e., it does not perform too poorly. Third, when WA yields the smallest ASE (third column in the table), it significantly outperforms both EA and the LES announcement. For example, for IN calls, ASE(EA) is 53% larger than ASE(WA) and ASE(LES) is more than three times ASE(WA).

In the e-companion (§EC.4.4), we also consider a data-based predictor which exploits information about the queue-length seen by a delayed customer upon arrival in a single-class system, and illustrate that this prediction is outperformed by the WA prediction too.

6.3. An Additional Data Set from a Larger Call Center

The real-life data set that we have analyzed in the previous section is taken from a small call center (number of agents is less than 15). To check the robustness of our results, we now consider an additional data set taken from a larger call center, where calls are handled by a pool of 200 agents. In particular, we use data from the call center of a Dutch company which specializes in delivering business solutions to its clients.

IN Call Type (single-class queue with a single priority)

	Ratios of (row) ASE to (column) winner ASE			
	EA wins (28%)	WA wins (66%)	LES wins (6%)	Overall
EA	1	1.53	2.38	1.43
WA	1.05	1	1.14	1.02
LES	2.09	3.20	1	2.75

PS Call Type (low-priority callers)

	Ratios of (row) ASE to (column) winner ASE			
	EA wins (8%)	WA wins (69%)	LES wins (23%)	Overall
EA	1	1.15	2.26	1.39
WA	1.02	1	1.70	1.16
LES	2.25	1.58	1	1.50

Table 8 Comparison of the ASE's of EA, WA, and LES in August-December for IN and PS customers. In each column, we report estimates of the ratio of the ASE of the prediction in the corresponding row, relative to the ASE of the predictor in the corresponding column.

6.3.1. Description of the Data. There are 11 different queues in the call center, and each queue corresponds to either one or two call types. We focus on one such queue: Queue-30170. We select this queue because it corresponds to a single call type. However, it is important to note that it is served by an agent pool which may be serving other call types at the same time. The total number of agents who serve Queue-30170 is equal to 148. Our data does not contain information about the routing policy for any of the queues. Thus, it reflects a realistic scenario, where the manager of the call center has information about the waiting times of customers, but the routing itself may be done in an ad-hoc manner. The call center is closed on Sundays. The average wait time for delayed customers is close to 90 seconds, and the probability of abandonment is close to 6%.

Our data for Queue-30170 is from May 9, 2012 until September 29, 2012. There are close to 50,000 delayed customers in that set. For each delayed customer, we proceed as before and calculate the LES prediction, the running average wait-time prediction, and the WA prediction which is based on a point estimate of the correlation. We calculate out-of-sample estimates for the correlation and the average waiting time based on a sample of 10,000 delayed customers. We then discard this sample from consideration when calculating the errors corresponding to our alternative predictions. As such, we are left with predictions for a total of 98 consecutive days. For each of the average wait-time and correlation estimates, we include a weekday effect. To compare the performances of our alternative predictions, we focused on delays which exceeded 5 seconds.

6.3.2. Accuracy of the WA Prediction. We present in Table 9 results which parallel those that we reported in Table 8. Based on Table 9, we can make the following observations. First, the

LES prediction is increasingly accurate as the size of the system grows. This is to be expected, and is in concert with previous theoretical results establishing the asymptotic accuracy of the LES prediction in large queueing systems, e.g., see Ibrahim et al. (2016). In particular, the LES prediction performs generally better than the average-waiting-time prediction in this case. Second, our new proposed WA prediction has a clear superior performance compared to both the LES and EA predictions. Indeed, WA "wins" on close to 75% of the days in our sample. Third, by restricting attention to days where the LES announcement yields the smallest daily error on average (first column in Table 9), we see that it outperforms the EA prediction by a lot in this case: Indeed, the ASE for EA is roughly 13 times the ASE for the LES announcement in that case. In other words, announcing the average can lead to considerable errors. Finally, we have further evidence that WA is a good announcement in practice: *Even when WA is not the most accurate prediction, it remains competitive, i.e., there do not exist days when it is dramatically outperformed by either the LES announcement or EA.*

Queue-30170

	Ratios of (row) ASE to (column) winner ASE			
	EA wins (8%)	WA wins (74%)	LES wins (17%)	Overall
EA	1	2.42	13.1	4.09
WA	1.30	1	2.40	1.25
LES	2.03	1.43	1	1.39

Table 9 Comparison of the ASE's of EA, WA, and LES for Queue-30170. In each column, we report estimates of the ratio of the ASE of the prediction in the corresponding row, relative to the ASE of the predictor in the corresponding column.

7. Conclusions and Future Research

In this paper, we compared the performance of the LES and static delay announcements. We developed a new *correlation-based assessment* which enables an easy comparison of the performances of static and dynamic announcements across several queueing models. The main takeaway from our analysis is that it is indeed justifiable to resort to simple static announcements, in certain cases, as is commonly done in practice. Indeed, even though the LES announcement takes real-time information into account, it may have worse accuracy than the simple static announcement.

Our asymptotic results, on the values of correlations, provide insights on how the performances of the LES announcement and EA depend on the traffic intensity in the system. In general, static announcements are appropriate in low to moderately congested systems when there is little or no abandonment, and under heavy congestion when there is considerable abandonment. Our data analysis revealed that they are especially useful in small systems, but that they may be significantly outperformed by dynamic announcements, such as the LES announcement, in large systems. Our numerical study suggests that these results continue to hold in multi-class systems with abandonment, and with time-varying arrival rates as well. Our theoretical, empirical, and numerical results all support the superiority of our new WA prediction.

There remains several extensions which are worth exploring in future research. Extensions of our results for the two-class priority system to systems with multiple classes is a direct extension. The system with time-varying arrivals, to which we only presented numerical results at this stage, is worth exploring as well. It is also of interest to incorporate customer response to the announcements in our framework, e.g., in the spirit of Ibrahim et al. (2016), and to explore the impact of customer response on correlations in those settings.

Acknowledgments

The authors are grateful to colleagues from the SEE Lab at the Technion for making the data publicly available, and for the detailed supporting documentation. The authors are also grateful to Ger Koole for providing them with the second data set.

References

Akşin, Zeynep, Baris Ata, Seyed Morteza Emadi, Che-Lin Su. 2016. Impact of delay announcements in call centers: An empirical approach. *Operations Research* **65**(1) 242–265.

Ang, Erjie, Sara Kwasnick, Mohsen Bayati, Erica L Plambeck, Michael Aratow. 2015. Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* **18**(1) 141–156.

Armony, Mor, Constantinos Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.

Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.

Bitran, Gabriel, René Caldentey. 2002. Two-class priority queueing system with state-dependent arrivals. *Queueing Systems* **40**(4) 355–382.

Brown, Lawrence, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, Linda Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association* **100**(469) 36–50.

Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2017. The impact of delay announcements on hospital network coordination and waiting times. Northwestern University, working paper.

Eick, Stephen G, William A Massey, Ward Whitt. 1993. $m_t/g/\infty$ queues with sinusoidal arrival rates. *Management Science* **39**(2) 241–252.

Gal, Avigdor, Avishai Mandelbaum, François Schnitzler, Arik Senderovich, Matthias Weidlich. 2015. Traveling time prediction in scheduled transportation with journey segments. *Information Systems* .

Gal, Avigdor, Avishai Mandelbaum, François Schnitzler, Arik Senderovich, Matthias Weidlich. 2017. Traveling time prediction in scheduled transportation with journey segments. *Information Systems* **64**(C) 266–280.

Garnett, Ofer, Avi Mandelbaum, Martin Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* **29**(3) 567–588.

Holt, Charles C. 2004. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting* **20**(1) 5–10.

Ibrahim, Rouba. 2010. Real-time delay prediction in customer service systems. Ph.D. thesis, COLUMBIA UNIVERSITY.

Ibrahim, Rouba. 2017. Sharing delay information in queueing systems: A literature survey. Tech. rep.

Ibrahim, Rouba, Mor Armony, Achal Bassamboo. 2016. Does the past predict the future? the case of delay announcements in service systems. *Management Science* .

Ibrahim, Rouba, Ward Whitt. 2009. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management* **11**(3) 397–415.

Ibrahim, Rouba, Ward Whitt. 2011. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and Operations Management* **20**(5) 654–667.

Jouini, Oualid, O Zeynep Akşin, Fikri Karaesmen, M Salah Aguir, Yves Dallery. 2015. Call center delay announcement using a newsvendor-like performance criterion. *Production and Operations Management* **24**(4) 587–604.

Jouini, Oualid, Yves Dallery, Zeynep Akşin. 2009. Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics* **120**(2) 389–399.

Kleinrock, Leonard. 1975. *Queueing systems, volume 2: Computer applications*, vol. 66.

Nakibly, Efrat. 2002. Predicting waiting times in telephone service systems. Ph.D. thesis, Technion–Israel Institute of Technology.

Pender, Jamol, Richard Rand, Elizabeth Wesson. 2017. Queues with choice via delay dierential equations. Tech. rep.

Reiman, Martin I. 1984. Open queueing networks in heavy traffic. *Mathematics of operations research* **9**(3) 441–458.

Roubos, Alex, Oualid Jouini. 2013. Call centers with hyperexponential patience modeling. *International Journal of Production Economics* **141**(1) 307–315.

Senderovich, Arik, Matthias Weidlich, Avigdor Gal, Avishai Mandelbaum. 2014. Queue mining–predicting delays in service processes. *International Conference on Advanced Information Systems Engineering*. Springer, 42–57.

Senderovich, Arik, Matthias Weidlich, Avigdor Gal, Avishai Mandelbaum. 2015. Queue mining for delay prediction in multi-class service processes. *Information Systems* **53** 278–295.

Singh, Siddharth Prakash, Mohammad Delasay, Alan Scheller-Wolf. 2017. Evaluating the first-movers advantage in announcing real-time delay information. Tech. rep.

Talreja, Rishi, Ward Whitt. 2009. Heavy-traffic limits for waiting times in many-server queues with abandonment. *The Annals of Applied Probability* 2137–2175.

Thiongane, Mamadou, Wyean Chan, Pierre L'Ecuyer. 2015. Waiting time predictors for multi-skill call centers. *Proceedings of the 2015 Winter Simulation Conference*. IEEE Press, 3073–3084.

Ward, A. R., W. Whitt. 2000. Predicting response times in processor-sharing queues. *Analysis of Communication Networks: Call Centres, Traffic, and Performance* **28** 1.

Whitt, Ward. 1999. Predicting queueing delays. *Management Science* **45**(6) 870–888.

Whitt, Ward. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.

Yu, Qiuping, Gad Allon, Achal Bassamboo. 2017. The reference effect of delay announcements: A field experiment. Tech. rep.

Electronic Companion:

In this e-companion, we present supporting material to the main paper. In §EC.1, we prove technical lemmas 1-4. In §EC.4, we present additional empirical support which corresponds to analyzing our first call-center data set, described in §3.

EC.1. Proofs of Technical Lemmas

EC.1.1. Proof of Lemma 1

Proof. First, we prove part (a). The statement holds trivially for $x = 0$. Now, for $x > 0$:

$$\begin{aligned}\mathbb{P}(W^n(\tau_t^n) \geq x) &= \mathbb{P}(W^n(\tau_t^n) \geq x | W^n(\tau_t^n) > 0)\mathbb{P}(W^n(\tau_t^n) > 0)\\ &= \mathbb{P}(W_\infty^n \geq x | W_\infty^n > 0, D^n, E^n)\mathbb{P}(W^n(\tau_t^n) > 0),\end{aligned}$$

where we define the following events:

- D^n: next arrival after entry to service is delayed
- E^n: next arrival is before next entry to service

We also note that:

$$\mathbb{P}(W^n(\tau_t^n) > 0) = 1 - \mathbb{P}(\text{LES customer finds exactly } n-1 \text{ customers in the system upon arrival}).$$

This is so because an LES customer must find at least $n-1$ customers in the system upon arrival. To see why, assume, aiming at a contradiction, that LES encounters $k < n-1$ customers in the system upon arrival. Then, it must be that the next arriving customer is not delayed, i.e., the current customer could not be an LES customer; this is a contradiction. Further,

$$\begin{aligned}\mathbb{P}(W_\infty^n \geq x | W_\infty^n > 0, D^n, E^n) &= \frac{\mathbb{P}(E^n, D^n | W_\infty^n \geq x, W_\infty^n > 0)\mathbb{P}(W_\infty^n \geq x, W_\infty^n > 0)}{\mathbb{P}(W_\infty^n > 0, E^n, D^n)}\\ &= \frac{\mathbb{P}(E^n, D^n | W_\infty^n \geq x)\mathbb{P}(W_\infty^n \geq x | W_\infty^n > 0)}{\mathbb{P}(E^n, D^n | W_\infty^n > 0)}\\ &= \mathbb{P}(W_\infty^n \geq x | W_\infty^n > 0),\end{aligned}$$

since $\mathbb{P}(D^n, E^n | W_\infty^n \geq x) = \mathbb{P}(D^n, E^n | W_\infty^n > 0)$. This is so because:

$$\begin{aligned}\mathbb{P}(D^n, E^n | W_\infty^n \geq x) &= \mathbb{P}(D^n | E^n, W_\infty^n \geq x)\mathbb{P}(E^n | W_\infty^n \geq x)\\ &= \mathbb{P}(E^n | W_\infty^n \geq x) \text{ since } \mathbb{P}(D^n | E^n, W_\infty^n \geq x) = 1\\ &= \frac{\lambda^n}{\lambda^n + n\mu}\\ &= \mathbb{P}(E^n | W_\infty^n > 0)\\ &= \mathbb{P}(D^n | E^n, W_\infty^n > 0)\mathbb{P}(E^n | W_\infty^n > 0) \text{ since } \mathbb{P}(D^n | E^n, W_\infty^n > 0) = 1\\ &= \mathbb{P}(D^n, E^n | W_\infty^n > 0).\end{aligned}$$

Finally, letting π_{n-1} denote the probability that the LES customer encounters $n-1$ customers in the system upon arrival, we must have that $\pi^n_{n-1} \to 0$ as $n \uparrow \infty$. This is so because for every n fixed we have $\sum_{k=0}^{\infty} \pi^n_k = 1$ which implies that $\pi^n_k \to 0$ as $k \to \infty$. Thus, for $k \geq M_\epsilon$, we have: $\pi^n_k < \epsilon$. This implies that for $n \geq M_\epsilon + 1$, we must also have that $\pi^n_{n-1} < \epsilon$. That is, $\pi^n_{n-1} \to 0$ as $n \to \infty$. This implies that $\mathbb{P}(W^n(\tau^n_t) > 0) \to 1$. Thus, we obtain:

$$\lim_{n\to\infty} \mathbb{P}(W^n(\tau^n_t) \geq x) = \lim_{n\to\infty} \mathbb{P}(W^n(t) \geq x) = \mathbb{P}(W_\infty \geq x | W_\infty > 0),$$

as desired, where $[W_\infty | W_\infty > 0]$ is the corresponding limiting steady-state distribution which is proper under both limiting regimes. To show that for every t:

$$\lim_{n\to\infty} \text{Var}[W^n(\tau^n_t)] = \lim_{n\to\infty} \text{Var}[W^n(t)] = \text{Var}[W_\infty \geq x | W_\infty > 0],$$

we need to show that uniform integrability of the sequence $\{W^n(\tau^n_t), n \geq 1\}$ holds. To do so, note that for any $x \geq 0$:

$$\begin{aligned}
\mathbb{P}[W^n(\tau^n_t) \geq x] &= \mathbb{P}[W^n_\infty \geq x | E^n, D^n] \\
&\leq \mathbb{P}[W^n_\infty \geq x | E^n, D^n, W^n_\infty > 0] \\
&= \frac{\mathbb{P}[W^n_\infty \geq x, E^n, D^n | W^n_\infty > 0]}{\mathbb{P}[E^n, D^n | W^n_\infty > 0]} \\
&\leq \frac{\mathbb{P}[W^n_\infty \geq x | W^n_\infty > 0]}{\mathbb{P}[E^n, D^n | W^n_\infty > 0]} \\
&\leq \frac{\mathbb{P}[W^n_\infty \geq x | W^n_\infty > 0]}{\frac{\rho^n}{\rho^n + 1}} \\
&= \mathbb{P}[W^n_\infty \geq x | W^n_\infty > 0] \frac{\rho^n + 1}{\rho^n}.
\end{aligned}$$

Thus, for any $x \geq 0$,

$$x \cdot \mathbb{P}[W^n(\tau^n_t) \geq x] \leq x \cdot \mathbb{P}[W^n_\infty \geq x | W^n_\infty > 0] \frac{\rho^n + 1}{\rho^n}.$$

This implies,

$$\begin{aligned}
\mathbb{E}[(W^n(\tau^n_t))^2] = \int_0^\infty 2x \mathbb{P}[W^n(\tau^n_t) \geq x] dx &\leq \frac{\rho^n + 1}{\rho^n} \int_0^\infty 2x \mathbb{P}[W^n_\infty \geq x | W^n_\infty > 0] dx \\
&= \frac{\rho^n + 1}{\rho^n} \mathbb{E}[(W^n_\infty)^2 | W^n_\infty > 0] \\
&< M,
\end{aligned}$$

for $M < \infty$ large enough. Thus, $\sup_{n\geq 1}\{\mathbb{E}[(W^n(\tau^n_t))^2]\} < \infty$, and the sequence $\{W^n(\tau^n_t), n \geq 1\}$ is uniformly integrable: Convergence of moments then follows from the convergence in distribution together with the established uniform integrability. ■

EC.1.2. Proof of Lemma 2

Proof. First, note that $W^n(\tau_t^n) =^{\mathcal{D}} [W^n|S^n, A^n, E^n]$ where

- Event S^n: "customer is served"
- Event A^n: "next arrival after current entry to service is before next entry to service"
- Event E^n: "next arrival is delayed"

Part (a): The lemma holds trivially at $w = 0$. Letting $w > 0$, we have:

$$\begin{aligned}
&\mathbb{P}(W^n(\tau^n) \geq w)\\
&= \int_w^\infty f_{W|A,S,E}(x|A^n, S^n, E^n)dx\\
&= \int_w^\infty \frac{f_{W|S,E}(x|S^n, E^n)\mathbb{P}(A^n|W^n = x, S^n, E^n)\mathbb{P}(S^n, E^n)}{\mathbb{P}(A^n, S^n, E^n)}dx\\
&= \frac{\mathbb{P}(S^n, E^n)}{\mathbb{P}(A^n, S^n, E^n)} \times\\
&\quad \int_w^\infty f_{W|S,E}(x|S^n, E^n)(\mathbb{P}(A^n|Q^n = 0, W^n = x, S^n, E^n)\mathbb{P}(Q^n = 0|W^n = x, S^n, E^n)\\
&+ \mathbb{P}(A^n|Q^n > 0, W^n = x, S^n, E^n)\mathbb{P}(Q^n > 0|W^n = x, S^n, E^n))dx.
\end{aligned}$$

Since $\mathbb{P}(E^n) \to 1$, we can remove it hereafter from the conditioning event since it does not matter asymptotically. Let Q^n be the number of customers left in queue when LES enters service. Note that $[Q^n|W^n = x, S^n]$ is distributed as the number of customers at time x in an $M/M/\infty$ queue starting out empty a time 0 which is Poisson distributed with rate $\frac{\lambda^n}{\theta}(1 - e^{-\theta x})$. Thus, for $x > 0$,

$$\lim_{n\to\infty} \mathbb{P}(Q^n = 0|W^n = x, S^n) = 0 \text{ and } \lim_{n\to\infty} \mathbb{P}(Q^n > 0|W^n = x, S^n) = 1.$$

Also, note that $\mathbb{P}(A^n|Q^n > 0, W^n = x, S^n) = \frac{\lambda^n}{\lambda^n + n\mu} = \frac{\rho}{\rho+1}$. We now show that $\mathbb{P}(A^n|S^n, E^n) \to \frac{\rho}{\rho+1}$ as well. This is obtained by a similar conditioning argument, as follows.

$$\begin{aligned}
&\mathbb{P}(A^n|S^n)\\
&= \int_0^\infty \mathbb{P}(A^n|S^n, W^n = x)f_W(x|S^n)dx\\
&= \int_{0^+}^\infty \mathbb{P}(A^n|S^n, W^n = x, Q^n = 0)\mathbb{P}(Q^n = 0|S^n, W^n = x)f_W(x|S^n)dx\\
&+ \int_{0^+}^\infty \mathbb{P}(A^n|S^n, W^n = x, Q^n > 0)\mathbb{P}(Q^n > 0|S^n, W^n = x)f_W(x|S^n)dx\\
&+ \mathbb{P}(W^n = 0|S^n)\mathbb{P}(A^n|S^n, W^n = 0)\mathbb{P}(Q^n = 0|S^n, W^n = 0)\\
&= \frac{\rho}{\rho+1}\int_{0^+}^\infty \mathbb{P}(Q^n > 0|S^n, W^n = x)f_W(x|S^n)dx + \mathbb{P}(W^n = 0|S^n)\\
&\to \frac{\rho}{\rho+1}.
\end{aligned}$$

Thus,

$$\lim_{n\to\infty} \mathbb{P}(W^n(\tau_t^n) \geq w) = \lim_{n\to\infty} \mathbb{P}(W^n(t) \geq w|S^n) = \lim_{n\to\infty} \mathbb{P}(W_S^n \geq w).$$

Part (b) We use three lemmas.

LEMMA EC.1. *As $n \to \infty$,*

$$\sqrt{n}\left(\sum_{i=0}^{\lfloor nt \rfloor} Y_{i,n} - c(t)\right) \Rightarrow \text{Nor}(0, d(t)), \tag{EC.1}$$

where $c(t) \equiv \frac{1}{\theta}\ln\left(1+\frac{\theta t}{\mu}\right)$ and $d(t) \equiv \frac{t}{\mu(\mu+\theta t)}$ for all $t \geq 0$.

Proof. Let $m_i \equiv \mathbb{E}[Y_{i,n}] = \frac{1}{n\mu+(i+1)\theta}$ and $\sigma_i^2 \equiv \text{Var}(Y_{i,n}) = \left(\frac{1}{n\mu+(i+1)\theta}\right)^2$. Then, for any $t \geq 0$:

$$\frac{\max_{0\leq j\leq \lfloor nt \rfloor} m_j^2}{\sum_{j=0}^{\lfloor nt \rfloor} m_j^2} \to 0 \text{ as } n \to \infty.$$

By Lemma EC.3 (which applies the Lindeberg-Feller CLT), we must have that

$$\frac{\sum_{i=0}^{\lfloor nt \rfloor}(Y_{i,n} - m_i)}{\sqrt{\sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2}} \Rightarrow \text{Nor}(0,1). \tag{EC.2}$$

To obtain (EC.1), note that for every $t \geq 0$:

$$\sum_{i=0}^{\lfloor nt \rfloor} m_i \to c(t) \equiv \frac{1}{\theta}\ln\left(1+\frac{\theta t}{\mu}\right) \text{ and } n\sum_{i=0}^{\lfloor nt \rfloor}\sigma_i^2 \to d(t) \equiv \frac{t}{\mu(\mu+\theta t)}.$$

Therefore, by the continuous mapping theorem,

$$\frac{\sum_{i=0}^{\lfloor nt \rfloor}(Y_{i,n} - m_i)}{\sqrt{\sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2}} = \sqrt{n}\frac{\sum_{i=0}^{\lfloor nt \rfloor}(Y_{i,n} - m_i)}{\sqrt{n\sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2}} \Rightarrow \text{Nor}(0,1) \text{ and } \sqrt{n}\left(\sum_{i=0}^{\lfloor nt \rfloor} Y_{i,n} - c(t)\right) \Rightarrow \text{Nor}(0, d(t)). \tag{EC.3}$$

■

We can use Lemma EC.1 to prove that W_S^n and W^n have asymptotically the same distribution, as follows.

LEMMA EC.2. *As $n \to \infty$,*

$$\sqrt{n}(W_S^n - w) \Rightarrow \text{Nor}(0, \frac{1}{\theta\mu}),$$

where $w \equiv \frac{1}{\theta}\ln(\rho)$.

Proof. We can write:

$$W_S^n = \sum_{i=0}^{Q^n}[Y_{i,n}|Y_{i,n} < T] \overset{D}{=} \sum_{i=0}^{Q^n} Y_{i,n},$$

where $T \sim \text{Exp}(\theta)$, since the rank ordering of exponentials and their minimum are independent. Now, using Lemma EC.1 and applying Theorem 6.4 of Talreja and Whitt (2009) yields the convergence. Since we also have that $\sqrt{n}(W^n - w) \Rightarrow \text{Nor}(0, \frac{1}{\theta\mu})$, W_S^n and W^n have the same distribution, asymptotically.

■

Above, we used the following lemma.

Lemma EC.3. *(Resnick, Chap. 9, problem 19) Let $U_1, U_2, \cdots, U_k$ be a sequence of independent exponential random variables with respective means $m_i, 1 \le i \le k$. If*

$$\frac{\max_{1\le i\le k} m_i^2}{\sum_{j=1}^k m_j^2} \to 0 \text{ as } k \to \infty,$$

then

$$\frac{\sum_{j=1}^k (U_j - m_j)}{\sqrt{\sum_{j=1}^k m_j^2}} \Rightarrow \mathrm{Nor}(0,1).$$

■

EC.1.3. Proof of Lemma 3

Proof. Note that $W^n(\tau^n) =^{\mathcal{D}} [W^n | S^n, A^n, E^n]$ where

- Event S^n: "customer is served"
- Event A^n: "next arrival after current entry to service is before next entry to service"
- Event E^n: "next arrival is delayed"

Part (a). For $w \ge 0$,

$$\begin{aligned}
&\mathbb{P}(\sqrt{n}W^n(\tau^n) \ge w)\\
=\ & \frac{1}{\sqrt{n}} \int_w^\infty f_{W|A,S,E}(x/\sqrt{n}|A^n, S^n, E^n)dx\\
=\ & \frac{1}{\sqrt{n}} \int_w^\infty \frac{f_{W|S,E}(x/\sqrt{n}|S^n)\mathbb{P}(A^n|W^n = x/\sqrt{n}, S^n, E^n)\mathbb{P}(S^n, E^n)}{\mathbb{P}(A^n, S^n, E^n)}dx\\
=\ & \frac{1}{\sqrt{n}} \frac{\mathbb{P}(S^n, E^n)}{\mathbb{P}(A^n, S^n, E^n)} \times\\
& [\int_w^\infty f_{W|S}(x/\sqrt{n}|S^n, E^n)\mathbb{P}(A^n|Q^n/\sqrt{n} = 0, W^n = x/\sqrt{n}, S^n, E^n)\mathbb{P}(Q^n/\sqrt{n} = 0|W^n = x/\sqrt{n}, S^n, E^n)dx\\
+\ & \int_w^\infty f_{W|S}(x/\sqrt{n}|S^n, E^n)\mathbb{P}(A^n|Q^n/\sqrt{n} > 0, W^n = x/\sqrt{n}, S^n, E^n)\mathbb{P}(Q^n/\sqrt{n} > 0|W^n = x/\sqrt{n}, S^n, E^n))dx].
\end{aligned}$$

Note that as $n \to \infty$,

$$\mathbb{P}[Q^n/\sqrt{n} = 0|W^n = x/\sqrt{n}, S^n, E^n] = \mathbb{P}[Q^n = 0|W^n = x/\sqrt{n}, S^n] = Exp\left(-\frac{\lambda^n}{\theta}\left(1 - e^{-\theta x/\sqrt{n}}\right)\right) \to 1$$

and $\mathbb{P}(A^n|Q^n/\sqrt{n} = 0, W^n = x/\sqrt{n}, S^n, E^n) = 1$. There remains to show that $\frac{1}{\sqrt{n}}\frac{\mathbb{P}(S^n,E^n)}{\mathbb{P}(A^n,S^n,E^n)} \to 1$. To obtain this,

$$\begin{aligned}
&\sqrt{n}\mathbb{P}(A^n|S^n, E^n)\\
=\ & \sqrt{n}\frac{1}{\sqrt{n}} \int_0^\infty \mathbb{P}(A^n|S^n, W^n = x/\sqrt{n}, E^n)f_W(x/\sqrt{n}|S^n, E^n)dx\\
=\ & \int_0^\infty \mathbb{P}(A^n|S^n, W^n = x/\sqrt{n}, Q^n/\sqrt{n} = 0, E^n)\mathbb{P}(Q^n/\sqrt{n} = 0|S^n, W^n = x/\sqrt{n}, E^n)f_W(x/\sqrt{n}|S^n, E^n)dx\\
+\ & \int_0^\infty \mathbb{P}(A^n|S^n, W^n = x/\sqrt{n}, Q^n/\sqrt{n} > 0, E^n)\mathbb{P}(Q^n/\sqrt{n} > 0|S^n, W^n = x/\sqrt{n}, E^n)f_W(x/\sqrt{n}|S^n, E^n)dx\\
\to\ & 1.
\end{aligned}$$

Note that we are conditioning on the probability that the *next* customer is delayed, not the current one. However, the difference between the arrival times is asymptotically negligible so that the current customer must have been delayed as well.

Part (b). Let $w \geq 0$,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\sqrt{n}W_S^n \geq w | W_S^n > 0) &= \lim_{n\to\infty} \mathbb{P}(\sqrt{n}W^n \geq w | S^n, W^n > 0) \\ &= \lim_{n\to\infty} \frac{\mathbb{P}(\sqrt{n}W^n \geq w, S^n | W^n > 0)}{\mathbb{P}(S^n | W^n > 0)} \\ &= \lim_{n\to\infty} \mathbb{P}(\sqrt{n}W^n \geq w | W^n > 0) \text{ since } \mathbb{P}(S^n) \to 1.\end{aligned}$$

■

EC.1.4. Proof of Lemma 4

Proof. We focus on delays for low-priority customers in what follows. We let γ_t^n denote the time of entry of the LES customer (of the low type) to service. For $x > 0$:

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}(W^n(\tau_t^n) \geq x) \\ &= \lim_{n\to\infty} (\mathbb{P}(W^n(\tau_t^n) \geq x | W^n(\tau_t^n) > 0)\mathbb{P}(W^n(\tau_t^n) > 0)) \\ &= \lim_{n\to\infty} \mathbb{P}(W_\infty^n \geq x | W_\infty^n > 0).\end{aligned}$$

The last step proceeds similarly to our proof for the single-class $M/M/n$ queue, so we omit the relevant details. There remains to show that $\mathbb{P}(W^n(\tau_t^n) > 0) \to 1$. For this, we resort to Lemma EC.4 (below) which implies that $\mathbb{P}(t^n - \gamma_t^n < M) \to 1$ for any $M > 0$ because convergence in distribution to a constant implies convergence in probability. This implies:

$$\begin{aligned}&\mathbb{P}(W^n(\tau_t^n) = 0) \\ &= \sum_{k=0}^{n-1} \mathbb{P}(\text{LES customers finds } k \text{ customers in service at } \tau_t^n) \\ &\leq \sum_{k=0}^{n-1} \mathbb{P}(\text{at least } n-k-1 \text{ H arrivals in } (\tau_t^n, t^n)) \\ &= \left[\sum_{k=0}^{n-1} \mathbb{P}(\text{at least } n-k-1 \text{ H arrivals in } (\tau_t^n, t^n) | t^n - \gamma^n < M)\right] \mathbb{P}(t^n - \gamma_t^n < M) \\ &+ \left[\sum_{k=0}^{n-1} \mathbb{P}(\text{at least } n-k-1 \text{ H arrivals in } (\tau_t^n, t^n) | t^n - \gamma^n \geq M)\right] \mathbb{P}(t^n - \gamma_t^n \geq M) \\ &\leq e^{-Cn} + e^{-Kn} \quad \text{for some } C > 0 \text{ by Chernoff bound and } K > 0 \text{ by proof of Lemma EC.4} \\ &\to 0,\end{aligned}$$

so that $\mathbb{P}(\text{LES customer was delayed}) = \mathbb{P}(W^n(\tau_t^n) > 0) \to 1$. ■

LEMMA EC.4. *As $n \to \infty$,*

$$t^n - \gamma_t^n \Rightarrow 0.$$

Proof. Let $\xi_t^n \equiv t^n - \gamma_t^n$ and calculate $\lim_{n\to\infty} \mathbb{P}(\xi_t^n > x)$ for $x \geq 0$. Note that there cannot be H customers in queue at time γ_t^n. We can write:

$$\mathbb{P}(\xi_t^n > x) = \sum_{i=1}^{2} \mathbb{P}(\xi_t^n > x | A_i)\mathbb{P}(A_i),$$

where Q_L denotes the number of L customers in queue and:

- $A_1 \equiv \{Q_L > 0 \text{ at } \gamma_t^n\}$; then

$$\begin{aligned}\mathbb{P}(\xi_t^n > x|A_1) &\leq \mathbb{P}(\text{at least as many H arrivals as service completions in } (0,\xi_t^n), \xi_t^n > x|A_1)\\ &\leq \mathbb{P}(\text{at least as many H arrivals as service completions in } (0,x)|A_1).\end{aligned}$$

  By a slight abuse of notation: Conditional on all servers being busy in $(0,\xi_t^n)$, and in particular in $(0,x)$: # service completions $\sim$ Poiss($n\mu x$) and # H arrivals $\sim$ Poiss($\lambda_H^n x$), where $\lambda_H^n < n\mu$. Indeed, $T^n \equiv$ (# H arrivals - # service completions) has a Skellam ($\lambda_H^n x, n\mu x$) distribution so that, by a bound on its weight at 0:

$$\mathbb{P}(\text{at least as many H arrivals as service completions in } (0,x)|A_1) = \mathbb{P}(T^n \geq 0) \leq e^{-n(\sqrt{\rho_H x}-\sqrt{x})^2} \to 0.$$

- $A_2 \equiv \{Q_L = 0 \text{ at } \gamma_t^n\}$; then

$$\begin{aligned}\mathbb{P}(\xi_t^n > x|A_2) &= \mathbb{P}(\xi_t^n > x, \text{ no L arrivals in } (0,x)|A_2)\\ &\quad + \mathbb{P}(\xi_t^n > x, \text{ at least one L arrival in } (0,x)|A_2).\end{aligned}$$

  For the first part, note that:

$$\mathbb{P}(\xi_t^n > x, \text{ no L arrivals in } (0,x)|A_2) \leq \mathbb{P}(\text{no L arrivals in } (0,x)) \leq e^{-\lambda_L^n x} \to 0.$$

  For the second part, note that:

$$\begin{aligned}&\mathbb{P}(\xi_t^n > x, \text{at least one L arrival in } (0,x)|A_2)\\ &= \mathbb{P}(\xi_t^n > x|\text{at least one L arrival in } (0,x), A_2)\\ &\times \mathbb{P}(\text{at least one L arrival in } (0,x)|A_2).\end{aligned}$$

  Now, let s^n denote the time of the first L arrival in (γ_t^n, t^n), and define the events:
  — $E^n \equiv \{\text{at least one L arrival in } (0,x), A_2\}$
  — $F^n \equiv \{\text{at least 1 new H arrival remaining in queue at } s^n\}$
  — $\bar{F}^n$ is the complement of F^n

Then, for any s^n:

$$\begin{aligned}\mathbb{P}(\xi_t^n > x|E^n) &= \mathbb{P}(\xi_t^n > x|E^n, F^n)\mathbb{P}(F^n|E^n) + \mathbb{P}(\xi_t^n > x|E^n, \bar{F}^n)\mathbb{P}(\bar{F}^n|E^n)\\ &\leq \mathbb{P}(\xi_t^n > x|E^n, F^n)\mathbb{P}(F^n|E^n)\\ &+ \mathbb{P}(\#\text{SC in } (s^n, t^n) \leq \#\text{H arrivals in } (s^n, t^n)|E^n, \bar{F}^n)\mathbb{P}(\bar{F}^n|E^n)\\ &\to 0,\end{aligned}$$

where SC denotes "service completions". This is so because $\mathbb{P}(F^n|E^n) \to 0$ and #H arrivals in (s^n, t^n) − #service completions in (s^n, t^n) has a Skellam $(\lambda_H^n x, n\mu x)$ distribution. Thus, $\mathbb{P}(\xi_t^n > x|A_2) \to 0$.

Combining the above steps, we must have that:

$$t^n - \gamma_t^n \Rightarrow 0.$$

■

EC.2. Proof of Proposition 6

Proof. Our proof for Proposition 6 makes use of a coupling argument. Before we get to the technical details, we begin by presenting the intuition behind our reasoning. The correlation in (16) can be written as $1/\bar{\rho}$, where we define $\bar{\rho} \equiv \rho_L/(1-\rho_H)$. That is, it is of the same form as the correlation expression in (13), for the overloaded single-class queue with abandonment. This suggests that, from the standpoint of low-priority customers, the system can be approximated by an overloaded single-class $M/M/n+M$ queue where a fraction $1-\rho_H$ of the available capacity is unavailable (consistently busy serving H customers). In concert with that intuition, our proof couples the original system with two bounding systems which, asymptotically, can both be approximated by overloaded single-class queues with traffic intensity $\bar{\rho}$; we then rely on a sandwiching argument to obtain the desired convergence for the correlation in (16).

Lower-bound system. We consider a system with two dedicated server pools, L and H, of respective sizes $n_L = n - \lambda_H$ and $n_H = \lambda_H$. (Here, we ignore integrality for the number of servers, which is justifiable in large systems.) Let $N_H(t)$, $N_L(t)$ be the numbers in service, and $Q_H(t)$, $Q_L(t)$ be the numbers in queue, at time t, for the H and L classes, respectively. There is sharing between the two pools as follows: An arrival of type L may only occupy a server in the H pool if all servers in the L pool are busy and $N_H(t) < n_H$, i.e., there is an idle server in the H pool and no H customers waiting for it; we assume the same condition for H customers to be served by a server in the L pool. Otherwise, L customers are served by the L pool, and H customers by the H pool. The service discipline is work-conserving, i.e., we do not allow a server to idle if there are customers waiting in line, and we use a FCFS discipline within each class. We couple arrivals in this and the original system. We assign service times to servers not to customers, and we randomly create new patience times for all waiting customers at each departure epoch (we can do so because of the exponential assumption on abandonment times); we do so identically in both systems. We initiate both systems empty. The lower-bound and original systems will have identical sample paths until a time epoch t_0 where: There is a departure from service from the L pool at t_0, and there are customers of both types waiting in queue. In the original system, an H customer must be served next since she takes priority over L customers. In the lower-bound system, an L customer must be served instead. We generate the same service time for both customers. We also regenerate the patience times of all customers in queue. Thus, the total number of customers remains identical in both systems, and only the identity of the customers in service changes. We repeat the same argument for similar subsequent epochs. In so doing, we ensure that the number of served L customers, at every point in time, is at least as high in the lower-bound system as in the original system. Therefore, the waiting time of L customers in the lower-bound system will be smaller, in a stochastic ordering sense.

Upper bound system. Once more, we consider a system with two server pools, L and H, of respective sizes $n_L = n - \lambda_H$ and $n_H = \lambda_H$. We now assume that H customers have non-preemptive priority over L customers in the L pool of servers. We also assume that L customers are never allowed in the H pool. That is, if at time t we have $N_L(t) = n_L$, $Q_L(t) > 0$, $N_H(t) < n_H$, and a departure from service occurs in the H pool, then the newly freed server waits for subsequent H customers; i.e., we allow for idling in the H pool. The original and upper bound systems have identical samples paths until a time epoch t_0 where there is an L customer waiting, all servers in the L pool are busy, no H customers are in queue, and at least one server in the H pool is idle (this can either be at a departure epoch from the H pool, or an arrival epoch for an L customer). At this point, we serve the L customer in the original system, and we keep her waiting in the upper bound system. We regenerate patience times for all customers waiting in either system, and all service times for customers in service. Proceeding as such at every subsequent such epoch guarantees that the number of L customers in queue in the original system, at every point in time, is at least as large in the upper-bound system as in the original system. Therefore, the waiting time of L customers in the upper-bound system will be larger, in a stochastic ordering sense.

Analysis in the bounding systems. We index processes in the lower-bound system by I, and in the upper-bound system by II. By the analysis above, the following holds at t^n:

$$W_I^n(t^n) \leq_{st} W^n(t^n) \leq_{st} W_{II}^n(t^n),$$

where $\leq_{st}$ denotes first-order stochastic dominance. Since the LES customer is some served customer, the following must also hold:

$$W_I^n(\tau_t^n) \leq_{st} W^n(\tau_t^n) \leq_{st} W_{II}^n(\tau_t^n).$$

This implies:

$$W_I^n(t^n) \cdot W_I^n(\tau_t^n) \leq_{st} W^n(t^n) \cdot W^n(\tau_t^n) \leq_{st} W_{II}^n(t^n) \cdot W_{II}^n(\tau_t^n),$$

and, taking expectations, we must also have:

$$\mathbb{E}[W_I^n(t^n) \cdot W_I^n(\tau_t^n)] \leq \mathbb{E}[W^n(t^n) \cdot W^n(\tau_t^n)] \leq \mathbb{E}[W_{II}^n(t^n) \cdot W_{II}^n(\tau_t^n)].$$

We now turn to our asymptotic analysis. We let all systems run long enough to reach steady state and let $n \uparrow \infty$. Consider a single-class $M/M/n_{LB} + M$ system, dedicated to L customers, which we denote by "LB". We let $n_{LB} = n_L + n_L^{1/2+\delta}$ servers, for some $\delta > 0$, and identical parameters for the L class as in our original system. Thus, the "LB" system is an overloaded queue with traffic intensity $\bar{\rho} \equiv \rho_L/(1-\rho_H)$. For large n, it is readily seen that $W_{LB}^n(t^n) \leq_{st} W_I^n(t^n)$. Similarly, we consider a single-class $M/M/n_{UB} + M$ system, dedicated to L customers, which we denote by "UB". We let $n_{UB} = n_L - n_L^{1/2+\delta'}$ servers, for some $\delta' > 0$, and identical parameters for the L class as in our original system. The "UB" system is also an overloaded queue with traffic intensity $\bar{\rho} \equiv \rho_L/(1-\rho_H)$. For large n, it is readily seen that $W_{II}^n(t^n) \leq_{st} W_{UB}^n(t^n)$. Using Whitt (2004), the following convergence holds in steady state:

$$W_{LB}^n(\infty) \Rightarrow \frac{1}{\theta}\ln(\bar{\rho}) \quad \text{and} \quad W_{UB}^n(\infty) \Rightarrow \frac{1}{\theta}\ln(\bar{\rho}), \quad \text{as} \quad n \to \infty,$$

where "$\Rightarrow$" denotes convergence in distribution. By a sandwiching argument, noting that the covariance $\mathrm{Cov}[W_I^n(t^n), W_I^n(\tau_t^n)] = \mathbb{E}[W_I^n(t^n) \cdot W_I^n(\tau_t^n)] - \mathbb{E}[W_I^n(t^n)]\mathbb{E}[W_I^n(\tau_t^n)]$, we obtain the desired:

$$r[W^n(\tau_t^n), W^n(t^n)] \to \frac{1}{\bar{\rho}} = \frac{1-\rho_H}{\rho_L} \text{ as } n \to \infty.$$

■

EC.3. Additional Numerical Results: Time-Varying Arrivals

We now consider time-varying arrival rates. This is practically important because arrival processes to service systems, in real life, typically vary significantly over time. We consider a sinusoidal arrival-rate intensity function to mimic cyclic behavior that is common in arrival processes to service systems:

$$\lambda(u) = \bar{\lambda} + \bar{\lambda}\alpha \sin(\gamma u), \text{ for } 0 \leq u < \infty \ , \tag{EC.4}$$

where $\bar{\lambda}$ is the average arrival rate and α is the relative amplitude. Given an appropriate constant staffing level, this arrival-rate function corresponds to alternating periods of underload and overload in the system. As pointed out by Eick et al. (1993), the parameters of the arrival-rate intensity function, $\lambda(u)$ in (EC.4), should be interpreted relative to the mean service time. Then, we speak of γ as the relative frequency. Table EC.1 displays values of the relative frequency as a function of the mean service time, assuming a daily cycle. For interpretation, we also will specify the associated mean service time in minutes, given a daily cycle. Small (large) values of γ correspond to slow (fast) time-variability in the arrival process, relative to the service times.

γ	Cycle length	Mean service time
0.0436	144	10 minutes
0.262	24	1 hour
1.571	4	6 hours

Table EC.1 The relative frequency is the frequency computed with measuring units so that the mean service time is equal to 1.

In Table EC.2, we consider a two-class queueing system with time-varying arrivals, and focus on low-priority customers, as before. We hold the values of ρ_H and ρ_L fixed, and vary γ to increase the frequency in the time-varying arrivals. We let the amplitude be fixed as well: $\alpha = 0.3$. We consider Markovian queues only, to focus on the effect of the time variation in the arrival rates.

Based on Ibrahim and Whitt (2011), we know that the LES prediction can be biased with time-varying arrivals, because delays then vary systematically over time. Thus, the assumptions of Proposition 1, namely that the LES prediction is unbiased and has the same variance as the steady-state delay, fail to hold. Therefore, it is not clear whether the superior performance of the WA predictor, as derived in §4.3, will continue to hold in this case. Indeed, inspecting the point estimates of the correlation in Table EC.2, we find that these estimates vary considerably with γ, even when the traffic intensities in the system are held fixed. Thus, we do not expect simple expressions for the correlations, such as those derived in §5, to continue holding with time-varying arrivals. In comparing ASE(WA) and ASE(WA-run), we find that, while these two predictors remain generally close, WA-run performs slightly better than WA, particularly when γ is large. Interestingly, we find that both WA and WA-run remain superior to both the LES announcement and EA, in almost all cases considered (the only exception is for $\gamma = 1.571$ and no abandonment, in which case EA is superior). This shows that our new WA predictor is robust to time-variation in the arrival rates as well.

$M_t/M/30$ with two classes and sinusoidal arrival rates

γ	ρ_L	ρ_H	LES	EA	WA	WA-run	EXP	Corr	$\mathbb{E}[W\|W>0]$
0	0.2	0.5	0.322	0.279	0.272	0.270	0.343	0.284	0.221
0.0436	0.2	0.5	0.773	0.905	0.746	0.714	0.903	0.637	0.724
0.262	0.2	0.5	0.668	0.646	0.588	0.587	0.671	0.450	0.549
1.571	0.2	0.5	0.384	0.313	0.316	0.310	0.392	0.183	0.271

$M_t/M/100+M$ with two classes and sinusoidal arrival rates

γ	ρ_L	ρ_H	LES	EA	WA	WA-run	EXP	Corr	$\mathbb{E}[W\|W>0]$
0	0.3	0.5	0.107	0.101	0.0946	0.0934	0.132	0.412	0.0857
0.0436	0.3	0.5	0.233	0.273	0.215	0.214	0.408	0.641	0.303
0.262	0.3	0.5	0.224	0.252	0.204	0.204	0.334	0.606	0.274
1.571	0.3	0.5	0.153	0.140	0.133	0.131	0.187	0.391	0.137

Table EC.2 Comparison of the square-root ASE's of the different predictions for low-priority customers and time-varying arrivals. We let $\alpha = 0.3$ and consider alternative values of γ.

EC.4. Additional Empirical Results

In this section, we describe additional results quantifying the performance of our alternative announcements with the small call-center data set analyzed in §6.2. In §EC.4.1, we consider an alternative EA prediction where we announce a running-average waiting time which we continuously update. That is, we do not consider a static announcement calculated out-of-sample as in the main paper. In §EC.4.2, we present yet another EA-type announcement which accounts for seasonal effects by incorporating a day-of-week effect. In §EC.4.3, we present results for tow additional call types, NE (stock exchange activity) and NW (potential new customers getting information). In §EC.4.4, we consider a data-based predictor which exploits information about the queue-length seen by a delayed customer upon arrival. We present additional tabular results in §EC.4.5.

EC.4.1. Running Average Waiting Time

In this section, we consider a continuously updated EA announcement; thus, we consider our entire data set and do not remove a sample where we compute an out-of-sample estimate for EA as we did in the main paper. We denote that new prediction by EA-C. In Table EC.3, we parallel Table 8 in §6.2: We take a closer look at performance and present data estimates for the ratios of the ASE's of our three predictors. The first sub-table corresponds to IN callers, whereas the second sub-table corresponds to low-priority PS callers. Each column in the table corresponds to days where one of the predictors yields the smallest ASE. For example, for the first column, we restrict attention to those days where the LES announcement yields the smallest ASE: For IN calls, the LES announcement yields the smallest ASE on only 4 days out of 106.

Table EC.3 shows that WA continues to have superior performance over both EA-C and the LES announcement. It is worth noting however that, unlike in Table 8, on the 4 days where the LES announcement yields superior performance over both WA and EA-C, it significantly outperforms those predictors (as can be seen from the first column of the table). Upon closer inspection, we see that, on those four days, customer delays were considerably shorter than usual: The average waiting time on those days is 57 seconds, whereas it is 140 seconds over the entire sample. (We note in passing that three our of these four days fell in the sample that we had initially removed to calculate an out-of-sample estimate of EA.) Thus, since customers have short delays on those days, both the WA and EA-C predictions perform poorly since they fail to capture that these days have unusually short delays.

IN Call Type

	EA-C wins (32.0%)	WA wins (64.2%)	LES wins (3.8%)
EA-C/winner	1	1.22	5.48
WA/winner	1.04	1	3.50
LES/winner	2.12	1.70	1

PS Call Type

	EA-C wins (13.1%)	WA wins (76.6%)	LES wins (10.2%)
EA-C/winner	1	1.49	2.38
WA/winner	1.03	1	1.59
LES/winner	2.02	3.82	1

Table EC.3 Comparison of the ASE's of EA-C, WA, and LES in August-December for IN and PS customers. For EA-C, we use a running average that is continuously updated (not calculated out of sample). In each column, we report estimates of the ratio of the ASE of the prediction in the corresponding row, relative to the ASE of the predictor in the corresponding column.

EC.4.2. Day-Of-Week Effect

In this section, we present results for yet another EA-based prediction. In particular, we account for day-of-week seasonality in calculating the EA announcement, i.e., we make a different EA announcement based on the day of week. We present detailed results in Tables EC.6 and EC.7, which we relegate to §EC.4.5. When inspecting the results of those tables, we noticed that the seasonally-adjusted EA announcement, which we denote by EA-DOW, does not perform better than the EA announcement for IN callers. To explain why that is the case, we plot Figures EC.1 and EC.2 where the curves correspond to moving averages of delays on Mondays and Wednesdays

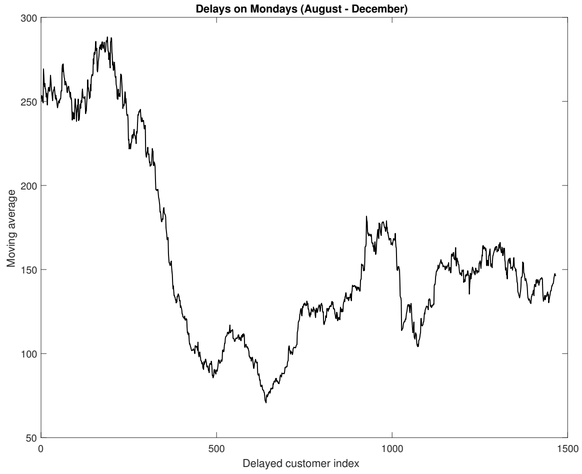

Figure EC.1 Moving average (window = 100) on Mondays.

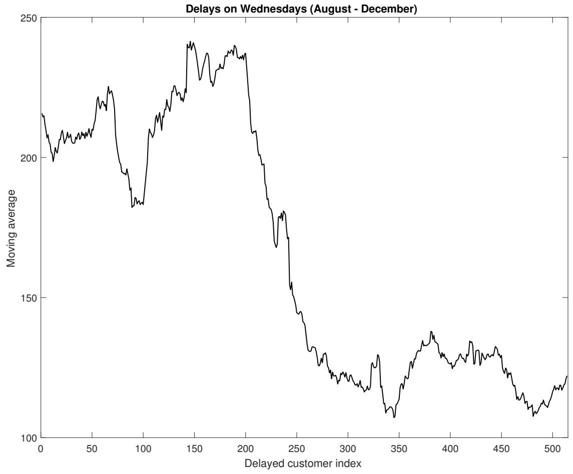

Figure EC.2 Moving average (window = 100) on Wednesdays.

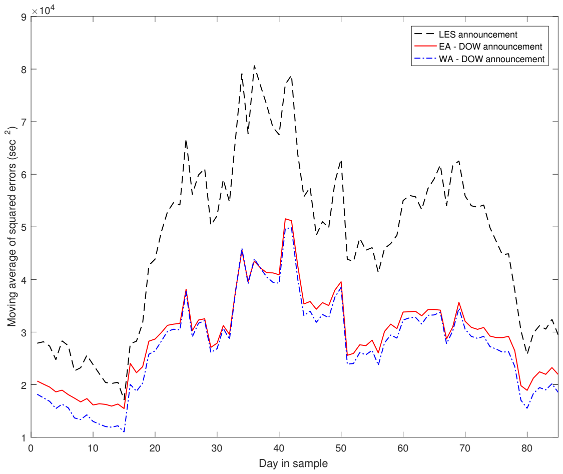

Figure EC.3 Moving average (window = 10) for squared errors of announcements.

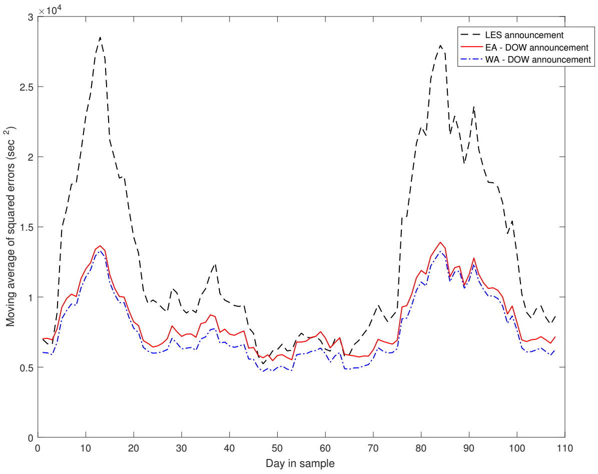

Figure EC.4 Moving average (window = 10) for squared errors of announcements.

for IN callers. In calculating the EA-DOW estimates, we take the first 2000 callers as our out of sample, which clearly correspond to longer delays, as the Figures show. Considering the overall EA average, as we did in the main paper, smoothes out those systematic changes in delays that were observed in the data (unfortunately, we do not have an explanation for what happened on those days). Nevertheless, Figures EC.3 and EC.4 show that we continue to observe the superiority of the WA-DOW prediction over both the LES announcement and EA-DOW. In these figures, we plot moving averages of the squared errors with a centered window of length 10.

EC.4.3. Results for Alternative Call Types

We include results for the NE and NW types in Table EC.4; this Table parallels Table EC.3: As before, we observe that WA is superior to the remaining predictions.

NE Call Type

	EA wins (20.0%)	WA wins (71.4%)	LES wins (8.5%)
EA/winner	1	1.35	4.69
WA/winner	1.19	1	2.23
LES/winner	2.63	1.82	1

NW Call Type

	EA wins (16.2%)	WA wins (79.0%)	LES wins (4.8%)
EA/winner	1	1.13	9.42
WA/winner	1.71	1	5.76
LES/winner	13.12	1.93	1

Table EC.4 Comparison of the ASE's of EA, WA, and LES in August-December for IN and PS customers. For EA, we use a running average that is continuously updated (not calculated out of sample). In each column, we report estimates of the ratio of the ASE of the prediction in the corresponding row, relative to the ASE of the predictor in the corresponding column.

EC.4.4. Queue-Length-Based Prediction

In this section, we consider a new data-based predictor which exploits information about the queue length seen by a delayed customer upon arrival to the system. It is well known that the MSE-minimizing prediction, conditional on the queue-length information, is the conditional expected waiting time, given that information. To estimate those conditional expectations in our data set, we resort to linear regression. Specifically, letting Q_i denote the queue-length seen upon arrival by customer i, and W_i her corresponding waiting time until either service or abandonment, we assume the following model:

$$W_i = \beta_0 + \beta_1 \times Q_i + \epsilon_i, \tag{EC.5}$$

where ϵ_i are i.i.d. normally distributed random variables with mean 0 and standard deviation σ_ϵ. We estimate the linear regression coefficients β_0 and β_1 in (EC.5) based on data for the first 2,000 customers in our data set. We restrict attention to single-class IN customers, since the waiting-time for low-priority customers (e.g., PS callers) is determined by both the queue-length seen upon arrival as well as future high-priority arrivals during the waiting time of the delayed customers. For IN callers, the estimates for β_0 and β_1 are given by:

$$\hat{\beta}_0 = 167.2 \quad \text{and} \quad \hat{\beta}_1 = 51.6.$$

That is, we define the Data-Based-Queue-Length (DB-QL) prediction for the i^{th} delayed customer as $\hat{\beta}_0 + \hat{\beta}_1 \times Q_i$. In Table EC.5, we compare the accuracy of the new DB estimator to EA, WA, and the LES announcement. Table EC.5 shows that while DB-QL is competitive, it is usually outperformed by the WA prediction.

IN Call Type

	EA wins (8.2%)	WA wins (54.1%)	LES wins (5.8%)	DB-QL wins (31.8%)
EA/winner	1	1.64	2.38	1.07
WA/winner	1.03	1	1.14	1.10
LES/winner	2.08	3.52	1	2.08
DB-QL/winner	1.06	1.56	2.25	1

Table EC.5 Comparison of the ASE's of EA, WA, LES, and DB-QL in August-December for IN customers. In each column, we report estimates of the ratio of the ASE of the prediction in the corresponding row, relative to the ASE of the predictor in the corresponding column.

EC.4.5. Detailed Tabular Results

Day index	LES	EA	EA-DOW	WA	WA-DOW
1	56989	30413	30333	32542	33205
2	9149	14570	19737	9322	12061
3	21244	17901	18061	13988	14067
4	29101	21381	22558	18790	19180
5	22898	14239	12708	12893	12328
6	29561	14403	17188	13155	14103
7	22349	16054	16171	12790	12836
8	6775	10679	12235	5189	5854
9	56844	21710	21579	24632	22952
10	19173	12449	10458	10074	9370
11	9167	18106	23629	11163	13951
12	15099	12453	12607	8886	8958
13	44859	21545	24403	21784	23097
14	11380	10315	10450	6544	6606
15	5776	13059	14910	6398	7271
16	12757	11324	16045	7203	9429
17	20390	14497	12866	12269	11573
18	9084	11520	16208	6336	8660
19	23289	12899	12995	11003	11028
20	125855	96237	96112	100463	99972
21	14161	2968	6075	170	954
22	49823	23798	24085	24543	24187
23	154173	76856	72893	80711	78190
24	23299	17270	14533	14187	12897
25	57191	25053	27055	24767	25553
26	51238	29925	29971	28429	28443
27	38426	14920	15120	15660	15529
28	4251	11966	17631	4651	7382
29	149194	76286	77503	82586	83961
30	19882	14764	17769	12556	13776

Day index	LES	EA	EA-DOW	WA	WA-DOW
31	51174	26238	26252	26889	26872
32	61803	26728	26551	29635	28926
33	46554	19895	18535	18792	17800
34	41546	20433	22460	19394	20137
35	125947	60480	60558	64087	64127
36	8402	11549	13156	6893	7658
37	162221	104087	99880	112858	108128
38	124690	90693	92573	86387	87901
39	35120	16553	18547	16280	16685
40	149411	56350	56349	61022	61017
41	12871	11591	13278	8155	9016
42	25160	14314	17467	11433	12253
43	3844	12427	18209	4905	7725
44	27726	20422	18928	18772	18367
45	221396	170461	166953	171166	167802
46	25876	9519	9580	9443	9447
47	12869	13052	14942	10040	11103
48	43232	16858	19186	16767	17713
49	52617	23327	23230	24956	25449
50	58500	39435	41608	39134	39482
51	38822	25770	25879	24029	24061
52	13208	10411	11888	5607	6157
53	91173	47784	47534	46936	46799
54	71329	36111	34857	37204	36787
55	30650	23900	27035	20902	21936
56	22084	13134	13257	10645	10688
57	56778	31018	31267	32185	31828
58	20469	14537	17599	12225	13679
59	57450	33307	33630	33045	33830
60	11077	13446	17389	9374	11226
61	84647	66742	66674	66620	66552
62	22474	24852	25807	20780	21098
63	107035	37727	39019	39615	40897
64	137283	65270	66412	70240	71105
65	40456	25570	27502	25493	26026
66	19517	14023	14134	11689	11728
67	32390	20234	22746	18438	18990
68	60359	29902	29259	30160	30430
69	75857	32682	33933	34358	34481
70	37744	16106	16192	16230	16256
71	7240	10308	11995	5973	6824
72	98209	48862	48386	45531	45191
73	116258	84542	85953	83132	84421
74	70747	29064	30358	30342	31129
75	21780	15826	15953	12197	12238
76	16756	9446	10359	7618	7961
77	36619	23854	26279	21799	22955
78	17524	14766	12963	11567	10875
79	50021	29775	31062	29769	29887
80	11892	15891	16067	10273	10354
81	9035	12558	14391	6799	7659
82	29201	17755	21226	15496	17241
83	39211	20788	19419	20408	20037
84	25180	18104	21442	14222	15804

Day index	LES	EA	EA-DOW	WA	WA-DOW
85	62359	39077	39116	40176	40155

Table EC.6: Daily ASE for all predictors from August to December for IN customers.

Day index	LES	EA	EA-DOW	WA	WA-DOW
1	3732	7223	6828	5695	5429
2	5132	6927	7383	5647	5953
3	9001	6523	5800	5980	5512
4	13441	10868	10946	9753	9759
5	3621	8200	4185	6317	3561
6	5042	7602	7173	6157	5861
7	6382	5913	6296	4881	5129
8	24498	13150	13129	12977	12934
9	63404	21849	21687	21949	21944
10	28442	15562	15450	14472	14261
11	21377	10176	9957	10360	10136
12	5992	6743	5693	5646	4889
13	31664	18537	18628	17933	17983
14	38433	17955	17784	17639	17567
15	19829	8632	8631	8782	8704
16	32876	15853	16783	15248	15736
17	18765	8886	8787	8992	8914
18	9141	9418	9734	8011	8220
19	5159	5325	4465	4371	3791
20	15636	7244	5889	7178	6100
21	7249	5808	3993	5168	4075
22	7468	5333	5233	4784	4902
23	7996	8499	8909	7271	7551
24	18617	9949	9783	9902	9767
25	8447	7760	6055	6706	5529
26	6573	6351	6081	5455	5272
27	8931	6503	6740	5986	6146
28	11781	7630	7192	7140	6822
29	2914	7489	5242	5486	3968
30	12035	8007	7787	7162	7003
31	4335	6667	7155	5201	5528
32	24707	14765	14480	13278	13017
33	4710	6525	4798	5181	4040
34	7640	6683	6386	5860	5666
35	4918	7167	7664	5705	6046
36	9188	6969	6165	6308	5769
37	6645	5676	4441	4925	4147
38	23812	16668	16819	14735	14805
39	7898	5965	6265	5370	5569
40	22735	13026	13081	12122	12066
41	11989	6594	5919	6603	6385
42	2917	7441	3260	5509	2656
43	25	13473	7082	9089	4745
44	5926	6052	2834	5231	2937
45	2783	7230	6800	5340	5053
46	8495	7480	7864	6418	6668

Day index	**LES**	**EA**	**EA-DOW**	**WA**	**WA-DOW**
47	7898	6613	5856	5861	5343
48	7426	5589	4521	5050	4369
49	3947	7236	6822	5551	5273
50	6398	6983	7424	5733	6020
51	6545	4967	4211	4460	3938
52	6842	6802	5373	5644	4661
53	5347	5296	2895	4397	2892
54	6857	6787	6526	5656	5498
55	6643	6911	7384	5772	6094
56	3697	6882	5785	5259	4519
57	8304	5363	4325	5053	4358
58	15705	14988	17133	13781	15517
59	7893	7045	6807	6165	6007
60	3423	7484	8064	5664	6058
61	6186	8079	6855	6388	5547
62	7918	7612	6151	6378	5360
63	2251	8781	6245	6328	4600
64	826	5497	2062	3808	1499
65	5184	2408	294	803	4
66	12187	9741	9570	9105	8991
67	9423	7267	7694	6552	6843
68	2372	6590	5364	4804	3982
69	9112	7702	6245	6407	5601
70	10558	8891	7514	7960	6994
71	9278	6264	6092	5776	5654
72	12634	6632	6866	6475	6653
73	6538	6881	6114	6100	5588
74	8889	8011	6566	7049	6018
75	13483	7857	7755	7634	7547
76	5408	7724	8201	6249	6570
77	4165	7912	6586	6419	5464
78	6243	5350	4454	4789	4242
79	13374	9041	9044	8510	8482
80	76935	30943	31261	28133	28253
81	9842	6659	6964	5962	6178
82	39061	14801	14921	15364	15396
83	32097	16025	17896	15633	16553
84	20959	11000	11831	11405	12047
85	7063	6785	5173	5688	4668
86	45734	20664	20818	20832	20890
87	18951	11810	11894	10765	10853
88	15405	9545	9240	9437	9244
89	7440	6126	4884	5492	4607
90	19011	10758	10770	10408	10411
91	23340	13578	13475	13254	13201
92	26642	15645	15975	15598	15817
93	10057	5137	4277	5053	4456
94	36332	16951	19512	16447	17722
95	33057	16867	16911	16258	16266
96	15172	9294	9387	8903	8972
97	5512	6802	5679	5452	4695
98	5235	6782	5226	5377	4404
99	7115	7417	5334	6303	4916
100	15914	9077	8981	8491	8411

Day index	**LES**	**EA**	**EA-DOW**	**WA**	**WA-DOW**
101	12079	8351	8636	7724	7924
102	4632	6737	3905	5386	3449
103	19149	10097	9970	9891	9793
104	12351	7504	7680	7246	7380
105	4976	5462	4581	4368	3804
106	2060	9829	8217	7142	6031
107	1238	9705	6964	6786	4917
108	11935	6064	5628	5963	5592

Table EC.7: Daily ASE for all predictors from August to December for low priority PS customers.