Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

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. We propose a novel correlation-based approach which is theoretically appealing because it allows for a comparison of the accuracies 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 queueing-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 delays 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 prevalence of average-wait-

\(^1\) http://www.bayareahospital.org/Emergency-Department-Wait-Times.aspx
\(^2\) http://appreviewtimes.com/
\(^3\) http://www.cbsa-asfc.gc.ca/bwt-taf/
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. In this paper, we compare the accuracies 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.

1.1. Comparing Static and Delay-History-Based Announcements

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. In §3, by presenting statistical evidence, based on the analysis of a call-by-call data set from an Israeli bank’s call center, we show 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.

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; e.g., see Figures 3 and 4 in Ibrahim and Whitt (2009a). 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 §6.3 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, as defined above. Importantly, our objective is to focus on reasonable, easy-to-implement announcements, and to formulate a framework that enables cross-model assessment of the announcements.
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?

1.2. The Role of Correlations
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 the accuracies of delay predictors across different queueing models. Herein lies one of the main contributions of our work: In order to study the accuracies 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
In this paper, 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.

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. 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. This is even the case in systems where customers respond to the announcements, e.g., as in Ibrahim et al. (2016), although this is not our main focus in this paper. This result also applies to announcements which are based on the queue-length information, which we do not focus on in this paper. 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 queueing model, our approach is a practically useful way to compare static and dynamic announcements.

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 and remarkably simple approximations, which depend solely on the traffic intensities in the system. 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. 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).

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. We supplement our theoretical analysis with simulation to consider more general settings. 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 has a superior accuracy.

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. We present results from an additional real-life data set in the e-companion. 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 (2018).

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 (2009a), 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 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. A closely related paper to ours is Shah et al. (2019), where the authors describe results from a simulation study which extends our results, e.g., by quantifying the impact of using future information for LES and considering an alternative Newsvendor-type error criterion. 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. Recent papers such as Yu et al. (2016) and Aksin et al. (2016) present data-driven models that capture the customer’s decision-making during his/her wait, in response to updated announcements during their waiting time. We do not consider such updated announcements in this paper.

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. (2017) 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. (2018) 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 that 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.

We also note that there is a stream of literature that studies the effect of delay announcements in the system by explicitly modelling the strategic behavior of customers in queues; e.g., we refer the reader to Guo and Zipkin (2007) and Hu et al. (2018), and references therein. While the approach in these papers is different from ours, their topic overlaps greatly with ours.

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. (2016) 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. We also note that there are average predictions proposed in Thiongane et al. (2016) that exploit the LES prediction, but that are different from the one we propose here.
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.

3.1. Definitions: EA and LES Announcements

We denote by $W_\infty$ 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 $E[W_\infty|W_\infty > 0]$. In our simulations, we use a point estimate of $E[W_\infty|W_\infty > 0]$ 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 $E[W_\infty|W_\infty > 0]$. With data, we approximate EA 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 $W_\infty \equiv [W_\infty|W_\infty > 0]$. Let $\tau_t$ denote the arrival time of the corresponding LES customer, whose delay was used in the announcement, i.e.,

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

$$ASE \equiv \frac{1}{k} \sum_{j=1}^{k} (a_j - p_j)^2,$$  \hspace{1cm} (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 mean-squared error (MSE) which is defined as the expected value of the square of the difference between the delay prediction and the virtual delay. While the MSE is useful for an assessment of average performance, it is also interesting to consider alternative error criteria which penalize over and under estimation of
delays in a non-symmetric manner. We do so in §EC.6 and demonstrate that our correlation-based framework extends to such criteria as well.

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). 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). In what follows, we focus on the IN and PS call types; we describe results for the remaining types in §EC.10.3 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 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 accuracies 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) 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).

---

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).
6 Sunday to Thursday are regular work days in Israel.
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 6, we present summary statistics for the waiting times of IN and PS callers during that same month, August 1999. Table 6 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.

<table>
<thead>
<tr>
<th>Performance measures</th>
<th>IN</th>
<th>PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected wait time $\mathbb{E}[W]$</td>
<td>135s</td>
<td>47s</td>
</tr>
<tr>
<td>Expected wait time conditional on positive wait $\mathbb{E}[W</td>
<td>W &gt; 0]$</td>
<td>185s</td>
</tr>
<tr>
<td>Probability of delay $\mathbb{P}(W &gt; 0)$</td>
<td>73%</td>
<td>41%</td>
</tr>
</tbody>
</table>

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

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 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.10.2). 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 accuracies 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 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 accuracies of the LES announcement and EA. As such, we derive structural insights which hold broadly, based on our numerics. Then, we introduce the 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 $\lambda$. 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 $\mu$; 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 $\theta$. The abandonment, service, and arrival processes are mutually independent. The traffic intensity, $\rho$, 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 $\rho$. 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 $\lambda$ to vary $\rho$. In Figures 5 and 6, we present plots of the RASE’s of EA and the LES announcement, as a function of $\rho$, 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 $\rho$, 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 $\rho$, 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 $\lambda_L$; increasing $\lambda_L$ 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 $\rho_L/\rho_H$, which measures the relative congestion due to $L$ arrivals; recalling that $\rho_H = 0.5$ is held constant, the higher $\rho_L/\rho_H$, 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 $\rho_L/\rho_H$, but is marginally more accurate than EA for intermediate values of $\rho_L/\rho_H$.

In short, Figures 5-8 paint a complex picture: The accuracies 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 “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
A correlation-based approach

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 waiting time of a new delayed customer, to whom a delay prediction is made, is distributed as $W_D$. 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. 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.
MSE corresponding to delay predictor $P$ is given by:

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

Recall that the EA prediction is equal to $\mathbb{E}[W_D]$, so that $\text{MSE}(\text{EA}) = \mathbb{E}[(W_D - \mathbb{E}[W_D])^2] = \text{Var}[W_D]$. We let $C_{W_D} \equiv \sqrt{\text{Var}[W_D]/\mathbb{E}[W_D]}$ 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 $\text{Var}[P] = \beta \text{Var}[W_D]$, where $\gamma \geq 0$ and $\beta \geq 0$, then:

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

**Proof.** We have:

$$\text{MSE}(P) \equiv \mathbb{E}[(P - W_D)^2] = \mathbb{E}[(P - \mathbb{E}[P] + \mathbb{E}[P] - \mathbb{E}[W_D] + \mathbb{E}[W_D] - W_D)^2]$$

$$= \text{Var}[P] + \text{Var}[W_D] + (\gamma - 1)^2 (\mathbb{E}[W_D])^2 - 2 \cdot r[P,W_D] \sqrt{\beta} \cdot \text{Var}[W_D]$$

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

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

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

$$\text{MSE(EA)} \leq \text{MSE}(P) \text{ if, and only if, } r[P,W_D] \leq 1/2.$$ \hfill (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. This result applies broadly, irrespective of the specific queueing model at hand, provided that some mild conditions are satisfied. This is even the case in systems where customers respond to the announcements, e.g., as in Ibrahim et al. (2016). 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, $\text{WA}(P)$, be defined as a weighted average between $P$ and EA. Specifically, for a prediction $P$ as defined in Corollary 1 and a scalar $\alpha$, let:

$$\text{WA}(P) \equiv \alpha \mathbb{E}[W_D] + (1 - \alpha)P.$$ \hfill (6)
We assume that \( E[P] = E[W_D] \) and that \( \text{Var}[W_D] = \text{Var}[P] \), and choose the form in (6) to guarantee that \( \text{WA}(P) \) is unbiased as well, i.e., \( E[\text{WA}(P)] = E[W_D] = E[P] \). We also ignore, for now, the non-negativity restriction on \( \text{WA}(P) \) in (6). We can then calculate the MSE-minimizing \( \alpha \):

\[
\text{MSE}(\text{WA}(P)) \equiv E[(\alpha E[W_D] + (1 - \alpha)P - W_D)^2],
\]

\[
= E[(\alpha E[W_D] + (1 - \alpha)P + (1 - \alpha)E[P] - (1 - \alpha)E[P] - W_D)^2],
\]

\[
= E[((E[W_D] - W_D) + (1 - \alpha)(P - E[P]))^2],
\]

\[
= [\alpha^2 - 2(1 - r[P,W_D]) \cdot \alpha + 2(1 - r[P,W_D])] \cdot \text{Var}[W_D],
\]

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

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

We note that the WA announcement as defined necessarily outperforms both LES and EA since both of these predictors are special cases of the WA announcement for \( \alpha = 0 \) and \( \alpha = 1 \), respectively. We note that if \( 0 \leq r[P,W_D] \leq 1 \), which should hold for any reasonable predictor \( P \), then the \( \text{WA}^*(P) \) 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:

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

Based on (8), we see that, relative to EA, \( \text{WA}^*(P) \) is increasingly more accurate as \( r[P,W_D] \) increases (more weight is assigned to \( P \) in (6)). Conversely, relative to \( P \), \( \text{WA}^*(P) \) 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 \( r[\text{LES},W_D] \). Because direct analysis is prohibitively difficult, we focus, in the next section, on establishing many-server heavy-traffic limits for \( r[\text{LES},W_D] \) instead.

5. Asymptotic Analysis

In this section, we derive accompanying theory to support our correlation-based framework. While the framework itself is broadly applicable, we make additional assumptions in this section to enable a tractable analysis. In particular, 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 $\lambda^n$ denote the arrival rate in system $n$, and let $\lambda^n \to \infty$ as $n \to \infty$. 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 $\rho \leq 1$. 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 $\{\lambda^n, n \geq 1\}$ satisfy

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

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 $W^n_D \equiv [W^n_\infty | W^n_\infty > 0]$. Let $\tau^n_t$ denote the arrival time of the corresponding LES customer, whose delay was announced, i.e.,

$$\tau^n_t = \sup\{s \leq t^n : \text{There is an arrival at time } s \text{ and } s + W^n(s) \leq t^n\}. \quad (10)$$

Then, $W^n(\tau^n_t)$ is equal to the LES prediction. We let $\{N(s) : s \geq 0\}$ 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:

$$r[W^n(\tau^n_t), W^n(t^n)] \to \rho \quad \text{as} \quad n \to \infty. \quad (11)$$

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 (EC.1), if $\rho \leq 0.5$, 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 (e.g., LN rather than $E_2$ service times).
<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$M/M/100$</th>
<th>$M/LN(1,1)/100$</th>
<th>$M/E_2/100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.712</td>
<td>0.687</td>
<td>0.704</td>
</tr>
<tr>
<td>±$1.30 \times 10^{-3}$</td>
<td>±$4.8 \times 10^{-4}$</td>
<td>±0.015</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.726</td>
<td>0.700</td>
<td>0.654</td>
</tr>
<tr>
<td>±$2.92 \times 10^{-4}$</td>
<td>±$9.0 \times 10^{-4}$</td>
<td>±$4.0 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.779</td>
<td>0.782</td>
<td>0.728</td>
</tr>
<tr>
<td>±$4.6 \times 10^{-4}$</td>
<td>±$3.7 \times 10^{-4}$</td>
<td>±$2.4 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>0.837</td>
<td>0.831</td>
<td>0.799</td>
</tr>
<tr>
<td>±$7.7 \times 10^{-4}$</td>
<td>±$8.4 \times 10^{-4}$</td>
<td>±$9.9 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>0.894</td>
<td>0.896</td>
<td>0.869</td>
</tr>
<tr>
<td>±$1.7 \times 10^{-4}$</td>
<td>±$3.3 \times 10^{-5}$</td>
<td>±$5.6 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.949</td>
<td>0.955</td>
<td>0.938</td>
</tr>
<tr>
<td>±$2.6 \times 10^{-4}$</td>
<td>±$1.3 \times 10^{-4}$</td>
<td>±$5.0 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>0.98</td>
<td>0.981</td>
<td>0.981</td>
<td>0.973</td>
</tr>
<tr>
<td>±$1.5 \times 10^{-4}$</td>
<td>±$1.2 \times 10^{-5}$</td>
<td>±$5.6 \times 10^{-5}$</td>
<td></td>
</tr>
</tbody>
</table>

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

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 $\lambda^n$ 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:

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

(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:

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

We note that the asymptotic behavior with and without abandonment is different, so that it is not straightforward to write the no-abandonment setting as a special case of the abandonment setting with $\theta = 0$. Indeed, it is interesting to note that the correlation is independent of the individual abandonment rate, $\theta$, i.e., of the speed at which customers abandon from the system. Moreover, the correlation is decreasing in the traffic intensity, $\rho$. We now turn to 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:

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

Combining the results in Propositions 1 and 3 allows for a simple explanation of our numerical observations in Figure 6: Based on (EC.2), 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 $\rho$ in the absence (presence) of abandonment, the relative accuracies 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).
### Table 3

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

#### 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, $\lambda$, 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 $\lambda_L^n$ and $\lambda_H^n$ 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.

**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 $\rho_H + \rho_L < 1$:

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

**Proof.** For the proof of the proposition, we need the following lemma.
Lemma 1. For all \( x \geq 0 \),
\[
\lim_{n \to \infty} \mathbb{P}(W^n(\tau^n_t) \geq x) = \lim_{n \to \infty} \mathbb{P}(W^n(t^n) \geq x).
\]

Now, proceeding as in the proof for Proposition 2:
\[
W^n(t^n) = \sum_{i=1}^{N_L(W^n(\tau^n_t))} B^n_i + Z^n,
\]
where \( B^n_i \) is the length of a busy period in an M/M/1 queue with arrival rate \( \lambda^n_H \) and service rate \( n\mu \), and \( Z^n \) is independent of \( W^n(\tau^n_t) \). 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 \( \mathbb{E}[B^n_i] = 1/(n\mu - \lambda^n_H) \); e.g., see Kleinrock (1975). Thus,
\[
\text{Cov}[W^n(t^n),W^n(\tau^n_t)] = \text{Cov} \left[ \sum_{i=1}^{N_L(W^n(\tau^n_t))} B^n_i, W^n(\tau^n_t) \right]
\nonumber
= \mathbb{E} \left[ \left( \sum_{i=1}^{N_L(W^n(\tau^n_t))} B^n_i \right) W^n(\tau^n_t) \right] - \mathbb{E} \left[ \sum_{i=1}^{N_L(W^n(\tau^n_t))} B^n_i \right] \mathbb{E}[W^n(\tau^n_t)]
\nonumber
= \frac{\rho^H_L}{1 - \rho^H_H} \text{Var}[W^n(\tau^n_t)].
\]

By Lemma 1, assuming that appropriate uniform integrability holds, we obtain that
\[
\lim_{n \to \infty} \text{Var}[W^n(\tau^n_t)] = \lim_{n \to \infty} \text{Var}[W^n(t^n)].
\]
This implies that:
\[
r[W^n(\tau^n_t),W^n(t^n)] \to \frac{\rho_L}{1 - \rho_H} \quad \text{as} \quad n \to \infty.
\]

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 \( \lambda_L \) to increase \( \rho_L \), and keep \( \rho_H = 0.5 \) constant. Table 4 illustrates that our asymptotic results in Proposition 1 are useful in describing performance in systems with a general service-time distribution as well.

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.
<table>
<thead>
<tr>
<th>$\rho_L$</th>
<th>$\frac{\rho_L}{1-\rho_H}$</th>
<th>$M/M/100$</th>
<th>$M/LN(1,1)/100$</th>
<th>$M/E_2/100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.35</td>
<td>0.7</td>
<td>0.659</td>
<td>$\pm 7.1 \times 10^{-3}$</td>
<td>$\pm 2.5 \times 10^{-3}$</td>
</tr>
<tr>
<td>0.375</td>
<td>0.75</td>
<td>0.707</td>
<td>$\pm 4.1 \times 10^{-3}$</td>
<td>$\pm 3.5 \times 10^{-3}$</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8</td>
<td>0.771</td>
<td>$\pm 2.6 \times 10^{-3}$</td>
<td>$\pm 2.2 \times 10^{-3}$</td>
</tr>
<tr>
<td>0.425</td>
<td>0.85</td>
<td>0.8324</td>
<td>$\pm 3.2 \times 10^{-3}$</td>
<td>$\pm 2.1 \times 10^{-3}$</td>
</tr>
<tr>
<td>0.45</td>
<td>0.9</td>
<td>0.892</td>
<td>$\pm 1.9 \times 10^{-3}$</td>
<td>$\pm 3.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>0.475</td>
<td>0.95</td>
<td>0.947</td>
<td>$\pm 1.8 \times 10^{-3}$</td>
<td>$\pm 2.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>0.49</td>
<td>0.98</td>
<td>0.976</td>
<td>$\pm 1.3 \times 10^{-3}$</td>
<td>$\pm 8.7 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

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 $\rho_L$.

**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 $\rho_L + \rho_H > 1$:

$$r[W^n(\tau^n), W^n(t^n)] \rightarrow \frac{1-\rho_H}{\rho_L} \quad \text{as} \quad n \rightarrow \infty. \quad (16)$$

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 $\lambda_L$ to increase $\rho_L$, 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. We close this section with a table summarizing our main theoretical results; see Table 6. In the following section, we describe results from a numerical study which substantiates and extends our theoretical analysis.

### 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.8) 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. In this section, we focus on steady-state simulations. In the online supplement, we consider the transient state as well, and comment on the speed with which the system converges to steady state, depending on the parameters in the system. There, we also consider systems where the number of servers is moderately small, and show that our asymptotic results are useful in describing performance in such systems as well.
6.1. Performance of the WA Prediction

We begin by considering generally-distributed, i.e., non-exponential, service and abandonment times. In Table 7, we consider both exponential and lognormal ($LN(1,1)$) service times for a single-class model. We vary the value of the traffic intensity, $\rho$, and present a point estimate of the ASE for each delay announcement, for that value of $\rho$. (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, 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 7 shows that both WA and WA-run consistently outperform the remaining predictors, for all values of $\rho$ 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 $\rho$ is large enough: This illustrates that there is no advantage in averaging over previous LES delays in this case. In Table 8, 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 8 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.

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 $WA \equiv (1 - \hat{r}) \cdot EA + \hat{r} \cdot LES$, 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 $EA_{IN} = 173$ seconds and $EA_{PS} = 121$ 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 and EA. We take a closer look at performance in Table 9, 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 $\text{ASE(EA)}/\text{ASE(LES)}$ (first row) and $\text{ASE(WA)}/\text{ASE(LES)}$ (second row). Based on Table 9, we can make the following observations. First, it is clear that WA outperforms both the LES announcement and EA: The proportions of

| $\rho$ | LES | EA | WA | WA-run | EXP | $\mathbb{E}[W|W > 0]$ |
|-------|-----|----|----|--------|-----|----------------|
| 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 |
| $\rho$ | LES | EA | WA | WA-run | EXP | $\mathbb{E}[W|W > 0]$ |
|-------|-----|----|----|--------|-----|----------------|
| 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 7 Comparison of the square-root ASE’s for the different predictions in the $M/G/100$ model for alternative values of the traffic intensity, $\rho$. 


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.

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).

<table>
<thead>
<tr>
<th>Ratios of (row) ASE to (column) winner ASE</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA wins (28%)</td>
<td>1.43</td>
</tr>
<tr>
<td>WA wins (66%)</td>
<td>1.02</td>
</tr>
<tr>
<td>LES wins (6%)</td>
<td>2.75</td>
</tr>
</tbody>
</table>

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

Table 9 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. Queue-Length-Based Predictions

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. In our data set, agent service rates are time-varying and the numbers of available agents vary with time. Thus, we must resort to estimating these quantities in the data to be able to approximate the expected waiting time conditional on the queue-length 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:

$$W_i = \beta_0 + \beta_1 \times Q_i + \epsilon_i,$$

(17)

where $\epsilon_i$ are i.i.d. normally distributed random variables with mean 0 and standard deviation $\sigma_{\epsilon}$.

We estimate the linear regression coefficients $\beta_0$ and $\beta_1$ in (17) 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 $\beta_0$ and $\beta_1$ 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 10, we compare the accuracy of the new DB estimator to EA, WA, and the LES announcement. Table 10 shows that while DB-QL is competitive, it is usually outperformed by the WA prediction. In particular, we find that the WA prediction outperforms DB-QL most of the time. This is likely due to the fact that the queue-length-based prediction, i.e., the conditional expected value of the waiting time, cannot be calculated in closed form in our setting, so that even though it should have a superior performance in theory, it does not have a superior performance in practice. It is important to emphasize that our intention is not to downplay the importance and relevance of considering queue-length-based predictions in general, but rather to point that alternative predictions such as the ones considered in this paper are important to consider as well, and may even outperform queue-length-based predictions in practice.

6.4. Customer Response to the Announcements: Endogenous Demand

In this section, we consider a setting where customers respond to the announcements that they receive. In particular, we assume that an arriving customer balks in response to an announcement, $w$, with probability $\beta(w)$. We do so because of empirical support that customers respond to delay announcements in practice; e.g., see Aksin et al. (2016) and Yu et al. (2016). We begin by noting that our correlation-based framework continues to apply when customers respond to the announcements.
Indeed, Proposition EC.2 does not make any assumptions about the distributions of $P$ and $W_D$ and, in particular, continues to apply to systems where customers react to the announcements. Such systems involve a complex equilibrium where the announcement and the delay experienced must coincide, but that does not matter. Additionally, it was shown in Ibrahim et al. (2016) that, under some regularity assumptions, if LES announcements are made and customers respond to the announcements, then asymptotically in large systems, under both the QED and ED regimes, the conditions of Corollary 1 hold and we have $\beta = \gamma = 1$; see Theorems 1 and 2 in Ibrahim et al. (2016). The only caveat is that, while our proposed correlation framework itself applies, deriving analytical expressions for the correlation $r[LES, W_D]$ is quite complicated when customers respond to the announcements, because the system itself is no longer Markovian.

To gain insight into expressions for correlations with customer response, we rely on simulation. In Table 11, we present point estimates of correlations and ASE’s in the $M/M/100 + M$ model with varying $\rho$ and $\mu = 1$ and $\theta = 0.5$, for two functional forms for $\beta(w)$: $\beta_1(w) = 1 - \exp(-\theta w)$ and $\beta_2(w) = w/(1+w)$. We choose these two functional forms to ensure that the regularity conditions of Ibrahim et al. (2016) are satisfied (similar functions were considered in that paper). We estimated the corresponding 95% confidence intervals and found them to be sufficiently small; thus, we do not report them in the table. In Table EC.5 in the e-companion, we consider an alternative functional form for the balking probability, $\beta_0(w)$ as specified in (EC.8), for which the regularity conditions of Ibrahim et al. (2016) do not hold. Then, the mean and variance of the LES predictions may diverge from those of the virtual waiting times, so that Corollary 1 no longer holds. Nevertheless, Proposition 1 continues to hold, as we illustrate in Table EC.5. We find that the obtained correlations remain remarkably close to $1/\rho$, consistently with our theoretical results in §5.2, where customers do not respond to the announcements. However, we do not have a proof for this at present. In the last column of Table 11, we draw attention to a possible pitfall: There, we consider $\beta_3(w) = 1 - \exp(-\theta \mathbb{E}[W|W > 0])$ where $\mathbb{E}[W|W > 0]$ denotes the expected virtual waiting time in the system (we obtain a point estimate for $\mathbb{E}[W|W > 0]$ in an initial run of the system). Since the system experiences an equilibrium where announced and experienced delays roughly coincide,
it may be tempting to assume that the system with state-dependent announcements has roughly
the same performance as when all customers are announced an equilibrium announcement, equal
to $\exp(-\theta \mathbb{E}[W|W > 0])$. However, comparing the first and third column of Table 11 illustrates
that the two systems, with state-dependent and equilibrium fixed announcements, behave quite
differently, particularly when $\rho$ is large.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$1/\rho$</th>
<th>Corr</th>
<th>LES</th>
<th>EA</th>
<th>Corr</th>
<th>LES</th>
<th>EA</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>0.83</td>
<td>0.831</td>
<td>0.00407</td>
<td>0.00972</td>
<td>0.836</td>
<td>0.00240</td>
<td>0.00451</td>
<td>0.863</td>
</tr>
<tr>
<td>1.4</td>
<td>0.71</td>
<td>0.711</td>
<td>0.00699</td>
<td>0.0117</td>
<td>0.723</td>
<td>0.00375</td>
<td>0.00608</td>
<td>0.835</td>
</tr>
<tr>
<td>1.6</td>
<td>0.63</td>
<td>0.621</td>
<td>0.00967</td>
<td>0.0125</td>
<td>0.643</td>
<td>0.00512</td>
<td>0.00669</td>
<td>0.789</td>
</tr>
<tr>
<td>1.8</td>
<td>0.56</td>
<td>0.548</td>
<td>0.0121</td>
<td>0.0131</td>
<td>0.580</td>
<td>0.00631</td>
<td>0.00706</td>
<td>0.743</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5</td>
<td>0.491</td>
<td>0.0141</td>
<td>0.0136</td>
<td>0.528</td>
<td>0.00738</td>
<td>0.00737</td>
<td>0.703</td>
</tr>
</tbody>
</table>

Table 11  Point estimates of correlations and ASE’s in the heavily-loaded $M/M/100 + M$ queue where
customers balk with probability $\beta(w)$ in response to an LES delay $w$. We let $\beta_1(w) = 1 - \exp(-\theta w)$,
$\beta_2(w) = w/(1 + w)$, and $\beta_3(w) = 1 - \exp(-\theta \mathbb{E}[W|W > 0])$.

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 accuracies
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 accuracies 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 abandon-
ment, 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. Finally, we focused here on the case where a single announcement is given to a delayed customer upon arrival. It would be interesting to consider an alternative setting where multiple announcements are given to delayed customers during their wait. Exploring this extension remains an interesting topic for future research.

References


Electronic Companion:

In this e-companion, we present supporting material to the main paper. In §??, we present the proof of Proposition 2. In §EC.2, we prove Proposition 3. In §EC.3, we prove Proposition .

In §EC.4, we prove technical lemmas ??-1. In §EC.10, we present additional empirical support which corresponds to analyzing our first call-center data set, described in §3.

EC.1. Proof of Proposition 2

Proposition EC.1. 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:

$$r[W^n(\tau^n_i), W^n(t^n)] \to \rho \quad \text{as} \quad n \to \infty.$$

(EC.1)

Proof. For the proof of the proposition, we need the following Lemma EC.1 which we prove in the following section.

Lemma EC.1. For all $x \geq 0$,

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

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

We let $\gamma^n_i$ denote the time of entry of the LES customer to service, and let $\xi^n_i \equiv t^n - \gamma^n_i$ denote the time between the entry to service of the LES customer and the new arrival time. We note that the following analysis holds irrespective of the initial state of the system; in particular, we do not need to be in steady state. We can write

$$W^n(t^n) = \sum_{i=1}^{N(t^n) - N(\gamma^n_i)} X_i = \sum_{i=1}^{N(W^n(\tau^n_i)) + N(\xi^n_i) - 1} X_i + \sum_{i=1}^{N(W^n(\tau^n_i))} X_i = \sum_{i=1}^{N(W^n(\tau^n_i))} X_i + Y^n,$$

where $X_i \sim \text{Exp}(\eta \mu)$ and $Y^n \equiv \sum_{i=1}^{N(t^n) - N(\gamma^n_i) - 1} X_i$ is independent of $W^n(\tau^n_i)$. This is so because $\gamma^n_i$ 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 $Y$:

$$\text{Cov}[W^n(\tau^n_i), W^n(t^n)] = \text{Cov} \left[ \sum_{i=1}^{N(W^n(\tau^n_i))} X_i, W^n(\tau^n_i) \right]$$

$$= \mathbb{E} \left[ \left( \sum_{i=1}^{N(W^n(\tau^n_i))} X_i \right) \cdot W^n(\tau^n_i) \right] - \mathbb{E} \left[ \left( \sum_{i=1}^{N(W^n(\tau^n_i))} X_i \right) \right] \cdot \mathbb{E}[W^n(\tau^n_i)]$$
\[
\begin{align*}
&= \mathbb{E} \left[ \mathbb{E} \left[ \left( \sum_{i=1}^{N(W^n(\tau^n))} X_i \right) \cdot W^n(\tau^n) \right] \right] \\
&\quad - \mathbb{E} \left[ \mathbb{E} \left[ \left( \sum_{i=1}^{N(W^n(\tau^n))} X_i \right) \cdot W(\tau^n) \right] \cdot \mathbb{E} \left[ W^n(\tau^n) \right] \right] \\
&= \mathbb{E} \left[ \frac{\lambda^n W(\tau^n)}{n\mu} \cdot W(\tau^n) \right] - \mathbb{E} \left[ \frac{\lambda^n W(\tau^n)}{n\mu} \right] \mathbb{E} \left[ W(\tau^n) \right] \\
&= \rho^n \mathbb{E} \left[ (W^n(\tau^n))^2 \right] - \rho^n \left( \mathbb{E} \left[ W^n(\tau^n) \right] \right)^2 = \rho^n \text{Var}[W^n(\tau^n)],
\end{align*}
\]

where we used the law of iterated expectations in the previous expressions. Therefore, we deduce that:

\[
r[W^n(\tau^n), W^n(t^n)] = \frac{\rho^n \text{Var}[W^n(\tau^n)]}{\sqrt{\text{Var}[W^n(\tau^n)] \text{Var}[W^n(t^n)]}} \rightarrow \rho \quad \text{as} \quad n \rightarrow \infty.
\]

**EC.2. Proof of Proposition 3**

**Proposition EC.2.** For the M/M/n + M model in the ED many-server heavy-traffic limiting regime, we have that:

\[
r[W^n(\tau^n), W^n(t^n)] \rightarrow \frac{1}{\rho} \quad \text{as} \quad n \rightarrow \infty.
\]

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

**Lemma EC.2.** In the ED regime,

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

\[
\lim_{n \rightarrow \infty} P(W^n(\tau^n) \geq w) = \lim_{n \rightarrow \infty} P(W^n_S \geq w),
\]

where \(W^n_S\) 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 \(n \rightarrow \infty\),

\[
\sqrt{n}(W^n_S - \bar{w}) \Rightarrow \text{Nor} \left( 0, \frac{1}{\theta \mu} \right),
\]

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:

\[
r[W^n(t^n), W^n(\tau^n)] = \frac{\text{Cov}[W^n(t^n), W^n(\tau^n)]}{\sqrt{\text{Var}[W^n(t^n)] \text{Var}[W^n(\tau^n)]}}.
\]

We have,

\[
\text{Cov}[W^n(t^n), W^n(\tau^n)] = \frac{1}{2} (\text{Var}[W^n(t^n)] + \text{Var}[W^n(\tau^n)] - \text{Var}[W^n(t) - W^n(\tau^n)])
\]

\[
= \frac{1}{2} (\text{Var}[W^n(t)] + \text{Var}[W^n(\tau^n)] - \frac{1}{2n} \text{Var}[\sqrt{n}(W^n(t) - W^n(\tau^n))]).
\]
By Theorem 4 of Ibrahim and Whitt (2009a), we have that

\[
\text{Var}[\sqrt{n}(W^n(t_n) - W^n(\tau^n_t))] \to \frac{2(\rho - 1)}{\rho \theta} \text{ as } n \to \infty.
\]

Also, we have that that \(n \text{Var}[W^n(t_n)] \to \frac{1}{\rho}\), where we use Theorem 6.4 of Talreja and Whitt (2009)\(^7\).

Lemma EC.2 implies that \(n \text{Var}[W^n(\tau^n_t)] \to \frac{1}{\rho}\). Thus,

\[

r[W^n(t_n), W^n(\tau^n_t)] = \frac{\frac{\rho}{2} \text{Var}[W^n(t_n)] + \frac{\rho}{2} \text{Var}[W^n(\tau^n_t)] - \frac{1}{2} \text{Var}[\sqrt{n}(W^n(t_n) - W^n(\tau^n_t))]}{\sqrt{n \text{Var}[W^n(t_n)] \cdot n \text{Var}[W^n(\tau^n_t)]}}
\]

\[
\to 1 - \frac{\rho - 1}{\rho}.
\]

That is,

\[
r[W^n(t_n), W^n(\tau^n_t)] \to \frac{1}{\rho} \text{ in the ED regime.}
\]

**EC.3. Proof of Proposition**

**Proposition EC.3.** For the \(M/M/n + M\) model in the many-server heavy-traffic QED regime:

\[
r[W^n(t_n), W^n(\tau^n_t)] \to 1 \text{ as } n \to \infty.
\]  \((EC.3)\)

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

**Lemma EC.3.** In the QED regime,

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

\[
\lim_{n \to \infty} \text{P}(\sqrt{n} W^n(\tau^n_t) \geq w) = \lim_{n \to \infty} \text{P}(\sqrt{n} W^n_S \geq w),
\]

where \(W^n_S\) is a random variable with the distribution of a served customer who is delayed.

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

\[
\lim_{n \to \infty} \text{P}(\sqrt{n} W^n(t_n) \geq w) = \lim_{n \to \infty} \text{P}(\sqrt{n} W^n_S \geq w).
\]

To prove the proposition, we can write:

\[
r(W^n(t_n), W^n(\tau^n_t)) = \frac{\text{Cov}[\sqrt{n} W^n(t_n), \sqrt{n} W^n(\tau^n_t)]}{\sqrt{\text{Var}[\sqrt{n} W^n(t_n)] \text{Var}[\sqrt{n} W^n(\tau^n_t)]}}.
\]

We also have that,

\[
\text{Cov}[\sqrt{n} W^n(t_n), \sqrt{n} W^n(\tau^n_t)] = \frac{1}{2} \left( \text{Var}[\sqrt{n} W^n(t_n)] + \text{Var}[\sqrt{n} W^n(\tau^n_t)] - \text{Var}[\sqrt{n}(W^n(t_n) - W^n(\tau^n_t))] \right).
\]

\(^7\) Talreja and Whitt (2009) do not condition on \(W^n(t_n) > 0\). However, \(\sqrt{n}(W^n(t_n) - \bar{w})\) and \(\sqrt{n}(W^n(t_n) - \bar{w})\) have the same distribution asymptotically in the ED regime, since \(\text{P}(W^n(t_n) > 0) \to 1\).
We know that $\sqrt{n}(W^n(\tau^n_t) - W^n(t^n)) \Rightarrow 0$ as $n \to \infty$, which follows from Theorem 1 of Ibrahim et al. (2016), so that $\text{Var}[\sqrt{n}(W^n(\tau^n_t) - W^n(t^n))] \to 0$. By Lemma EC.3, together with uniform integrability, we have:

$$\lim_{n \to \infty} \text{Var}[\sqrt{n}W^n(t^n)] = \lim_{n \to \infty} \text{Var}[\sqrt{n}W^n(\tau^n_t)].$$

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,

$$r[W^n(t^n), W^n(\tau^n_t)] \to 1$$

in the QED regime.

**EC.4. Proofs of Technical Lemmas**

**EC.4.1. Proof of Lemma EC.1**

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

$$\mathbb{P}(W^n(\tau^n_t) \geq x) = \mathbb{P}(W^n(\tau^n_t) \geq x | W^n(\tau^n_t) > 0)\mathbb{P}(W^n(\tau^n_t) > 0)$$

$$= \mathbb{P}(W^n_{\infty} \geq x | W^n_{\infty} > 0, D^n, E^n)\mathbb{P}(W^n(\tau^n_t) > 0),$$

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:

$$\mathbb{P}(W^n(\tau^n_t) > 0) = 1 - \mathbb{P}($$ LES customer finds exactly $n - 1$ 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,

$$\mathbb{P}(W^n_{\infty} \geq x | W^n_{\infty} > 0, D^n, E^n) = \frac{\mathbb{P}(E^n, D^n | W^n_{\infty} \geq x, W^n_{\infty} > 0)\mathbb{P}(W^n_{\infty} \geq x, W^n_{\infty} > 0)}{\mathbb{P}(W^n_{\infty} > 0, E^n, D^n)}$$

$$= \frac{\mathbb{P}(E^n, D^n | W^n_{\infty} \geq x)\mathbb{P}(W^n_{\infty} \geq x | W^n_{\infty} > 0)}{\mathbb{P}(E^n, D^n | W^n_{\infty} > 0)}$$

$$= \mathbb{P}(W^n_{\infty} \geq x | W^n_{\infty} > 0).$$

---

8 We establish the uniform integrability for $\{\sqrt{n}W^n(\tau^n_t), n \geq 1\}$ in the proof of Lemma EC.3.
since \( P(D^n, E^n | W^n_\infty \geq x) = P(D^n, E^n | W^n_\infty > 0) \). This is so because:

\[
P(D^n, E^n | W^n_\infty \geq x) = P(D^n | E^n, W^n_\infty \geq x) P(E^n | W^n_\infty \geq x)
\]

\[
= P(E^n | W^n_\infty \geq x) \quad \text{since} \quad P(D^n | E^n, W^n_\infty \geq x) = 1
\]

\[
= \frac{\lambda^n}{\lambda^n + n\mu}
\]

\[
= P(E^n | W^n_\infty > 0)
\]

\[
= P(D^n | E^n, W^n_\infty > 0) P(E^n | W^n_\infty > 0) \quad \text{since} \quad P(D^n | E^n, W^n_\infty > 0) = 1
\]

\[
= P(D^n, E^n | W^n_\infty > 0).
\]

Finally, letting \( \pi_{n-1} \) 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 \( n \to \infty \). 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 \( P(W^n(\tau^n_\epsilon) > 0) \to 1 \). Thus, we obtain:

\[
\lim_{n \to \infty} P(W^n(\tau^n_\epsilon) \geq x) = \lim_{n \to \infty} P(W^n(t) \geq x) = P(W_\infty \geq x | W_\infty > 0),
\]

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 \( t \):

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

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

\[
P[W^n(\tau^n_\epsilon) \geq x] = P[W^n_\infty \geq x | E^n, D^n]
\]

\[
\leq P[W^n_\infty \geq x | E^n, D^n, W^n_\infty > 0]
\]

\[
= \frac{P[W^n_\infty \geq x, E^n, D^n | W^n_\infty > 0]}{P[E^n, D^n | W^n_\infty > 0]}
\]

\[
\leq \frac{P[W^n_\infty \geq x | W^n_\infty > 0]}{\rho^n}
\]

\[
\leq \frac{P[W^n_\infty \geq x | W^n_\infty > 0]}{\rho^n + 1}
\]

\[
= P[W^n_\infty \geq x | W^n_\infty > 0] \frac{\rho^n + 1}{\rho^n}.
\]

Thus, for any \( x \geq 0 \),

\[
x \cdot P[W^n(\tau^n_\epsilon) \geq x] \leq x \cdot P[W^n_\infty \geq x | W^n_\infty > 0] \frac{\rho^n + 1}{\rho^n}.
\]
This implies,

\[
\mathbb{E}[(W^n(\tau^n))^2] = \int_0^\infty 2x \mathbb{P}[W^n(\tau^n) > x]dx \leq \frac{\rho^n + 1}{\rho^n} \int_0^\infty 2x \mathbb{P}[W^n > x|W^n > 0]dx
\]

\[
= \frac{\rho^n + 1}{\rho^n} \mathbb{E}[(W^n_\infty)^2|W^n_\infty > 0]
\]

\[
< M,
\]

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

\[\blacksquare\]

**EC.4.2. Proof of Lemma EC.2**

**Proof.** First, note that \( W^n(\tau^n)^2 \equiv [W^n|S^n, A^n, E^n] \) 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:

\[
\mathbb{P}(W^n(\tau^n) \geq w) = \int_w^\infty f_{W^n|A^n,S^n,E^n}(x|A^n,S^n,E^n)dx
\]

\[
= \int_w^\infty f_{W^n|S^n,E^n}(x|S^n,E^n)\mathbb{P}(A^n|W^n = x, S^n, E^n)\mathbb{P}(S^n, E^n)dx
\]

\[
= \frac{\mathbb{P}(S^n, E^n)}{\mathbb{P}(A^n, S^n, E^n)} \times
\]

\[
\int_w^\infty f_{W^n|S^n,E^n}(x|S^n,E^n)\mathbb{P}(A^n|Q^n = 0, W^n = x, S^n, E^n)\mathbb{P}(Q^n = 0|W^n = x, S^n, E^n)
\]

\[
+ \mathbb{P}(A^n|Q^n > 0, W^n = x, S^n, E^n)\mathbb{P}(Q^n > 0|W^n = x, S^n, E^n)dx.
\]

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 \( \frac{\lambda^n}{\mu}(1 - e^{-\theta x}) \). Thus, for \( x > 0 \),

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

Also, note that \( \mathbb{P}(A^n|Q^n > 0, W^n = x, S^n) = \frac{\lambda^n}{\lambda^n + \mu} = \frac{\rho}{\rho + 1} \). 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.

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

Thus,
\[
\lim_{n \to \infty} \mathbb{P}(W^n(\tau^n) \geq w) = \lim_{n \to \infty} \mathbb{P}(W^n(t) \geq w | S^n) = \lim_{n \to \infty} \mathbb{P}(W^n_S \geq w).
\]

We note that we can establish the uniform integrability of \( \{W^n(\tau^n), n \geq 1\} \) by the following argument, building on the above:
\[
\mathbb{P}(W^n(\tau^n) \geq w) = \mathbb{P}(W^n \geq w | A^n, S^n, E^n) \leq \frac{\mathbb{P}(W^n \geq w)}{\mathbb{P}(A^n, S^n, E^n)} \to (\rho + 1) \mathbb{P}(W^n \geq w),
\]
which follows because: \( \mathbb{P}(A^n, S^n, E^n) = \mathbb{P}(A^n | S^n, E^n) \cdot \mathbb{P}(S^n, E^n) \to \frac{\rho}{\rho + 1} \frac{1}{\rho} = \frac{1}{\rho}. \)

Thus, we deduce that \( \mathbb{E}[W^n(\tau^n)^2] = \int_0^\infty 2x \mathbb{P}(W^n(\tau^n) \geq x) \leq (\rho + 1) \mathbb{E}[(W^n)^2] < M \) for some large \( M < \infty \). Thus, uniform integrability follows.

**Part (b)** We use three lemmas.

**Lemma EC.4.** As \( n \to \infty \),
\[
\sqrt{n} \left( \sum_{i=0}^{\lfloor nt \rfloor} Y_{i,n} - c(t) \right) \Rightarrow \text{Nor}(0, d(t)),
\]
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 \( t \geq 0 \):
\[
\frac{\max_{0 \leq j \leq \lfloor nt \rfloor} m_j^2}{\sum_{j=0}^{\lfloor nt \rfloor} m_j^2} \to 0 \quad \text{as} \quad n \to \infty.
\]

By Lemma EC.6 (which applies the Lindeberg-Feller CLT), we must have that
\[
\frac{\sum_{i=0}^{\lfloor nt \rfloor} (Y_{i,n} - m_i)}{\sqrt{\sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2}} \Rightarrow \text{Nor}(0, 1).
\]

To obtain (EC.4), note that for every \( t \geq 0 \):
\[
\sum_{i=0}^{\lfloor nt \rfloor} m_i \to c(t) \equiv \frac{1}{\theta} \ln \left( 1 + \frac{\theta t}{\mu} \right) \quad \text{and} \quad n \sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2 \to d(t) \equiv \frac{t}{\mu(\mu + \theta t)}.
\]
Therefore, by the continuous mapping theorem,
\[
\frac{\sum_{i=0}^{\lfloor nt \rfloor} (Y_{i,n} - m_i)}{\sqrt{\sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2}} = \sqrt{n} \frac{\sum_{i=0}^{\lfloor nt \rfloor} (Y_{i,n} - m_i)}{\sqrt{n \sum_{i=0}^{\lfloor nt \rfloor} \sigma_i^2}} \Rightarrow \text{Nor}(0, 1) \quad \text{and} \quad \sqrt{n} \left( \sum_{i=0}^{\lfloor nt \rfloor} Y_{i,n} - c(t) \right) \Rightarrow \text{Nor}(0, d(t)).
\]
(EC.6)

We can use Lemma EC.4 to prove that \( W^n_S \) and \( W^n \) have asymptotically the same distribution, as follows.

**Lemma EC.5.** As \( n \to \infty \),
\[
\sqrt{n}(W^n_S - w) \Rightarrow \text{Nor}(0, \frac{1}{\theta \mu}),
\]
where \( w \equiv \frac{1}{\theta} \ln(\rho) \).

**Proof.** We can write:
\[
W^n_S = \sum_{i=0}^{Q^n} [Y_{i,n} | Y_{i,n} < T] \overset{D}{=} \sum_{i=0}^{Q^n} Y_{i,n},
\]
where \( T \sim \text{Exp}(\theta) \), since the rank ordering of exponentials and their minimum are independent. Now, using Lemma EC.4 and applying Theorem 6.4 of Talreja and Whitt (2009) yields the convergence. Since we also have that \( \sqrt{n}(W^n - w) \Rightarrow \text{Nor}(0, \frac{1}{\theta \mu}) \), \( W^n_S \) and \( W^n \) have the same distribution, asymptotically.

Above, we used the following lemma.

**Lemma EC.6.** (Resnick, Chap. 9, problem 19) Let \( U_1, U_2, \cdots, U_k \) be a sequence of independent exponential random variables with respective means \( m_i, 1 \leq i \leq k \). If
\[
\frac{\max_{1 \leq i \leq k} m_i^2}{\sum_{j=1}^{k} m_j^2} \to 0 \quad \text{as} \quad k \to \infty,
\]
then
\[
\frac{\sum_{j=1}^{k} (U_j - m_j)}{\sqrt{\sum_{j=1}^{k} m_j^2}} \Rightarrow \text{Nor}(0, 1).
\]

**EC.4.3. Proof of Lemma EC.3**

**Proof.** Note that \( W^n(\tau^n) = D [W^n | S^n, A^n, E^n] \) 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 \( w \geq 0 \),

\[
P(\sqrt{n} W^n(x^n) \geq w) = \frac{1}{\sqrt{n}} \int_w^\infty f_{W|\sqrt{n}S^n}(x/\sqrt{n}|A^n,S^n,E^n) \, dx
\]

and

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

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. To establish uniform integrability of the sequence \( \{\sqrt{n}W(\tau^n), n \geq 1\} \), we note that:

\[
P(\sqrt{n} W^n(\tau^n) \geq w) \leq \frac{1}{\sqrt{n}} \int_0^\infty f_{W|\sqrt{n}S^n}(x/\sqrt{n}|S^n) P(A^n| W^n = x/\sqrt{n}, S^n, E^n) P(S^n, E^n) \, dx \to P(\sqrt{n} W^n(t^n) \geq w),
\]

so that we can bound the second moment of \( \sqrt{n} W^n(\tau^n) \) as in the proof of Lemmas ?? and EC.2.

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

\[
\lim_{n \to \infty} \mathbb{P}(\sqrt{n} W^n_S \geq w | W^n_S > 0) = \lim_{n \to \infty} \mathbb{P}(\sqrt{n} W^n \geq w | S^n, W^n > 0)
\]

\[
= \lim_{n \to \infty} \frac{\mathbb{P}(\sqrt{n} W^n \geq w, S^n | W^n > 0)}{\mathbb{P}(S^n | W^n > 0)}
\]

\[
= \lim_{n \to \infty} \mathbb{P}(\sqrt{n} W^n \geq w | S^n, W^n > 0) \text{ since } \mathbb{P}(S^n) \to 1.
\]
EC.4.4. Proof of Lemma 1

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

\[
\lim_{n \to \infty} P(W^n(\tau^n_t) \geq x) = \lim_{n \to \infty} (P(W^n(\tau^n_t) \geq x | W^n(\tau^n_t) > 0) P(W^n(\tau^n_t) > 0)) = \lim_{n \to \infty} P(W^n \geq x | W^n > 0).
\]

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 $P(W^n(\tau^n_t) > 0) \to 1$. For this, we resort to Lemma EC.7 (below) which implies that $P(t^n - \gamma^n_t < M) \to 1$ for any $M > 0$ because convergence in distribution to a constant implies convergence in probability. This implies:

\[
P(W^n(\tau^n_t) = 0) = \sum_{k=0}^{n-1} P(\text{LES customers finds } k \text{ customers in service at } \tau^n_t)
\]

\[
\leq \sum_{k=0}^{n-1} P(\text{at least } n - k - 1 \text{ H arrivals in } (\tau^n_t, t^n))
\]

\[
= \sum_{k=0}^{n-1} P(\text{at least } n - k - 1 \text{ H arrivals in } (\tau^n_t, t^n) | t^n - \gamma^n_t < M) P(t^n - \gamma^n_t < M)
\]

\[
+ \sum_{k=0}^{n-1} P(\text{at least } n - k - 1 \text{ H arrivals in } (\tau^n_t, t^n) | t^n - \gamma^n_t \geq M) P(t^n - \gamma^n_t \geq M)
\]

\[
\leq e^{-Cn} + e^{-Kn} \quad \text{for some } C > 0 \text{ by Chernoff bound and } K > 0 \text{ by proof of Lemma EC.4}
\]

\[
\to 0,
\]

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

Lemma EC.7. As $n \to \infty$,

\[
t^n - \gamma^n_t \Rightarrow 0.
\]

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

\[
P(\xi^n_t > x) = \sum_{i=1}^{2} P(\xi^n_t > x | A_i) P(A_i),
\]

where $Q_L$ denotes the number of L customers in queue and:
- $A_1 \equiv \{Q_L > 0 \text{ at } \gamma^n_t\}$; then

$$
P(\xi^n_t > x | A_1) \leq P(\text{at least as many H arrivals as service completions in } (0, \xi^n_t), \xi^n_t > x | A_1)$$

$$\leq P(\text{at least as many H arrivals as service completions in } (0, x) | A_1).$$

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

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

- $A_2 \equiv \{Q_L = 0 \text{ at } \gamma^n_t\}$; then

$$P(\xi^n_t > x | A_2) = P(\xi^n_t > x, \text{ no L arrivals in } (0, x) | A_2)$$

$$+ P(\xi^n_t > x, \text{ at least one L arrival in } (0, x) | A_2).$$

For the first part, note that:

$$P(\xi^n_t > x, \text{ no L arrivals in } (0, x) | A_2) \leq P(\text{no L arrivals in } (0, x)) \leq e^{-\lambda^n_L x} \to 0.$$

For the second part, note that:

$$P(\xi^n_t > x, \text{ at least one L arrival in } (0, x) | A_2)$$

$$= P(\xi^n_t > x | \text{at least one L arrival in } (0, x), A_2)$$

$$\times P(\text{at least one L arrival in } (0, x) | A_2).$$

Now, let $s^n$ denote the time of the first L arrival in $(\gamma^n_t, t^n)$, 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 $s^n$:

$$P(\xi^n_t > x | E^n) = P(\xi^n_t > x | E^n, F^n)P(F^n | E^n) + P(\xi^n_t > x | E^n, \bar{F}^n)P(\bar{F}^n | E^n)$$

$$\leq P(\xi^n_t > x | E^n, F^n)P(F^n | E^n)$$

$$+ P(\# \text{SC in } (s^n, t^n) \leq \# \text{H arrivals in } (s^n, t^n) | E^n, \bar{F}^n)P(\bar{F}^n | E^n)$$

$$\to 0,$$

where SC denotes “service completions”. This is so because $P(F^n | E^n) \to 0$ and $\# \text{H arrivals in } (s^n, t^n) - \# \text{service completions in } (s^n, t^n)$ has a Skellam ($\lambda^n_H x, n \mu x$) distribution. Thus, $P(\xi^n_t > x | A_2) \to 0.$
Combining the above steps, we must have that:

\[ t^n - \gamma^n_t \Rightarrow 0. \]

EC.5. 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 (EC.2), 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 \( t^n \):

\[
W^n_I(t^n) \leq_{st} W^n(t^n) \leq_{st} W^n_{II}(t^n),
\]

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

\[
W^n_I(\tau^n_I) \leq_{st} W^n(\tau^n_I) \leq_{st} W^n_{II}(\tau^n_I).
\]

This implies:

\[
W^n_I(t^n) \cdot W^n_I(\tau^n_I) \leq_{st} W^n(t^n) \cdot W^n(\tau^n_I) \leq_{st} W^n_{II}(t^n) \cdot W^n_{II}(\tau^n_I),
\]

and, taking expectations, we must also have:

\[
E[W^n_I(t^n) \cdot W^n_I(\tau^n_I)] \leq E[W^n(t^n) \cdot W^n(\tau^n_I)] \leq E[W^n_{II}(t^n) \cdot W^n_{II}(\tau^n_I)].
\]

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 $W^n_{LB}(t^n) \leq s_t W^n_L(t^n)$. 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 $W^n_U(t^n) \leq s_t W^n_{UB}(t^n)$. Using Whitt (2004), the following convergence holds in steady state:

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

where “$\Rightarrow$” denotes convergence in distribution. By a sandwiching argument, noting that the covariance $\text{Cov}[W^n_l(t^n), W^n_t(\tau^n_l)] = E[W^n_l(t^n) \cdot W^n_t(\tau^n_l)] - E[W^n_l(t^n)]E[W^n_t(\tau^n_l)]$, we obtain the desired:

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

**EC.6. Alternative Error Criterion**

In this section, we demonstrate that our correlation-based framework continues to apply when considering an alternative delay criterion which penalizes over and under estimation of delays in a non-symmetric manner. Shah et al. (2019) considers an alternative Newsvendor-type error criterion. We define for a predictor $P$ the error criterion:

$$ERR(P) = aE[((P - W)^+)^2] + bE[((W - P)^+)^2].$$

**Lemma EC.8.** If $X \equiv P - W$ has a symmetric distribution around 0, then

$$ERR(P) = \frac{a + b}{2} \text{MSE}(P).$$

**Proof.**

$$ERR(P) = aE[((P - W)^+)^2] + bE[((W - P)^+)^2]$$

$$= aE\left[\frac{1}{4}((P - W) + |P - W|)^2\right] + bE\left[\frac{1}{4}((W - P) + |W - P|)^2\right]$$

$$= \frac{a}{4} [E((P - W)^2) + E(|P - W|^2)] + \frac{b}{4} [E((W - P)^2) + E(|W - P|^2)]$$

$$+ \frac{a + b}{4} [E((W - P)^2) + E(|W - P|^2)] + 2E[(P - W)|P - W|]$$

$$= \frac{a + b}{2} \text{MSE}(P) + \frac{a - b}{2} E[(P - W)|P - W|].$$
Recall that $X \equiv (P - W)$ is symmetric around 0. Thus,

$$
E[(P - W)|P - W|] = E[X|X] = E[X^2 1(X \geq 0)] - E[X^2 1(X \leq 0)] = 0.
$$

Thus,

$$
ERR(P) = \frac{a + b}{2} MSE(P),
$$

as desired.

There remains to show that the LES and EA predictors satisfy the symmetry property above. We note that this is true in heavily-loaded large systems. For supporting theory, we refer the reader to earlier papers demonstrating the asymptotic normality of the waiting time conditional on the LES prediction in Ibrahim and Whitt (2009a) and Ibrahim and Whitt (2009b) (which consider both the ED and QED regimes).

**EC.6.1. Supporting Numerical Study**

In what follows, we present numerical results which substantiate Lemma EC.8 for the LES announcement. In particular, we consider the single-class $M/M/100$ model with $\rho = 0.98$, the $M/M/100 + M$ model with $\rho = 1.4$, $\theta = 0.5$ and $\mu = 1$, the two-class $M/M/100$ model with $\rho_L = 0.75$ and $\rho_H = 0.23$, and the two-class $M/M/100 + M$ model with $\rho_L = 1.1$, $\rho_H = 0.3$, $\theta_H = \theta_L = 0.5$, and $\mu_L = \mu_H = 1$. The point estimates in Table EC.1 are based on averaging 10 independent simulation runs of length 8 million arrivals each. We consider different values of $a$ and $b$ in Lemma EC.8, calculate the new errors in the Lemma, and relate them to $ASE(LES)$, also calculated in the same simulation runs. The values of $a$ and $b$ are chosen so that $(a + b)/2 = 0.5$, i.e., we expect to see that $ERR(LES)/ASE(LES)$ is roughly equal to 0.5. We note that ASE(LES) varies slightly for different values of $a$ and $b$ because of the different seeds in the simulations. Table EC.1 illustrates that Lemma EC.8 indeed holds for the LES announcement, particularly when $a$ and $b$ are not too asymmetric.

**EC.7. 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.
Table EC.1  New error criterion for the LES announcement in Lemma EC.8, for alternative values of \(a\) and \(b\).

EC.7.0.1. 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 accuracies of our alternative predictions, we focused on delays which exceeded 5 seconds.

**EC.7.0.2. Accuracy of the WA Prediction.** We present in Table EC.2 results which parallel those that we reported in Table 9. Based on Table EC.2, 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 EC.2), 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.*

<table>
<thead>
<tr>
<th>Queue-30170</th>
<th>Ratios of (row) ASE to (column) winner ASE</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA wins (8%)</td>
<td>WA wins (74%)</td>
<td>LES wins (17%)</td>
</tr>
<tr>
<td>EA</td>
<td>1</td>
<td>2.42</td>
</tr>
<tr>
<td>WA</td>
<td>1.30</td>
<td>1</td>
</tr>
<tr>
<td>LES</td>
<td>2.03</td>
<td>1.43</td>
</tr>
</tbody>
</table>

**Table EC.2** 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.
EC.8. 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:

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

(EC.7)

where $\bar{\lambda}$ is the average arrival rate and $\alpha$ 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.7), should be interpreted relative to the mean service time. Then, we speak of $\gamma$ as the relative frequency. Table EC.3 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 $\gamma$ correspond to slow (fast) time-variability in the arrival process, relative to the service times.

<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>Cycle length</th>
<th>Mean service time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0436</td>
<td>144</td>
<td>10 minutes</td>
</tr>
<tr>
<td>0.262</td>
<td>24</td>
<td>1 hour</td>
</tr>
<tr>
<td>1.571</td>
<td>4</td>
<td>6 hours</td>
</tr>
</tbody>
</table>

Table EC.3 The relative frequency is the frequency computed with measuring units so that the mean service time is equal to 1.

In Table EC.4, we consider a two-class queueing system with time-varying arrivals, and focus on low-priority customers, as before. We hold the values of $\rho_H$ and $\rho_L$ fixed, and vary $\gamma$ 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.4, we find that these estimates vary considerably with $\gamma$, 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 $\gamma$ 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.

\begin{table}
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline
$\gamma$ & $\rho_L$ & $\rho_H$ & LES & WA & WA-run & EXP & $E[W|W > 0]$ \\
\hline
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 \\
\hline
\end{tabular}
\caption{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 $\gamma$.}
\end{table}

\textbf{EC.9. Additional Numerical Results: Customer Response}

In this section, we consider a balking function, $\beta_0(w)$, for which the assumptions of Ibrahim et al. (2016) are violated. In particular, the function $\beta_0(w)$ is both non-monotone and discontinuous at $w = 0.1$, i.e., it violates both the continuity and monotonically increasing assumptions of Ibrahim et al. (2016):

\[
\beta_0(w) = \begin{cases} 
w/(1+w) & \text{if } w \leq 0.1 \\ 
\exp(-\theta w) & \text{otherwise.} 
\end{cases} \quad (EC.8)
\]

In this case, we investigate the performances of LES and EA in the $M/M/100 + M$ model for varying values of $\rho$, and report corresponding estimates for the correlations. First, we note that the conditions of Corollary 1 are violated, i.e., we do not have that $\beta = \gamma = 1$. Nevertheless, Proposition 1 continues to hold as we illustrate in Table EC.5. There, we report point estimates of $\beta$, $\gamma$, $C_{WP}$, and $r$ as given in the proposition (we add a hat notation to estimates). We also
let \( A \equiv \left( \hat{\beta} + 1 + \left( \frac{\hat{\gamma} - 1}{\hat{C}_W} \right)^2 \right) - 2 \cdot \hat{r} [P, W_D] \sqrt{\hat{\beta}}. \) According to Proposition 1, we should have that ASE(LES) is roughly equal to \( A \times \) ASE(EA). Table EC.5 shows that this is indeed the case. We also note that Corollary 1 does not hold in this case: For example, we see that for \( \rho = 1.2 \) and \( \rho = 1.4 \), we have that the correlation estimate \( \hat{r} > 0.5 \), yet ASE(LES) > ASE(EA).

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( 1/\rho )</th>
<th>ASE(LES)</th>
<th>ASE(EA)</th>
<th>( \hat{r} )</th>
<th>( \hat{\gamma} )</th>
<th>( \hat{\beta} )</th>
<th>( \hat{C}_W )</th>
<th>( A )</th>
<th>ASE(LES)/ASE(EA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>0.833</td>
<td>0.00175</td>
<td>0.00191</td>
<td>0.566</td>
<td>0.67</td>
<td>0.499</td>
<td>0.713</td>
<td>0.913</td>
<td>0.913</td>
</tr>
<tr>
<td>1.4</td>
<td>0.714</td>
<td>0.00255</td>
<td>0.00246</td>
<td>0.533</td>
<td>0.635</td>
<td>0.442</td>
<td>0.661</td>
<td>1.038</td>
<td>1.038</td>
</tr>
<tr>
<td>1.6</td>
<td>0.625</td>
<td>0.00356</td>
<td>0.00301</td>
<td>0.493</td>
<td>0.598</td>
<td>0.421</td>
<td>0.636</td>
<td>1.181</td>
<td>1.183</td>
</tr>
<tr>
<td>1.8</td>
<td>0.556</td>
<td>0.00478</td>
<td>0.00358</td>
<td>0.445</td>
<td>0.569</td>
<td>0.420</td>
<td>0.617</td>
<td>1.332</td>
<td>1.335</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5</td>
<td>0.00623</td>
<td>0.00419</td>
<td>0.390</td>
<td>0.544</td>
<td>0.433</td>
<td>0.607</td>
<td>1.485</td>
<td>1.486</td>
</tr>
</tbody>
</table>

Table EC.5: Point estimates of ASE’s and correlations in the heavily-loaded \( M/M/100 + M \) queue where customers balk with probability \( \beta_0(w) \) in (EC.8).

**EC.10. 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.10.1, 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.10.2, we present yet another EA-type announcement which accounts for seasonal effects by incorporating a day-of-week effect. In §EC.10.3, we present results for two additional call types, NE (stock exchange activity) and NW (potential new customers getting information). In §6.3, 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.10.4.

**EC.10.1. 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.6, we parallel Table 9 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.6 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 9, 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.

<table>
<thead>
<tr>
<th>IN Call Type</th>
<th>EA-C wins (32.0%)</th>
<th>WA wins (64.2%)</th>
<th>LES wins (3.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA-C/winner</td>
<td>1</td>
<td>1.22</td>
<td>5.48</td>
</tr>
<tr>
<td>WA/winner</td>
<td>1.04</td>
<td>1</td>
<td>3.50</td>
</tr>
<tr>
<td>LES/winner</td>
<td>2.12</td>
<td>1.70</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PS Call Type</th>
<th>EA-C wins (13.1%)</th>
<th>WA wins (76.6%)</th>
<th>LES wins (10.2%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA-C/winner</td>
<td>1</td>
<td>1.49</td>
<td>2.38</td>
</tr>
<tr>
<td>WA/winner</td>
<td>1.03</td>
<td>1</td>
<td>1.50</td>
</tr>
<tr>
<td>LES/winner</td>
<td>2.02</td>
<td>3.82</td>
<td>1</td>
</tr>
</tbody>
</table>

Table EC.6 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.10.2. 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.8 and EC.9, which we relegate to §EC.10.4. 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 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.10.3. Results for Alternative Call Types

We include results for the NE and NW types in Table EC.7; this Table parallels Table EC.6: As before, we observe that WA is superior to the remaining predictions.

<table>
<thead>
<tr>
<th>NE Call Type</th>
<th>EA wins (20.0%)</th>
<th>WA wins (71.4%)</th>
<th>LES wins (8.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA/winner</td>
<td>1</td>
<td>1.35</td>
<td>4.69</td>
</tr>
<tr>
<td>WA/winner</td>
<td>1.19</td>
<td>1</td>
<td>2.23</td>
</tr>
<tr>
<td>LES/winner</td>
<td>2.63</td>
<td>1.82</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NW Call Type</th>
<th>EA wins (16.2%)</th>
<th>WA wins (79.0%)</th>
<th>LES wins (4.8%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA/winner</td>
<td>1</td>
<td>1.13</td>
<td>9.42</td>
</tr>
<tr>
<td>WA/winner</td>
<td>1.71</td>
<td>1</td>
<td>5.76</td>
</tr>
<tr>
<td>LES/winner</td>
<td>13.12</td>
<td>1.93</td>
<td>1</td>
</tr>
</tbody>
</table>

Table EC.7 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.10.4. Detailed Tabular Results

<table>
<thead>
<tr>
<th>Day index</th>
<th>LES</th>
<th>EA</th>
<th>EA-DOW</th>
<th>WA</th>
<th>WA-DOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56989</td>
<td>30413</td>
<td>30333</td>
<td>32542</td>
<td>33205</td>
</tr>
<tr>
<td>2</td>
<td>9149</td>
<td>14570</td>
<td>19737</td>
<td>9322</td>
<td>12061</td>
</tr>
<tr>
<td>3</td>
<td>21244</td>
<td>17901</td>
<td>18061</td>
<td>13988</td>
<td>14067</td>
</tr>
<tr>
<td>4</td>
<td>29101</td>
<td>21381</td>
<td>22558</td>
<td>18790</td>
<td>19180</td>
</tr>
<tr>
<td>5</td>
<td>22898</td>
<td>14239</td>
<td>12708</td>
<td>12893</td>
<td>12328</td>
</tr>
<tr>
<td>6</td>
<td>29561</td>
<td>14403</td>
<td>17188</td>
<td>13155</td>
<td>14103</td>
</tr>
<tr>
<td>7</td>
<td>22349</td>
<td>16054</td>
<td>16171</td>
<td>12790</td>
<td>12836</td>
</tr>
<tr>
<td>8</td>
<td>6775</td>
<td>10679</td>
<td>12235</td>
<td>5189</td>
<td>5854</td>
</tr>
<tr>
<td>9</td>
<td>56844</td>
<td>21710</td>
<td>21579</td>
<td>24632</td>
<td>22952</td>
</tr>
<tr>
<td>10</td>
<td>19173</td>
<td>12449</td>
<td>10458</td>
<td>10074</td>
<td>9370</td>
</tr>
<tr>
<td>11</td>
<td>9167</td>
<td>18106</td>
<td>23629</td>
<td>11163</td>
<td>13951</td>
</tr>
<tr>
<td>12</td>
<td>15099</td>
<td>12453</td>
<td>12607</td>
<td>8886</td>
<td>8958</td>
</tr>
<tr>
<td>13</td>
<td>44859</td>
<td>21545</td>
<td>24403</td>
<td>21784</td>
<td>23097</td>
</tr>
<tr>
<td>14</td>
<td>11380</td>
<td>10315</td>
<td>10450</td>
<td>6544</td>
<td>6606</td>
</tr>
<tr>
<td>15</td>
<td>5776</td>
<td>13059</td>
<td>14910</td>
<td>6398</td>
<td>7271</td>
</tr>
<tr>
<td>16</td>
<td>12757</td>
<td>11324</td>
<td>16045</td>
<td>7203</td>
<td>9429</td>
</tr>
<tr>
<td>17</td>
<td>20390</td>
<td>14497</td>
<td>12866</td>
<td>12269</td>
<td>11573</td>
</tr>
<tr>
<td>18</td>
<td>9084</td>
<td>11520</td>
<td>16208</td>
<td>6336</td>
<td>8660</td>
</tr>
<tr>
<td>19</td>
<td>23289</td>
<td>12899</td>
<td>12995</td>
<td>11003</td>
<td>11028</td>
</tr>
<tr>
<td>20</td>
<td>125855</td>
<td>96237</td>
<td>96112</td>
<td>100463</td>
<td>99972</td>
</tr>
<tr>
<td>21</td>
<td>14161</td>
<td>2968</td>
<td>6075</td>
<td>170</td>
<td>954</td>
</tr>
<tr>
<td>22</td>
<td>49823</td>
<td>23798</td>
<td>24085</td>
<td>24543</td>
<td>24187</td>
</tr>
<tr>
<td>23</td>
<td>154173</td>
<td>76856</td>
<td>72893</td>
<td>80711</td>
<td>78190</td>
</tr>
<tr>
<td>24</td>
<td>23299</td>
<td>17270</td>
<td>14533</td>
<td>14187</td>
<td>12897</td>
</tr>
<tr>
<td>Day index</td>
<td>LES</td>
<td>EA</td>
<td>EA-DOW</td>
<td>WA</td>
<td>WA-DOW</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>25</td>
<td>57191</td>
<td>25053</td>
<td>27055</td>
<td>24767</td>
<td>25553</td>
</tr>
<tr>
<td>26</td>
<td>51238</td>
<td>29925</td>
<td>29971</td>
<td>28429</td>
<td>28443</td>
</tr>
<tr>
<td>27</td>
<td>38426</td>
<td>14920</td>
<td>15120</td>
<td>15660</td>
<td>15529</td>
</tr>
<tr>
<td>28</td>
<td>4251</td>
<td>11966</td>
<td>17631</td>
<td>4651</td>
<td>7382</td>
</tr>
<tr>
<td>29</td>
<td>149194</td>
<td>76286</td>
<td>77503</td>
<td>82586</td>
<td>83961</td>
</tr>
<tr>
<td>30</td>
<td>19882</td>
<td>14764</td>
<td>17769</td>
<td>12556</td>
<td>13776</td>
</tr>
<tr>
<td>31</td>
<td>51174</td>
<td>26238</td>
<td>26252</td>
<td>26889</td>
<td>20872</td>
</tr>
<tr>
<td>32</td>
<td>61803</td>
<td>26728</td>
<td>26551</td>
<td>29635</td>
<td>28926</td>
</tr>
<tr>
<td>33</td>
<td>46554</td>
<td>19895</td>
<td>18535</td>
<td>18792</td>
<td>17800</td>
</tr>
<tr>
<td>34</td>
<td>41546</td>
<td>20433</td>
<td>22460</td>
<td>19394</td>
<td>20137</td>
</tr>
<tr>
<td>35</td>
<td>125947</td>
<td>60486</td>
<td>60558</td>
<td>64087</td>
<td>64127</td>
</tr>
<tr>
<td>36</td>
<td>8402</td>
<td>11549</td>
<td>13156</td>
<td>6893</td>
<td>7658</td>
</tr>
<tr>
<td>37</td>
<td>162221</td>
<td>104087</td>
<td>99880</td>
<td>112858</td>
<td>108128</td>
</tr>
<tr>
<td>38</td>
<td>124690</td>
<td>90693</td>
<td>92573</td>
<td>86387</td>
<td>87901</td>
</tr>
<tr>
<td>39</td>
<td>35120</td>
<td>16553</td>
<td>18547</td>
<td>16280</td>
<td>16685</td>
</tr>
<tr>
<td>40</td>
<td>149411</td>
<td>56350</td>
<td>56349</td>
<td>61022</td>
<td>61017</td>
</tr>
<tr>
<td>41</td>
<td>12871</td>
<td>11591</td>
<td>13278</td>
<td>8155</td>
<td>9016</td>
</tr>
<tr>
<td>42</td>
<td>25160</td>
<td>14314</td>
<td>17467</td>
<td>11433</td>
<td>12253</td>
</tr>
<tr>
<td>43</td>
<td>5844</td>
<td>12427</td>
<td>18209</td>
<td>4905</td>
<td>7725</td>
</tr>
<tr>
<td>44</td>
<td>27726</td>
<td>20422</td>
<td>18928</td>
<td>18772</td>
<td>18367</td>
</tr>
<tr>
<td>45</td>
<td>221396</td>
<td>170461</td>
<td>166953</td>
<td>171166</td>
<td>167802</td>
</tr>
<tr>
<td>46</td>
<td>25876</td>
<td>9519</td>
<td>9580</td>
<td>9443</td>
<td>9447</td>
</tr>
<tr>
<td>47</td>
<td>12869</td>
<td>13052</td>
<td>14942</td>
<td>10040</td>
<td>11103</td>
</tr>
<tr>
<td>48</td>
<td>43232</td>
<td>16858</td>
<td>19186</td>
<td>16767</td>
<td>17713</td>
</tr>
<tr>
<td>49</td>
<td>52617</td>
<td>23327</td>
<td>23230</td>
<td>24956</td>
<td>25449</td>
</tr>
<tr>
<td>50</td>
<td>58500</td>
<td>39435</td>
<td>41608</td>
<td>39134</td>
<td>39482</td>
</tr>
<tr>
<td>51</td>
<td>38822</td>
<td>25770</td>
<td>25879</td>
<td>24029</td>
<td>24061</td>
</tr>
<tr>
<td>52</td>
<td>13208</td>
<td>10411</td>
<td>11888</td>
<td>5067</td>
<td>6157</td>
</tr>
<tr>
<td>53</td>
<td>91173</td>
<td>47784</td>
<td>47534</td>
<td>46936</td>
<td>46799</td>
</tr>
<tr>
<td>54</td>
<td>71329</td>
<td>36111</td>
<td>34857</td>
<td>37204</td>
<td>36787</td>
</tr>
<tr>
<td>55</td>
<td>30650</td>
<td>23900</td>
<td>27035</td>
<td>20902</td>
<td>21936</td>
</tr>
<tr>
<td>56</td>
<td>22084</td>
<td>13134</td>
<td>13257</td>
<td>10645</td>
<td>10688</td>
</tr>
<tr>
<td>57</td>
<td>56778</td>
<td>31018</td>
<td>31267</td>
<td>32185</td>
<td>31828</td>
</tr>
<tr>
<td>58</td>
<td>20469</td>
<td>14537</td>
<td>17599</td>
<td>12225</td>
<td>13679</td>
</tr>
<tr>
<td>59</td>
<td>57450</td>
<td>33307</td>
<td>33630</td>
<td>33045</td>
<td>33830</td>
</tr>
<tr>
<td>60</td>
<td>11077</td>
<td>13446</td>
<td>17389</td>
<td>9374</td>
<td>11226</td>
</tr>
<tr>
<td>61</td>
<td>84647</td>
<td>66742</td>
<td>66674</td>
<td>66620</td>
<td>66552</td>
</tr>
<tr>
<td>62</td>
<td>22474</td>
<td>24852</td>
<td>25807</td>
<td>20780</td>
<td>21098</td>
</tr>
<tr>
<td>63</td>
<td>107035</td>
<td>37727</td>
<td>39019</td>
<td>39615</td>
<td>40897</td>
</tr>
<tr>
<td>64</td>
<td>137283</td>
<td>65270</td>
<td>66412</td>
<td>70240</td>
<td>71105</td>
</tr>
<tr>
<td>65</td>
<td>40456</td>
<td>25570</td>
<td>27502</td>
<td>25493</td>
<td>26026</td>
</tr>
<tr>
<td>66</td>
<td>19517</td>
<td>14023</td>
<td>14134</td>
<td>11689</td>
<td>11728</td>
</tr>
<tr>
<td>67</td>
<td>32390</td>
<td>20234</td>
<td>22746</td>
<td>18438</td>
<td>18990</td>
</tr>
<tr>
<td>68</td>
<td>60359</td>
<td>29902</td>
<td>29259</td>
<td>30160</td>
<td>30430</td>
</tr>
<tr>
<td>69</td>
<td>75857</td>
<td>32682</td>
<td>33933</td>
<td>34358</td>
<td>34481</td>
</tr>
<tr>
<td>70</td>
<td>37744</td>
<td>16106</td>
<td>16192</td>
<td>16230</td>
<td>16256</td>
</tr>
<tr>
<td>71</td>
<td>7240</td>
<td>10308</td>
<td>11995</td>
<td>5973</td>
<td>6824</td>
</tr>
<tr>
<td>72</td>
<td>98209</td>
<td>48862</td>
<td>48386</td>
<td>45531</td>
<td>45191</td>
</tr>
<tr>
<td>73</td>
<td>116258</td>
<td>84542</td>
<td>85953</td>
<td>83132</td>
<td>84421</td>
</tr>
<tr>
<td>74</td>
<td>70747</td>
<td>29064</td>
<td>30358</td>
<td>30342</td>
<td>31129</td>
</tr>
<tr>
<td>75</td>
<td>21780</td>
<td>15826</td>
<td>15953</td>
<td>12197</td>
<td>12238</td>
</tr>
<tr>
<td>76</td>
<td>16756</td>
<td>9446</td>
<td>10359</td>
<td>7618</td>
<td>7961</td>
</tr>
<tr>
<td>77</td>
<td>36619</td>
<td>23854</td>
<td>26279</td>
<td>21799</td>
<td>22955</td>
</tr>
<tr>
<td>78</td>
<td>17524</td>
<td>14766</td>
<td>12963</td>
<td>11567</td>
<td>10875</td>
</tr>
</tbody>
</table>
Table EC.8: Daily ASE for all predictors from August to December for IN customers.

<table>
<thead>
<tr>
<th>Day index</th>
<th>LES</th>
<th>EA</th>
<th>EA-DOW</th>
<th>WA</th>
<th>WA-DOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>50021</td>
<td>29775</td>
<td>31062</td>
<td>29769</td>
<td>29887</td>
</tr>
<tr>
<td>80</td>
<td>11892</td>
<td>15891</td>
<td>16067</td>
<td>10273</td>
<td>10354</td>
</tr>
<tr>
<td>81</td>
<td>9035</td>
<td>12558</td>
<td>14391</td>
<td>6799</td>
<td>7659</td>
</tr>
<tr>
<td>82</td>
<td>29211</td>
<td>17755</td>
<td>21296</td>
<td>15496</td>
<td>17241</td>
</tr>
<tr>
<td>83</td>
<td>11892</td>
<td>15891</td>
<td>16067</td>
<td>10273</td>
<td>10354</td>
</tr>
<tr>
<td>84</td>
<td>25180</td>
<td>18104</td>
<td>21442</td>
<td>14222</td>
<td>15804</td>
</tr>
<tr>
<td>85</td>
<td>62359</td>
<td>39077</td>
<td>39116</td>
<td>40176</td>
<td>40155</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Day index</th>
<th>LES</th>
<th>EA</th>
<th>EA-DOW</th>
<th>WA</th>
<th>WA-DOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3732</td>
<td>7223</td>
<td>6828</td>
<td>5695</td>
<td>5429</td>
</tr>
<tr>
<td>2</td>
<td>5132</td>
<td>6927</td>
<td>7383</td>
<td>5647</td>
<td>5953</td>
</tr>
<tr>
<td>3</td>
<td>9001</td>
<td>6523</td>
<td>5800</td>
<td>5980</td>
<td>5512</td>
</tr>
<tr>
<td>4</td>
<td>13441</td>
<td>10868</td>
<td>10946</td>
<td>9753</td>
<td>9759</td>
</tr>
<tr>
<td>5</td>
<td>3621</td>
<td>8200</td>
<td>4185</td>
<td>6317</td>
<td>3561</td>
</tr>
<tr>
<td>6</td>
<td>5042</td>
<td>7602</td>
<td>7173</td>
<td>6157</td>
<td>5861</td>
</tr>
<tr>
<td>7</td>
<td>6382</td>
<td>5913</td>
<td>6296</td>
<td>4881</td>
<td>5129</td>
</tr>
<tr>
<td>8</td>
<td>24498</td>
<td>13150</td>
<td>13129</td>
<td>12977</td>
<td>12934</td>
</tr>
<tr>
<td>9</td>
<td>63404</td>
<td>21849</td>
<td>21687</td>
<td>21949</td>
<td>21944</td>
</tr>
<tr>
<td>10</td>
<td>28442</td>
<td>15562</td>
<td>15450</td>
<td>14472</td>
<td>14261</td>
</tr>
<tr>
<td>11</td>
<td>21377</td>
<td>10176</td>
<td>9957</td>
<td>10360</td>
<td>10136</td>
</tr>
<tr>
<td>12</td>
<td>5992</td>
<td>6743</td>
<td>5693</td>
<td>5646</td>
<td>4889</td>
</tr>
<tr>
<td>13</td>
<td>31664</td>
<td>18537</td>
<td>18628</td>
<td>17933</td>
<td>17983</td>
</tr>
<tr>
<td>14</td>
<td>38433</td>
<td>17955</td>
<td>17784</td>
<td>17639</td>
<td>17567</td>
</tr>
<tr>
<td>15</td>
<td>19829</td>
<td>8632</td>
<td>8631</td>
<td>8782</td>
<td>8704</td>
</tr>
<tr>
<td>16</td>
<td>32876</td>
<td>15853</td>
<td>16783</td>
<td>15248</td>
<td>15736</td>
</tr>
<tr>
<td>17</td>
<td>18765</td>
<td>8886</td>
<td>8787</td>
<td>8992</td>
<td>8914</td>
</tr>
<tr>
<td>18</td>
<td>9141</td>
<td>9418</td>
<td>9734</td>
<td>8011</td>
<td>8220</td>
</tr>
<tr>
<td>19</td>
<td>5159</td>
<td>5325</td>
<td>4465</td>
<td>4371</td>
<td>3791</td>
</tr>
<tr>
<td>20</td>
<td>15636</td>
<td>7244</td>
<td>5889</td>
<td>7178</td>
<td>6100</td>
</tr>
<tr>
<td>21</td>
<td>7249</td>
<td>5808</td>
<td>3993</td>
<td>5168</td>
<td>4075</td>
</tr>
<tr>
<td>22</td>
<td>7468</td>
<td>5333</td>
<td>5233</td>
<td>4784</td>
<td>4902</td>
</tr>
<tr>
<td>23</td>
<td>7996</td>
<td>8499</td>
<td>8909</td>
<td>7271</td>
<td>7551</td>
</tr>
<tr>
<td>24</td>
<td>18617</td>
<td>9949</td>
<td>9783</td>
<td>9902</td>
<td>9767</td>
</tr>
<tr>
<td>25</td>
<td>8447</td>
<td>7760</td>
<td>6055</td>
<td>6706</td>
<td>5529</td>
</tr>
<tr>
<td>26</td>
<td>6573</td>
<td>6351</td>
<td>6081</td>
<td>5455</td>
<td>5272</td>
</tr>
<tr>
<td>27</td>
<td>8931</td>
<td>6503</td>
<td>6740</td>
<td>5986</td>
<td>6146</td>
</tr>
<tr>
<td>28</td>
<td>11781</td>
<td>7630</td>
<td>7192</td>
<td>7140</td>
<td>6822</td>
</tr>
<tr>
<td>29</td>
<td>2914</td>
<td>7489</td>
<td>5242</td>
<td>5486</td>
<td>3968</td>
</tr>
<tr>
<td>30</td>
<td>12035</td>
<td>8007</td>
<td>7787</td>
<td>7162</td>
<td>7003</td>
</tr>
<tr>
<td>31</td>
<td>4335</td>
<td>6667</td>
<td>7155</td>
<td>5201</td>
<td>5528</td>
</tr>
<tr>
<td>32</td>
<td>24707</td>
<td>14765</td>
<td>14480</td>
<td>13278</td>
<td>13017</td>
</tr>
<tr>
<td>33</td>
<td>4710</td>
<td>6525</td>
<td>4798</td>
<td>5181</td>
<td>4040</td>
</tr>
<tr>
<td>34</td>
<td>7640</td>
<td>6683</td>
<td>6386</td>
<td>5860</td>
<td>5666</td>
</tr>
<tr>
<td>35</td>
<td>4918</td>
<td>7167</td>
<td>7664</td>
<td>5705</td>
<td>6046</td>
</tr>
<tr>
<td>36</td>
<td>9188</td>
<td>6969</td>
<td>6165</td>
<td>6308</td>
<td>5769</td>
</tr>
<tr>
<td>37</td>
<td>6645</td>
<td>5676</td>
<td>4441</td>
<td>4925</td>
<td>4147</td>
</tr>
<tr>
<td>38</td>
<td>23812</td>
<td>16668</td>
<td>16819</td>
<td>14735</td>
<td>14805</td>
</tr>
<tr>
<td>39</td>
<td>7898</td>
<td>5965</td>
<td>6265</td>
<td>5370</td>
<td>5569</td>
</tr>
<tr>
<td>40</td>
<td>22735</td>
<td>13026</td>
<td>13081</td>
<td>12122</td>
<td>12066</td>
</tr>
<tr>
<td>Day index</td>
<td>LES</td>
<td>EA</td>
<td>EA-DOW</td>
<td>WA</td>
<td>WA-DOW</td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>41</td>
<td>11989</td>
<td>6594</td>
<td>5919</td>
<td>6603</td>
<td>6385</td>
</tr>
<tr>
<td>42</td>
<td>2917</td>
<td>7441</td>
<td>3260</td>
<td>5509</td>
<td>2656</td>
</tr>
<tr>
<td>43</td>
<td>25</td>
<td>13473</td>
<td>7082</td>
<td>9089</td>
<td>4745</td>
</tr>
<tr>
<td>44</td>
<td>5926</td>
<td>6052</td>
<td>2834</td>
<td>5231</td>
<td>2937</td>
</tr>
<tr>
<td>45</td>
<td>2783</td>
<td>7230</td>
<td>6800</td>
<td>5340</td>
<td>5053</td>
</tr>
<tr>
<td>46</td>
<td>8495</td>
<td>7480</td>
<td>7864</td>
<td>6418</td>
<td>6688</td>
</tr>
<tr>
<td>47</td>
<td>7898</td>
<td>6613</td>
<td>5856</td>
<td>5861</td>
<td>5343</td>
</tr>
<tr>
<td>48</td>
<td>7426</td>
<td>5589</td>
<td>4521</td>
<td>5050</td>
<td>4309</td>
</tr>
<tr>
<td>49</td>
<td>3947</td>
<td>7236</td>
<td>6822</td>
<td>5551</td>
<td>5273</td>
</tr>
<tr>
<td>50</td>
<td>6398</td>
<td>6983</td>
<td>7424</td>
<td>5733</td>
<td>6020</td>
</tr>
<tr>
<td>51</td>
<td>6545</td>
<td>4967</td>
<td>4211</td>
<td>4460</td>
<td>3938</td>
</tr>
<tr>
<td>52</td>
<td>6842</td>
<td>6802</td>
<td>5373</td>
<td>5644</td>
<td>4661</td>
</tr>
<tr>
<td>53</td>
<td>5347</td>
<td>5296</td>
<td>2895</td>
<td>4397</td>
<td>2892</td>
</tr>
<tr>
<td>54</td>
<td>6857</td>
<td>6787</td>
<td>6526</td>
<td>5656</td>
<td>5498</td>
</tr>
<tr>
<td>55</td>
<td>6643</td>
<td>6911</td>
<td>7384</td>
<td>5772</td>
<td>6094</td>
</tr>
<tr>
<td>56</td>
<td>3697</td>
<td>6882</td>
<td>5785</td>
<td>5259</td>
<td>4519</td>
</tr>
<tr>
<td>57</td>
<td>8304</td>
<td>5363</td>
<td>4325</td>
<td>5053</td>
<td>4358</td>
</tr>
<tr>
<td>58</td>
<td>15705</td>
<td>14988</td>
<td>17133</td>
<td>13781</td>
<td>15517</td>
</tr>
<tr>
<td>59</td>
<td>7893</td>
<td>7045</td>
<td>6807</td>
<td>6165</td>
<td>6007</td>
</tr>
<tr>
<td>60</td>
<td>3423</td>
<td>7484</td>
<td>8064</td>
<td>5664</td>
<td>6058</td>
</tr>
<tr>
<td>61</td>
<td>6186</td>
<td>8079</td>
<td>6855</td>
<td>6388</td>
<td>5547</td>
</tr>
<tr>
<td>62</td>
<td>7918</td>
<td>7612</td>
<td>6151</td>
<td>6378</td>
<td>5360</td>
</tr>
<tr>
<td>63</td>
<td>2251</td>
<td>8781</td>
<td>6245</td>
<td>6328</td>
<td>4600</td>
</tr>
<tr>
<td>64</td>
<td>826</td>
<td>5497</td>
<td>2062</td>
<td>3808</td>
<td>1499</td>
</tr>
<tr>
<td>65</td>
<td>5184</td>
<td>2408</td>
<td>294</td>
<td>803</td>
<td>4</td>
</tr>
<tr>
<td>66</td>
<td>12187</td>
<td>9741</td>
<td>9570</td>
<td>9105</td>
<td>8991</td>
</tr>
<tr>
<td>67</td>
<td>9423</td>
<td>7267</td>
<td>7694</td>
<td>6552</td>
<td>6843</td>
</tr>
<tr>
<td>68</td>
<td>2372</td>
<td>6590</td>
<td>5364</td>
<td>4804</td>
<td>3982</td>
</tr>
<tr>
<td>69</td>
<td>9112</td>
<td>7702</td>
<td>6245</td>
<td>6407</td>
<td>5601</td>
</tr>
<tr>
<td>70</td>
<td>10558</td>
<td>8891</td>
<td>7514</td>
<td>7960</td>
<td>6994</td>
</tr>
<tr>
<td>71</td>
<td>9278</td>
<td>6264</td>
<td>6092</td>
<td>5776</td>
<td>5654</td>
</tr>
<tr>
<td>72</td>
<td>12634</td>
<td>6632</td>
<td>6866</td>
<td>6475</td>
<td>6653</td>
</tr>
<tr>
<td>73</td>
<td>6538</td>
<td>6881</td>
<td>6114</td>
<td>6100</td>
<td>5588</td>
</tr>
<tr>
<td>74</td>
<td>8889</td>
<td>8011</td>
<td>6566</td>
<td>7049</td>
<td>6018</td>
</tr>
<tr>
<td>75</td>
<td>13483</td>
<td>7857</td>
<td>7755</td>
<td>7634</td>
<td>7547</td>
</tr>
<tr>
<td>76</td>
<td>5408</td>
<td>7724</td>
<td>8201</td>
<td>6249</td>
<td>6570</td>
</tr>
<tr>
<td>77</td>
<td>4165</td>
<td>7912</td>
<td>6586</td>
<td>6419</td>
<td>5464</td>
</tr>
<tr>
<td>78</td>
<td>6243</td>
<td>5350</td>
<td>4454</td>
<td>4789</td>
<td>4242</td>
</tr>
<tr>
<td>79</td>
<td>13374</td>
<td>9041</td>
<td>9044</td>
<td>8510</td>
<td>8482</td>
</tr>
<tr>
<td>80</td>
<td>76935</td>
<td>30943</td>
<td>31261</td>
<td>28133</td>
<td>28253</td>
</tr>
<tr>
<td>81</td>
<td>9842</td>
<td>6659</td>
<td>6964</td>
<td>5962</td>
<td>6178</td>
</tr>
<tr>
<td>82</td>
<td>39061</td>
<td>14801</td>
<td>14921</td>
<td>15364</td>
<td>15396</td>
</tr>
<tr>
<td>83</td>
<td>32097</td>
<td>16025</td>
<td>17896</td>
<td>15633</td>
<td>16553</td>
</tr>
<tr>
<td>84</td>
<td>20959</td>
<td>11000</td>
<td>11831</td>
<td>11405</td>
<td>12047</td>
</tr>
<tr>
<td>85</td>
<td>7063</td>
<td>6785</td>
<td>5173</td>
<td>5688</td>
<td>4668</td>
</tr>
<tr>
<td>86</td>
<td>45734</td>
<td>20664</td>
<td>20818</td>
<td>20832</td>
<td>20890</td>
</tr>
<tr>
<td>87</td>
<td>18951</td>
<td>11810</td>
<td>11894</td>
<td>10765</td>
<td>10853</td>
</tr>
<tr>
<td>88</td>
<td>15405</td>
<td>9545</td>
<td>9240</td>
<td>9437</td>
<td>9244</td>
</tr>
<tr>
<td>89</td>
<td>7440</td>
<td>6126</td>
<td>4884</td>
<td>5492</td>
<td>4607</td>
</tr>
<tr>
<td>90</td>
<td>19011</td>
<td>10758</td>
<td>10770</td>
<td>10408</td>
<td>10411</td>
</tr>
<tr>
<td>91</td>
<td>23340</td>
<td>13578</td>
<td>13475</td>
<td>13254</td>
<td>13201</td>
</tr>
<tr>
<td>92</td>
<td>26642</td>
<td>15645</td>
<td>15975</td>
<td>15598</td>
<td>15817</td>
</tr>
<tr>
<td>93</td>
<td>10057</td>
<td>5137</td>
<td>4277</td>
<td>5053</td>
<td>4456</td>
</tr>
<tr>
<td>94</td>
<td>36332</td>
<td>16951</td>
<td>19512</td>
<td>16447</td>
<td>17722</td>
</tr>
<tr>
<td>Day index</td>
<td>LES</td>
<td>EA</td>
<td>EA-DOW</td>
<td>WA</td>
<td>WA-DOW</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>95</td>
<td>33057</td>
<td>16867</td>
<td>16911</td>
<td>16258</td>
<td>16266</td>
</tr>
<tr>
<td>96</td>
<td>15172</td>
<td>9294</td>
<td>9387</td>
<td>8903</td>
<td>8972</td>
</tr>
<tr>
<td>97</td>
<td>5512</td>
<td>6802</td>
<td>5679</td>
<td>5452</td>
<td>4695</td>
</tr>
<tr>
<td>98</td>
<td>5235</td>
<td>6782</td>
<td>5226</td>
<td>5377</td>
<td>4404</td>
</tr>
<tr>
<td>99</td>
<td>7115</td>
<td>7417</td>
<td>5334</td>
<td>6303</td>
<td>4916</td>
</tr>
<tr>
<td>100</td>
<td>15914</td>
<td>9077</td>
<td>8981</td>
<td>8491</td>
<td>8411</td>
</tr>
<tr>
<td>101</td>
<td>12079</td>
<td>8351</td>
<td>8636</td>
<td>7724</td>
<td>7924</td>
</tr>
<tr>
<td>102</td>
<td>4632</td>
<td>6737</td>
<td>3905</td>
<td>5386</td>
<td>3449</td>
</tr>
<tr>
<td>103</td>
<td>19149</td>
<td>10097</td>
<td>9970</td>
<td>9891</td>
<td>9793</td>
</tr>
<tr>
<td>104</td>
<td>12351</td>
<td>7504</td>
<td>7680</td>
<td>7246</td>
<td>7380</td>
</tr>
<tr>
<td>105</td>
<td>4976</td>
<td>5462</td>
<td>4581</td>
<td>4368</td>
<td>3804</td>
</tr>
<tr>
<td>106</td>
<td>2060</td>
<td>9829</td>
<td>8217</td>
<td>7142</td>
<td>6031</td>
</tr>
<tr>
<td>107</td>
<td>1238</td>
<td>9705</td>
<td>6964</td>
<td>6786</td>
<td>4917</td>
</tr>
<tr>
<td>108</td>
<td>11935</td>
<td>6064</td>
<td>5628</td>
<td>5963</td>
<td>5592</td>
</tr>
</tbody>
</table>

Table EC.9: Daily ASE for all predictors from August to December for low priority PS customers.