Does the Past Predict the Future? The Case of Delay Announcements in Service Systems

Rouba Ibrahim

School of Management, University College London, London WC1E 6BT

rouba.ibrahim@ucl.ac.uk

Mor Armony

Department of Information, Operations, and Management Sciences, Stern School of Business, New York University, NY 10012

marmony@stern.nyu.edu

Achal Bassamboo

Department of Managerial Economics and Decision Sciences, Kellogg School of Management, IL 60208

a-bassamboo@kellogg.northwestern.edu

Motivated by the recent interest in making delay announcements in large service systems, such as call centers,
we investigate the accuracy of announcing the waiting time of the Last customer to Enter Service (LES). In
practice, customers typically respond to delay announcements by either balking or by becoming more or less
impatient, and their response alters system performance. We study the accuracy of the LES announcement
in single-class multi-server Markovian queueing models with announcement-dependent customer behavior.
We show that, interestingly, even in this stylized setting, the LES announcement *may not* always be accu-
rate. This motivates the need to study its accuracy carefully, and to determine conditions under which it is
accurate. Since the direct analysis of the system with customer response is prohibitively difficult, we focus
on many-server heavy-traffic analysis instead. We consider the quality-and-efficiency-driven (QED) and the
efficiency-driven (ED) many-server heavy-traffic regimes and prove, under both regimes, that the LES pre-
diction is asymptotically accurate if, and only if, asymptotic fluctuations in the queue length process are
small as long as some regulatory conditions apply. This result provides an easy check for the accuracy of LES
in practice. We supplement our theoretical results with an extensive simulation study to generate practical
managerial insights.

Key words: delay prediction; delay announcements; call centers; many-server queues; heavy traffic

1. Introduction

We study the problem of making accurate delay announcements in large service systems where
customer behavior is affected by the announcements. Delay announcements are especially helpful
in settings where customers cannot observe the current state of the system. This is typically true

with unobservable queues, such as in telephone call centers. Delay announcements are also useful in other service settings. For example, they are instrumental in hospital emergency departments where queues may be observable, yet patients often lack the experience and knowledge needed to estimate their own delays; see Plambeck et al. (2015). Because it is useful to have a specific context in mind, we will generally focus here on call centers; e.g., see Aksin et al. (2007) for background.

1.1. Delay Announcements

System managers typically use delay announcements as a relatively inexpensive way of alleviating customer uncertainty about upcoming delays, thereby increasing the level of customer satisfaction with the service provided. Additionally, delay announcements have been shown to strongly impact customer behavior. For example, information about long upcoming delays may induce some customers to balk (hang up immediately). Customers who do not balk may change their abandonment behavior, depending on the delay information. Since delay announcements typically impact customer behavior, they may be used as levers of control in the system. For example, delay announcements may be used in a highly congested system to encourage the most impatient customers to balk or abandon, thereby decreasing the number of callers on hold, and reducing system congestion; e.g., see Whitt (1999a, b), Guo and Zipkin (2007), and Armony et al. (2009).

In this paper, we assume that delay announcements are made to customers upon arrival to the system. In order to make those announcements, we need effective ways of accurately predicting, in real time, the waiting times of delayed customers. We contend that making accurate announcements is important because inaccurate delay information may cause frustration for customers.

We focus on the last-to-enter-service (LES) delay announcement. The LES prediction is equal to the waiting time of the last customer to have entered service prior to the arrival time of the new delayed customer. The LES customer is the one who experienced the LES delay. For a detailed discussion of the LES announcement, see Ibrahim and Whitt (2009). We study the accuracy of the LES announcement in models with customer response. In particular, we assume that an arriving customer may balk upon arrival with a given probability, depending on the announcement. If he does not balk, then he may subsequently abandon the queue before receiving service, and his abandonment behavior is also dependent on the announcement. To the best of our knowledge, there are no studies of how customer response to individual delay announcements impacts the accuracy of these announcements. In this paper, we take a step towards filling that gap in the literature.

1.2. Customer Response

In systems with no customer response, the LES announcement was shown to be remarkably accurate, albeit under steady-state conditions only; see Ibrahim and Whitt (2009). When customers respond to the announcements, their behavior alters the performance of the system which, in turn,

affects the future delay announcements given. Therefore, studying customer response requires an equilibrium analysis of the system. However, it is not clear, a priori, when and whether such an equilibrium exists; there may even be multiple equilibria. Moreover, even if a unique equilibrium can be shown to exist, it is not clear how stochastic fluctuations around that equilibrium will impact the accuracy of the individual LES announcements. Thus, analyzing systems with customer response entails a complicated analysis. Herein lies the main technical contribution of this paper.

In Figure 1, we illustrate the main complexity in incorporating customer response to the announcements. We plot simulation sample paths of actual delays and LES predictions in an $M/M/N + M$ queueing model; see §3 for a description of this model. We let $N = 10,000$; we deliberately choose such a large number of servers to minimize the effect of stochastic noise in the system (however, a smaller number of servers, e.g., $N = 100$, also leads to similar results but the corresponding figures are not as clear). In the first subplot of Figure 1, we assume that customers do not respond to the announcements. In the second subplot, we assume that customers respond according to a linear abandonment-rate function; in the third subplot, we assume that customers respond according to a discontinuous abandonment-rate function (the specific functional forms of those abandonment-rate functions do not matter here and are therefore omitted). We choose system parameters so as to hold the average waiting time approximately constant across our three models. Clearly, system dynamics are very similar in the first and second subplots, but are very different in the third subplot. (Since all parameters are held constant across the three graphs except for the functional form of customer response, the change in system dynamics is due to this difference in customer response.) Indeed, actual delays and LES announcements closely match in the first two subplots, but are evidently out of sync in the third (with larger fluctuations as well). In particular, since the abandonment-rate function in the last subplot is discontinuous, small fluctuations around the point of discontinuity drive the abandonment behavior, and the waiting times in the system, to vary substantially in short time intervals. As a result, the two curves, corresponding to the LES and actual delays, are out-of-sync in the plot. As such, Figure 1 illustrates that system dynamics are intimately tied to whether and how customers respond to the announcements.

The accuracy of the LES announcements also depends on customer response. Indeed, the first and second subplots of Figure 1 illustrate the asymptotic accuracy of the LES announcements. Stochastic fluctuations, due to the randomness in the system, imply that the LES announcements are not exactly equal to actual delays; nevertheless, the resulting errors are of a small magnitude (this will be made more precise later). However, the third subplot of Figure 1 clearly illustrates that the LES announcement is not accurate, and consistently fluctuates between cycles of overestimation and underestimation of actual delays. This substantiates the need to formulate conditions under which the LES announcement will be accurate in systems with customer response, which is what

4

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

we do in this paper. This lies in contrast to systems without customer response, where the accuracy of the LES announcement was shown to hold in steady state irrespective of specific assumptions on system parameters; see Ibrahim and Whitt (2009).
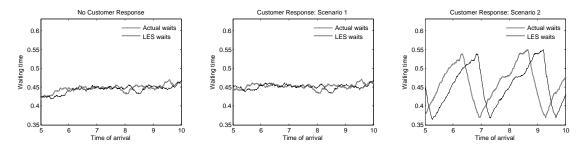

Figure 1 **Impact of customer response on the accuracy of the LES announcement.**

1.3. Asymptotic Regimes

In this paper, we investigate the accuracy of the LES delay announcement in a Markovian queueing model. Even though our modeling framework is relatively simple, explicit analysis of the underlying birth and death (BD) process is analytically complex. This is because balking probabilities and abandonment rates are all dependent on the announcements. Indeed, computing the transition rates of the BD process requires, at the minimum, keeping track of all customers in queue, and their respective announcements. Thus, instead of doing direct analysis, we focus on establishing many-server heavy-traffic limits which provide useful insights. In this paper, we focus on two such regimes: (i) The Quality-and-Efficiency-Driven (QED) or Halfin-Whitt regime (Halfin and Whitt 1981; Garnett et al. 2002), and (ii) the Efficiency-Driven (ED) regime (Whitt 2004).

The QED regime is particularly useful in describing large well-managed systems because it strikes a balance between service quality and operational efficiency. Even though waiting times in the QED regime are asymptotically small, studying the asymptotic accuracy of the LES announcement remains of practical importance in that setting. Indeed, the specific time scale under consideration is critical. For example, operating a hospital ward in the QED regime involves lengths of stay that are in the order of days, and waiting times that are in the order of hours; see Armony et al. (2015). As such, although waiting times are "small" compared to service times, predicting them accurately remains essential. The ED regime is useful in describing highly congested systems where customer waiting times tend to be long (in the order of service times), and virtually all customers are delayed before receiving service; see Whitt (2004). Delay announcements are especially important with such long waiting times. Through our asymptotic analysis in both regimes, we establish the *relative accuracy* of the LES announcement. By relative accuracy, we mean the difference between the LES

and actual waits, scaled by the appropriate order of magnitude of delays in the system. Since the asymptotic magnitude of waiting times in the ED regime is drastically different from the QED regime, the scaling that we use differs depending on the particular regime considered.

1.4. Main Insights and Contributions

1.4.1. A Result of Practical Importance.
In both the QED and ED regimes, we establish an important asymptotic result which unifies our analysis throughout: *The relative error in the LES prediction is small if, and only if, the relative error in the queue length is small*; e.g., see Theorems 1 and 2. By relative error in the queue length, we mean the difference in the queue lengths seen upon arrival by the LES customer and the newly arriving customer (to whom the announcement is made), scaled by the order of magnitude of the queue length in the system. We emphasize that our result concerns the experience of *individual* customers in the system; thus, it is stronger than a general result relating wait-time and queue-length *averages or distributions*, such as Little's law; e.g., see Little and Graves (2008) and Bertsimas and Nakazato (1995).

Our result provides a quick and easy check for the accuracy of the LES announcement in practice. At a high level, to be made more precise later, our result implies that the LES announcement will be accurate if the relative difference between the queue lengths seen upon arrival by the LES and newly arriving customers is not too large. Therefore, it is possible to check at the arrival epoch of a new customer (which is also the announcement epoch) whether or not the waiting time that he is about to experience will be close to the LES delay. In practice, a system manager may use this result to decide when to make LES delay announcements. This is particularly important since: (i) as indicated above, these announcements may not always be accurate, and (ii) real-time queue-length information is typically readily available in service systems, such as in amusement parks, banks, or hospitals, and is usually easier to keep track of than wait-time information.

We also performed simulation experiments to investigate, numerically, how the relative error in the queue length translates into the accuracy of the LES announcement. For example, based on our numerical results, we find that for a large and heavily-loaded system, a queue-length error of less than 5% corresponds to a median waiting time error that is about 4%, for continuous and strictly increasing abandonment-rate functions.

1.4.2. Contributions: The QED Regime.
With announcement-dependent abandonment, it is not clear how customer response, particularly for small wait-time values, will affect both the asymptotic behavior in the system and the accuracy of the LES announcement. In particular, it may be that discontinuous customer abandonment behavior at the origin could lead to asymptotically inaccurate LES announcements. We show that this is not the case, and that the LES announcement is asymptotically accurate in the QED regime, under relatively mild conditions and provided that the initial queue length in the system is tight around its fluid limit.

1.4.3.   Contributions: The ED Regime. With asymptotically non-negligible waiting times, the analysis of the system involves a non-trivial equilibrium. Armony et al. (2009) derived conditions guaranteeing the existence and uniqueness of such an equilibrium in an approximating deterministic fluid model of the system. In this fluid model, all delayed customers receive the same delay announcement at equilibrium, and they subsequently experience the same waiting time. In other words, the LES announcement is accurate, at equilibrium, in the fluid model.

In the stochastic queueing system, waiting times for served customers fluctuate around the equilibrium expected waiting time value (which is approximated by the deterministic fluid waiting time). Even if the system is at equilibrium at fluid scale, it is not clear how those stochastic fluctuations will impact the accuracy of the individual LES announcements. Armony et al. (2009) left the problem of "quantifying the impact of (such) stochastic fluctuations for future research" (p. 78). In this paper, we extend Armony et al. (2009) and establish the asymptotic accuracy of the LES announcement in the ED regime with customer response. In particular, we formulate sufficient conditions for which initializing the system at equilibrium (at fluid scale) guarantees the asymptotic accuracy of the LES announcement with customer response.

1.4.4.   Insights Based on Numerical Experiments. In §6, we describe results of simulation experiments which quantify the accuracy of the LES announcement. There, we further our understanding of how customer response affects the accuracy of the LES announcement. For example, we go beyond previous work which focused solely on steady-state conditions. We illustrate that the LES announcement may not be accurate in the transient state of the system, and derive heuristic adjustments that outperform the straightforward LES announcement in that state. We also consider examples where our main theoretical results fail to hold. As such, we provide more evidence of the importance to pay close attention to exactly how customers respond to delay announcements.

The remainder of this paper is organized as follows. In §2, we review the relevant literature. In §3, we introduce our model. In §4, we present theoretical and numerical results for the QED regime. In §5, we present theoretical and numerical results for the ED regime. In §6, we present simulation experiments which validate and extend our theoretical results. In §7, we draw conclusions. We present all proofs in the appendix.

2.   Literature Review

Asymptotic Analysis of Multiserver Queues. We perform an asymptotic analysis of queueing systems in this paper. In particular, we focus on both the QED and the ED heavy-traffic limiting regimes. The QED or Halfin-Whitt limiting regime was first formalized in the seminal paper by Halfin and Whitt (1981). The authors of that paper focused on the classical $GI/M/N$ model with a general renewal arrival process, exponential service times, and no customer abandonment; they

showed that the delay probability approaches a limit strictly between 0 and 1 if, and only if, the system is staffed according to the square-root staffing rule. Here is a sample of other work along similar lines. Jennings et al. (1996) used the QED regime to determine staffing levels in Markovian queues with a time-varying demand. Fleming et al. (1994) and Garnett et al. (2002) extended the QED framework and incorporated the phenomenon of customer abandonment into their models. Puhalskii and Reiman (2002) studied multiclass queueing systems with a renewal arrival process and phase-type service times, both with and without customer priorities. Talreja and Whitt (2009) extended Garnett et al. (2002), and established stochastic-process limits for waiting times in multi-server queueing models with generally distributed service times and times to abandon. For additional references, see Aksin et al. (2007).

The ED regime supports low-to-moderate quality of service, and often yields useful and simple approximations. Whitt (2004) derived stochastic-process limits for the $M/M/N + M$ model in the ED regime, and developed approximations based on those limits. Borst et al. (2004) investigated the staffing problem of large call centers in an asymptotic optimization framework. They focused on three operational regimes, including the QED and ED regimes. Whitt (2006) conjectured the existence of a deterministic fluid limit for the general $G/GI/N + GI$ model in the ED regime. That fluid limit was later established in Kang and Ramanan (2010) and Zhang (2013). Talreja and Whitt (2009) established stochastic-process limits for waiting times in the ED regime as well.

Delay Announcements. The most closely related works to the current paper are Armony et al. (2009) and Ibrahim and Whitt (2009). Armony et al. (2009) studied the performance impact of making LES delay announcements by analyzing an approximating fluid model. They discussed the motivation for the LES delay announcement, and modeled changes in customer behavior that result from such an announcement. However, unlike our work here, the authors of that paper did not establish the accuracy of the individual announcements. Ibrahim and Whitt (2009) established the accuracy of the LES announcement in many-server Markovian models, in the ED regime, but they did not consider customer response to the announcements. They also focused solely on steady-state behavior in their models. Some other references related to delay announcements include Whitt (1999a, b), Armony and Maglaras (2004), Guo and Zipkin (2007), Ibrahim and Whitt (2009), Jouini et al. (2011), Allon et al. (2012a, b), Jouini et al. (2015), and references therein. The recent work in Senderovich et al. (2015) takes an empirical process mining approach to study the accuracy of snapshot-based predictions (essentially delay-history-based predictions such as LES). The authors provide evidence of the accuracy of these predictions with real-life data. Some of the published literature on delay announcements focused on the problem of determining "the best" wait-time quote (by assuming appropriate cost structures) and studied the advantages of both overestimating

8

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

and underestimating anticipated delays, e.g., see Jouini et al. (2015). In this paper, we focus on the problem of accurately predicting anticipated delays in the system instead.

Several recent papers emphasize the importance of incorporating customer response to the announcements, and demonstrate empirically that customers respond to delay announcements in practice. Mandelbaum and Zeltyn (2013) quantified the effect of the announcements by statistically estimating the hazard-rate of the abandonment-time distribution. Aksin et al. (2015) modeled customer abandonment decisions with delay announcements. They used an empirical approach which estimates the parameters of the abandonment distribution from data, and studied the effect of customer behavioural changes in a queueing setting. Yu et al. (2014) explored the impact of delay announcements using an empirical approach. Interestingly, they found that delay announcements affect customer abandonment behavior in a complex way, and that they directly affect the waiting costs of delayed customers. Acknowledging the importance of customer response to the announcements, Huang et al. (2015) studied the optimal timing of delay announcements and optimal staffing decisions in an asymptotic framework which accounts for the impact of delay announcements on the abandonment-time distribution.

3. Modelling Framework

In this paper, we consider single class $M/M/N + M$ queues, also known as Erlang A, with announcement-dependent balking and abandonment. We let the times between successive arrival epochs be independent and identically distributed (i.i.d.) exponential random variables with rate λ. We assume that there are N homogeneous servers working in parallel. We let service times be i.i.d. exponential random variables with rate μ. We let the times to abandon be i.i.d. exponential random variables with rate θ. The traffic intensity, ρ, is given by $\rho \equiv \lambda/N\mu$. There is unlimited waiting space and we use the first-come-first-served (FCFS) service discipline.

We envision that each delayed customer is given, upon arrival, a single-number prediction of his waiting time before entering service. A delayed customer, arriving to the system at time t, receives a delay announcement w_t and may balk, upon arrival, with probability $b(w_t)$. If that customer does not balk, then he will abandon the queue before being served if his waiting time exceeds an exponentially distributed random variable with rate $\theta(w_t)$. That is, individual balking probabilities and abandonment rates depend on the announcements.

We are now ready to give a precise definition of the LES announcement. Let t denote the arrival epoch of a new customer. Let the patience of that customer be denoted by $K(t)$. Let the virtual waiting time, at time t, be denoted by $W(t)$, i.e., $W(t)$ is the waiting time of a hypothetical infinitely patient customer arriving to the system at time t. Let τ_t^N be the arrival time of the last customer to have entered service prior to t, which is defined as:

$$\tau_t^N = \sup\{s \leq t : \text{There is an arrival at time } s, \ s + W(s) \leq t, \text{ and } K(s) > W(s)\}; \qquad (1)$$

the customer arriving to the system at time τ_t^N is the LES customer at time t.

4.   Asymptotic Accuracy of LES in the QED Regime

Waiting times in the QED regime are asymptotically small, converging to zero at a rate which is proportional to $1/\sqrt{N}$, as $N \to \infty$, where N is the number of servers. Given that the magnitude of waiting times is asymptotically negligible, it seems natural to conclude that only the abandonment response behavior at the origin should matter asymptotically. In our setting, if system dynamics could be well approximated by assuming a constant abandonment rate, equal to $\theta(0)$, then the asymptotic accuracy of LES should carry through from previously established results, which do not assume any customer response to the announcements (Ibrahim and Whitt, 2009).

However, customer abandonment response may very well be rapidly changing around zero. In particular, we may have discontinuous customer abandonment behavior at the origin. For example, this may arise in practice when customers are "extremely" impatient to any waiting so that there is a jump in their impatience in response to being announced a positive delay; e.g., see Figure 12 in Mandelbaum and Zeltyn (2013). More generally, customer abandonment response may be irregular in a real-life context: Yu et al. (2015) present empirical evidence supporting that delay announcements are influential on customer abandonment times, but that there are "no particular patterns" (p. 11) for how announcements impact those abandonment times. When customer abandonment behavior changes rapidly around zero, approximating system performance by using the abandonment rate at the origin is no longer appropriate. With such abandonment behavior, it is not clear, a priori, how customer response, particularly around the origin, will affect both the asymptotic behavior in the system and the asymptotic accuracy of the LES announcement.

Similar ideas about the importance of customer abandonment behavior for small wait-time values were advanced in Reed and Ward (2008) and Reed and Tezcan (2012). These authors proposed heavy-traffic limits which capture rapidly-changing abandonment behavior at the origin. Their limits result from scaling the abandonment hazard-rate function appropriately, and involve the entire abandonment-time distribution. They showed that the superiority of their new heavy-traffic approximations is most pronounced when the hazard rate changes rapidly around zero. That is, they showed that simply approximating system performance with abandonment behavior at the origin may lead to poor approximations. In our setting, this means that the asymptotic accuracy of LES is not obvious, and cannot be simply deduced from the existing literature. In this section, we demonstrate that the LES announcement is asymptotically accurate in the QED regime, irrespective of customer abandonment response to the announcements (around the origin or elsewhere) provided that the abandonment-rate function is bounded. Interestingly, we show that this asymptotic accuracy continues to hold for non-monotonic and/or discontinuous customer abandonment

10

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

behavior (in contrast, the third subplot in Figure 1 corresponds to an overloaded system, where additional initial conditions on customer abandonment behavior are needed).

4.1. Asymptotic Framework

Consider a sequence of queueing systems indexed by N, and let $N \to \infty$. Let the arrival rate in the Nth system be given by λ^N. There are N servers in the Nth system, each having the same service rate μ. As in Garnett et al. (2002), and consistently with the QED regime, we assume that:

$$\lim_{N \to \infty} \sqrt{N} \left(1 - \frac{\lambda^N}{N\mu} \right) = \beta, \text{ for } \beta \in (-\infty, \infty). \tag{2}$$

We now consider the Nth system in that sequence. At time t, the LES delay announcement is given by $W^N(\tau_t^N)$ for τ_t^N in (1). We let $b : \mathbb{R}_+ \to [0,1]$, where $\mathbb{R}_+$ denotes the set of nonnegative real numbers. We assume that $b(\cdot)$ is a Lipschitz continuous, monotone non-decreasing function with $b(0) = 0$. We also let $\theta : \mathbb{R}_+ \to [\underline{\theta}, \bar{\theta}]$ with $\bar{\theta} > \underline{\theta} \geq 0$. A new arrival at time t balks with probability $b(W^N(\tau_t^N))$. If the customer does not balk, then he may abandon the queue prior to beginning service, and his abandonment rate is given by $\theta(W^N(\tau_t^N))$. We assume that $\underline{\theta} > 0$ if $\beta \leq 0$ in (2), in order to guarantee the stability of the system. We note that $\theta(\cdot)$ and $b(\cdot)$ do not scale with N. We let $\Rightarrow$ denote convergence in distribution; see Whitt (2002). Next, we state our main theorem and outline its proof. We relegate the details of the proof to the appendix. In this section, we consider an exponential patience distribution. In §9, we go beyond this assumption and consider a general patience distribution; this is important because there is empirical evidence showing that the patience distribution is usually non-exponential in practice; e.g., see Brown et al. (2005).

4.2. Main Theorem and Outline of Proof

Let $Z^N(s)$ denote the number of customers at time s in the Nth system. We assume that the sequence $\{(Z^N(0) - N)/\sqrt{N}\}_{N \geq 1}$ is tight; for more on tightness, see §5 of Pang et al. (2007).

THEOREM 1. *For the $M/M/N + M$ model in the QED many-server heavy-traffic regime: if $\{(Z^N(0) - N)/\sqrt{N}\}_{N \geq 1}$ is tight, then*

$$\sqrt{N} |W^N(t) - W^N(\tau_t^N)| \Rightarrow 0 \quad in \quad \mathbb{R}, \tag{3}$$

for every fixed time point t, as $N \to \infty$.

Theorem 1 specifies an initial condition which guarantees the asymptotic relative accuracy of the LES announcement at time t. This condition implies that the initial number of customers in the queue and the initial number of idle servers are not too large, which is consistent with QED characteristics and is a common assumption often made in the literature, e.g., see Garnett et al. (2002).

[1] Since delays are asymptotically of the order of $O_p(1/\sqrt{N})$ [2], we divide the absolute difference in (3) by $1/\sqrt{N}$. The key to proving Theorem 1 lies in establishing that the LES announcement is relatively accurate if, and only if, the relative difference between the queue lengths seen upon arrival by a customer and his corresponding LES customer is asymptotically negligible.

Technical challenges. In a system with no customer response, the asymptotic accuracy of the LES announcement follows directly from the snapshot principle; e.g., see Reiman (1982) and Puhalskii and Reiman (2000). In the existing literature, diffusion limits for waiting times have been usually proven using Puhalskii's invariance principle for first-passage times (Puhalskii, 1994) together with established diffusion limits for queue-length processes; e.g., see Puhalskii and Rieman (2000). However, employing a similar proof technique is prohibitively difficult in our system; perhaps even impossible. Indeed, our proof technique does not rely on establishing diffusion limits for the (scaled) queue-length and wait-time processes in a system with customer response. Instead of establishing that convergence directly in the original system (with customer response), we show that the snapshot principle holds in two bounding auxiliary queueing systems; see §8. As such, we show that the snapshot principle must hold in the original system too. The bounding arguments that we rely on require that the scaled state at zero has a limit as N goes to infinity. In particular, this is needed in order to be able to apply the results of Garnett et al. (2002) to the upper and lower bound processes (that bound our original system). Our assumed tightness at the origin implies that any sequence of diffusion-scaled states at time zero has at least one converging subsequence. Then, we can apply results from Garnett et al. (2002) with respect to each such subsequence, and the asymptotic accuracy of LES in our original system follows. It is important to emphasize that our proof technique does not amount to a standard sandwiching argument, i.e., to showing that the scaled wait-time and queue-length processes in the upper and lower bound systems converge to the same limit. Indeed, the bounding processes *do not* converge to the same limit; additionally, the processes in the original system need not converge at all.

Proof outline. To prove Theorem 1, we proceed as follows. First, we show that the time between the arrival epochs of the LES and current customer is negligible in the heavy-traffic limit. Then, we show that the snapshot principle holds, i.e., that the queue length (system state) changes negligibly between the arrival epochs of the LES and current customer. When establishing that the relative error in the queue length is small, we scale the difference in queue lengths by $\sqrt{N}$ since this is the order of magnitude of queue lengths in the QED limiting regime; see Garnett et al. (2002). Finally,

---

[1] For a back-of-the-envelope example which violates this initial condition, consider a system where all servers are busy and $Q^N(0)$ is a constant that grows with N faster than $\sqrt{N}$, e.g., $Q^N(0) = N^{3/4}$. Then, it is clear that our initial tightness condition does *not* hold.

[2] Let $\{X_n, n \geq 1\}$ be a sequence of random variables and $\{a_n, n \geq 1\}$ be a sequence of real numbers. We say that $X_n = O_p(a_n)$ if for every $\epsilon > 0$ there exists a finite $M > 0$ such that $P(|X_n/a_n| > M) < \epsilon$ for all n.

we establish an asymptotic relation between the queue length and the waiting time. Combining those results yields Theorem 1.

4.3. Supporting Numerical Results

In this section, we describe results of a simulation study which quantifies the accuracy of the LES announcement. Our objective is to substantiate our theoretical results by considering many-server $M/M/N + M$ queues in the QED regime. To quantify the accuracy of the LES announcement, we use the *average-squared-error* (ASE): $ASE \equiv (1/n) \sum_{j=1}^{n} (a_j - p_j)^2$, where $a_j > 0$ is the virtual delay of customer j, p_j is his predicted delay, and n 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 delay prediction and actual delay. 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.

Unless stated otherwise, our simulation results throughout are based on 10 independent replications of 2 million events each, where an event is either a service completion, an arrival, or an abandonment from the system. Our simulations are steady-state simulations. For this, we exclude from each simulation run the first 5,000 events so as to remove the effect of the initial transient period. In Figure 2, we vary the number of servers, N, and consider values ranging from $N = 10$ to $N = 1000$. Without loss of generality, we assume that the service rate is $\mu = 1$. That is, we measure time in units of mean service time. We define the balking probability, $b(w)$, as follows:

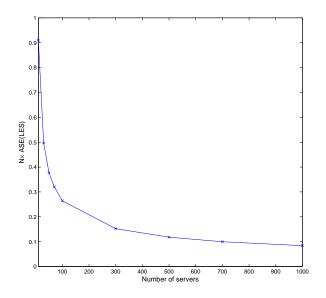$$b(w) = \frac{1}{10} - \frac{1}{10}e^{-w} \text{ for } w \geq 0 . \tag{4}$$

This balking function yields a balking proportion of roughly 6% in response to a delay announcement $w = 1$, i.e., to an announcement equal to the mean service time in the system. We let the abandonment rate of a customer who does not balk be defined as:

$$\theta(w) = \frac{3}{4} - \frac{1}{2}e^{-w} \text{ for } w \geq 0 . \tag{5}$$

Then, $\theta(0) = 1/4$ is the abandonment rate corresponding to a delay announcement $w = 0$.

In Figure 2, we plot $N \times$ ASE(LES) as a function of N. We fix $\rho = 1$. For $\rho = 1$ and relatively large values of N, QED approximations are relatively accurate; see Garnett et al. (2002). Theorem 1 shows that the LES announcement is asymptotically accurate in this case. Figure 2 shows that $N \times$ ASE(LES) decreases as N increases, and converges to 0 for large N. This is consistent with our theoretical results where we show that ASE(LES) is roughly $o(1/N)$ in the QED regime [3]. Figure 2 illustrates that the LES announcement performs relatively poorly with a very small number of servers ($N = 10$), but its accuracy improves rapidly as the number of servers increases, e.g., causing a sharp decrease for $N \times$ ASE(LES) in going from $N = 10$ to $N = 30$.

---

[3] Let f and g be two functions defined on some subset of the real numbers. Then, $f(n) = o(g(n))$ as $n \to \infty$ if for all $\epsilon > 0$, there exists N such that $|f(n)| \leq \epsilon |g(n)|$ for all $n \geq N$.
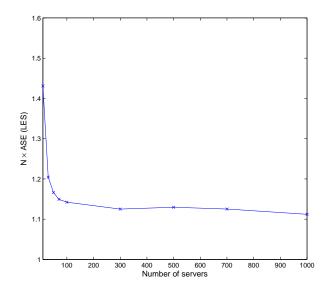
Figure 2 $N \times$ **ASE(LES)** in the $M/M/N + M$ model in the QED regime with customer response given by (4) and (5).

Figure 3 $N \times$ **ASE(LES)** in the $M/M/N + M$ model in the ED regime with customer response given by (4) and (5).

5.   Asymptotic Accuracy of LES in the ED Regime

We now focus on overloaded scenarios, in which the arrival rate exceeds the maximum possible total service rate. In particular, we consider the ED limiting regime where the asymptotic magnitude of waiting times is non-negligible. It is practically important to consider the ED regime because we are primarily interested in making delay announcements when delays are large.

Establishing the asymptotic accuracy of the LES announcement in the ED regime is complicated. Essentially, since waiting times are asymptotically long, the state of the system may change significantly during the LES delay, and the LES delay announcement may not be close to the new arrival's delay. With non-negligible waiting times and customer response to the announcements, the analysis of the system involves a complex equilibrium. Armony et al. (2009) derived conditions guaranteeing the existence and uniqueness of that equilibrium in an approximating deterministic fluid model of the system. In this fluid model, all delayed customers receive *the same* delay announcement at equilibrium, and they subsequently experience the same waiting time. In other words, the LES announcement is accurate, at equilibrium, in the fluid model.

In the stochastic queueing system, waiting times for served customers fluctuate around the equilibrium expected waiting time value (which is approximated by the deterministic fluid waiting time). Even if the system is at equilibrium at fluid scale, it is not clear how those stochastic fluctuations will impact the accuracy of the individual LES announcements. Armony et al. (2009) left the problem of "quantifying the impact of (such) stochastic fluctuations" for future research

14

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the manuscript number!)

(p. 78). In this section, we extend Armony et al. (2009) and establish the asymptotic accuracy of the LES announcement in the ED regime. We show, in Theorem 2, that initializing the system at equilibrium at fluid scale is sufficient to guarantee that asymptotic accuracy.

5.1.   Asymptotic Framework

We consider a sequence of queueing systems indexed by N, where N is the number of servers. The arrival rate in the Nth system is given by $\lambda^N = N\lambda$. We let $N \to \infty$ while holding the traffic intensity $\rho > 1$ fixed. For every system, we fix the service rate and let it be equal to μ, independently of N. Let $\bar{b}(w)$ denote the probability of joining the system (not balking) when receiving an LES delay announcement equal to w. Then, $\bar{b}(w) = 1 - b(w)$, where $b(w)$ is the corresponding probability of balking. We assume that $\bar{b}(0) = 1$, $\bar{b}(w) \to 0$ as $w \to \infty$, and $\bar{b}(\cdot)$ is a strictly decreasing and continuous function. We also assume that $\theta(\cdot)$ is a continuous and strictly increasing function. These assumptions on $b(\cdot)$ and $\theta(\cdot)$ are sufficient to guarantee the existence and uniqueness of a fluid equilibrium in the system, as we show in §5.2 [4]. We need those additional assumptions on $b(\cdot)$ and $\theta(\cdot)$ because the magnitude of the equilibrium waiting time in the ED limit is not asymptotically negligible, unlike in the QED limit. In our proofs, we also assume that $b(\cdot)$ and $\theta(\cdot)$ are differentiable at that unique fluid equilibrium point. We note that $\theta(\cdot)$ and $b(\cdot)$ do not scale with N.

5.2.   Fluid Steady-State Equilibrium

We begin by considering the fluid model approximation of the system. At equilibrium, the announced delay must be consistent with the actual delay for served customers, after customer response. Let $\bar{w}$ denote an equilibrium waiting time in the fluid model. Let $\bar{z}$ denote an equilibrium fluid content in the system. Then, $\bar{w}$ and $\bar{z}$ must satisfy the two following equations:

$$\lambda\bar{b}(\bar{w}) = \mu + \theta(\bar{w})(\bar{z} - 1), \tag{6}$$

$$\bar{w} = \frac{1}{\theta(\bar{w})} \ln\left(1 + \frac{\theta(\bar{w})(\bar{z} - 1)}{\mu}\right). \tag{7}$$

Equation (6) is a balance equation which follows since the long-run rate into the system must equal the long-run rate out of the system, by service or abandonment. Equation (7) follows from the relation between the waiting time and the queue content in the fluid model; e.g., see Equation (3.7) of Whitt (2006). In the ED regime, we must have that $\lambda > \mu$. The continuity assumptions on $\bar{b}(\cdot)$ and $\bar{\theta}(\cdot)$, along with the boundary conditions on $\bar{b}(\cdot)$, guarantee the existence of an equilibrium. The monotonicity assumptions on those two functions guarantee the uniqueness of that equilibrium. Thus, under our assumptions, there is a unique solution $(\bar{w}, \bar{z})$ that satisfies (6) and (7).

---

[4] The assumptions on $b(\cdot)$ and $\theta(\cdot)$ are only needed to guarantee the existence and uniqueness of a fluid equilibrium. Thus, we could also assume instead that such an equilibrium exists and is unique.

5.3. Main Theorem and Outline of Proof

In this section, we focus on a special case of customer response: We assume that customer abandonment behavior is unaffected by delay announcements; customers may still balk upon arrival, and their balking probabilities depend on the announcements. Theorem 2 is our main theorem for this case (we provide its proof in the appendix): It states that, under a mild technical condition on the function $\bar{b}$, LES is asymptotically accurate if the system is initialized at equilibrium at fluid scale. That is, we impose convergence of $Z^N(0)/N$ to its fluid limit $\bar{z}$ in Theorem 2. For a back-of-the-envelope example where this initial condition does *not* hold, consider an initial system state where $Z^N(0) = N(\bar{z} + \delta)$ for some $\delta > 0$. Then, the system state will change considerably until the system reaches its equilibrium, and the LES announcement will not be accurate as it will overestimate the actual delay at all times. Establishing the case with both announcement-dependent balking and abandonment is more complicated algebraically; that is why we relegate the relevant theorem and proof to the appendix. Recall that $Z^N(s)$ is the number of customers in the system at time s. We also let $T > 0$.

THEOREM 2. *For the $M/M/N + M$ model in the ED heavy-traffic limiting regime with announcement-dependent balking and a constant abandonment rate θ,*

If

$$\frac{Z^N(0)}{N} \Rightarrow \bar{z} \quad in \quad (6) \quad and \quad (7) \quad as \quad N \to \infty, \tag{8}$$

then

$$\left|\left|W^N(t) - W^N(\tau_t^N)\right|\right|_{[0,T]} \to 0 \text{ almost surely as } N \to \infty \quad, \tag{9}$$

under the condition that

$$\left|\frac{\bar{b}'(\bar{w})}{\bar{b}(\bar{w})}\right| < \theta. \tag{10}$$

It is readily seen that the condition in (10) is satisfied when $\bar{b}$ is equal to an exponentially decaying function whose rate is smaller than θ. It is important to stress that we focus on the *relative* accuracy of the LES announcement in this paper. By relative accuracy, we mean the difference between the LES and actual delays, *scaled by the appropriate asymptotic order of magnitude of delays in the system*. As a result of this, the expressions for asymptotic accuracy in the QED and ED regimes are different; e.g., compare (3) with (9). The $\sqrt{N}$ factor in (3) is due to dividing the difference of the waiting times by $1/\sqrt{N}$, which is the asymptotic order of magnitude of the waiting times in the system. Thus, the $\sqrt{N}$ factor is a reflection of the smaller magnitude of waiting times.

Technical challenges. As in the proof of Theorem 1 for the QED regime, the main step is to show that the relative error in the waiting times is asymptotically negligible if, and only if, the relative error in the queue lengths is asymptotically negligible. In the ED regime, the queue length is $O_p(N)$ and the waiting time is $O_p(1)$; that is why we use these scalings in (8) and (9), respectively. Since customer response complicates system dynamics, it is difficult to characterize limits for the wait-time process on $[0, T]$ directly, as in (9). Indeed, stochastic fluctuations in the waiting times affect the LES announcements made, which in turn affect balking probabilities. These probabilities determine both the number of customers in the system and subsequent waiting times. To circumvent that difficulty, we devise a stopping-time argument instead, as in Gurvich and Whitt (2009).

Proof outline. Our proof proceeds as follows. We begin by bounding the stochastic fluctuations of the scaled number of customers in the system up to a given stopping time, σ^N. We restrict attention to the bounded stopping time, $\alpha^N = \min\{\sigma^N, T\}$. Then, we show that if the stopped number of customers in the system is close to its equilibrium value up to α^N, then the stopped waiting time will be close to its equilibrium value up to α^N as well, i.e., we establish the stochastic boundedness of the stopped waiting times. We do so by proving an asymptotic relationship between the waiting time in the system and a function of the number of customers in the system, and then exploiting a Taylor series expansion argument. Next, we show that $T < \sigma^N$, for every $T > 0$; since T can be made arbitrarily large, we obtain that the stopping time itself diverges to ∞. For this, we establish that the scaled number of customers in the system is stochastically bounded *at* α^N as well. To do so, we exploit results on the convergence of the scaled number of customers in a system with state-dependent arrival rates (since balking probabilities depend on the delay announcements made), together with a bounding argument and the additional technical condition in (10). Consequently, drawing on the analysis above, (9) must hold too. In other words, the relative errors in both the queue lengths and the waiting times are asymptotically negligible, provided that the system is initialized at its equilibrium fluid steady state.

5.4. Supporting Numerical Results

5.4.1. Validating Theorem 2. We substantiate our theoretical results by considering many-server $M/M/N + M$ queues in the ED regime. For $b(w)$ and $\theta(w)$, we consider the functions in (4) and (5). With those balking and abandonment-rate functions, it can be readily checked that a unique equilibrium exists in the system, as per (6) and (7).

In Figure 3, we let $\rho = 1.4$ and consider the same values for N as in Figure 2. With $\rho = 1.4$ and large N, ED approximations are relatively accurate; see Whitt (2004). Figure 3 shows that $N \times$ ASE(LES) is roughly constant as N increases. This suggests that the LES announcement is asymptotically accurate in the ED regime and that ASE(LES) converges to 0 at a rate which is

inversely proportional to N. This substantiates and supplements our theoretical results. Indeed, in Theorem 2 (and Theorem 4) we show that ASE(LES) is asymptotically negligible in the ED regime, but do not specify the rate at which it converges to 0. Figure 3 suggests that ASE(LES) is $O(1/N)$ [5]. Consistent with Figure 2 for the QED regime, Figure 3 shows that the accuracy of LES is poor with a very small number of servers ($N = 10$), but quickly improves as the number of servers increases. Indeed, Figure 3 shows that $N \times$ ASE(LES) does not vary by much for $N \geq 30$.

5.4.2. Customer Response and Fluid Equilibrium. As illustrated in Figure 1, introducing customer response to the announcements may significantly complicate system dynamics. The main complexity in incorporating customer response to the announcements lies in the existence, or possibly lack thereof, of a unique equilibrium of the system. This equilibrium is non-trivial when waiting times in the system are long, e.g., as in the ED limiting regime. Theorem 2 shows that if there exists a unique equilibrium of the system, then initializing the system at that equilibrium, at fluid scale, is sufficient to ensure the asymptotic accuracy of the LES announcement. There, we imposed continuity and strict monotonicity assumptions on $\theta(\cdot)$ and $b(\cdot)$ which guarantee both the existence and uniqueness of that equilibrium. We also showed that the relative error in the wait times is asymptotically negligible if, and only if, the relative error in the queue lengths is negligible.

We now consider abandonment-rate response functions for which: (i) there does not exist an equilibrium of the system, or (ii) there exist multiple equilibria of the system. We investigate whether our previous results continue to hold in such scenarios. Interestingly, we show that this may not be the case. This is in contrast to systems without customer response, where the asymptotic accuracy of the LES announcement was shown to hold irrespective of specific assumptions on system parameters; e.g., see Ibrahim and Whitt (2009).

Our objective is twofold: (i) to investigate, numerically, how the relative error in the queue length translates into the accuracy of the LES announcement in systems where there exists a unique equilibrium; and (ii) to show that the equivalence between small wait-time and small queue-length errors may not hold more generally, specifically when an equilibrium does not exist. Therefore, our results show that the existence (or lack thereof) of a unique equilibrium strongly affects the asymptotic accuracy of the LES announcement, and illustrate how the respective magnitudes of wait-time and queue-length errors are affected.

Stability of Wait-Time and Queue-Length Errors. Point (i) above is important from a practical perspective so that system managers, who may typically observe the queue-length, are able to quantify the errors in the LES announcements based on the queue-length errors that they observe.

---

[5] Let f and g be two functions defined on some subset of the real numbers. Then, $f(n) = O(g(n))$ as $n \to \infty$ if there exists $M > 0$ and $N > 0$ such that $|f(n)| \leq M|g(n)|$ for $n \geq N$.

In our simulation experiments, we collect the relative queue-length errors reported (differences between the queue lengths seen by the new and LES customers, scaled appropriately), and partition these into the following intervals:

$$(0, 0.05), (0.05, 0.1), (0.1, 0.2), (0.2, 0.3), (0.3, 0.4), (0.4, 0.5), \text{ and } (0.5, 1).$$

For example, the first interval corresponds to queue-length errors that are smaller than 5%, while the second interval corresponds to queue-length errors which are between 5% and 10%. For each interval, we collect the corresponding relative wait-time errors in the simulation run. For example, we collect all relative wait-time errors which correspond to queue-length errors that are smaller than 5% (first interval), or those which correspond to queue-length errors that are between 5% and 10% (second interval), and so on. We then compute the median of those wait-time errors to assess precisely how the error in the queue length translates into the wait-time error.

We consider two forms for the abandonment-rate response function, and assume that there is no customer balking in the system. We consider $\theta_1(w)$ given by:

$$\theta_1(w) = b - e^{-aw} \text{ where } a, b > 0. \tag{11}$$

With $\theta_1(w)$, there exists a unique equilibrium of the system, and our theoretical results continue to hold. We vary a and b in (11) to consistently have that $\bar{w} = \ln(\rho) = \ln(1.4)$; this is the steady-state fluid waiting time in a system with no customer response to the announcements, and with $\theta(w) = 1$ for all w. Increasing a amounts to increasing the intensity of customer response to the announcements. Second, we violate the continuity assumption and let

$$\theta_0(w) = \begin{cases} 0.5, & \text{if } w \leq 0.5, \\ 1.5, & \text{otherwise}, \end{cases} \tag{12}$$

so that $\theta_0(w)$ has a discontinuity at $w = 0.5$. Then, it is not hard to show that there do not exist $\bar{w}$ and $\bar{z}$ which simultaneously solve (6) and (7); thus there is no equilibrium of the system. In the online supplement, we present more simulation results for various models, in particular we consider alternative abandonment-rate functions and alternative system sizes and congestion levels.

In Table 1, we report our results for $N = 1000$ and $\rho = 1.4$. Table 1 shows that the order of magnitudes of the queue-length and wait-time errors are generally close for $\theta_1(\cdot)$, irrespective of a and b. For example, for queue-length errors that are smaller than 5%, the median of corresponding wait-time errors is about 4%. Table 1 also shows that wait-time errors fluctuate less extremely than queue-length errors. For example, for queue-length errors that are in $(0.1, 0.2)$, the median of corresponding wait-time errors, under $\theta_1(\cdot)$, remains around 6%. This suggests that the LES announcement will be accurate in practice, even when the queue-length error is not too small.

Queue Length	$\theta_1(w)$ in (11)					$\theta_0(w)$ in (12)
	$a=0, b=2$	$a=0.5, b=1.85$	$a=1, b=1.71$	$a=1.5, b=1.6$	$a=2, b=1.51$	
<0.05	0.0428	0.0434	0.0436	0.0437	0.0439	0.103
$\in (0.05, 0.1)$	0.0479	0.0487	0.0490	0.0495	0.0497	0.101
$\in (0.1, 0.2)$	0.0618	0.0621	0.0627	0.0630	0.0630	0.115
$\in (0.2, 0.3)$	0.105	0.1076	0.109	0.108	0.108	0.153
$\in (0.3, 0.4)$	0.176	0.169	0.175	0.167	0.165	0.202
$\in (0.4, 0.5)$	0.185	0.210	0.202	0.237	0.231	0.229
>0.5	0.405	0.407	0.407	0.414	0.421	0.413

Table 1   Relative errors for the queue length and median wait-time estimates for the $M/M/1000 + M$ model
with $\rho = 1.4$, $\theta_1(w) = b - e^{-aw}$, and $\theta_0(w)$ in (12).

In contrast, Table 1 shows that, for $\theta_0(\cdot)$, "large" wait-time errors may correspond to "small" queue-length errors. For example, a median wait-time error of over 10% corresponds to queue-length errors which are smaller than 5%. Indeed, because of the discontinuity in $\theta(\cdot)$, it is possible that two customers who encounter, upon arrival, the same queue lengths in the system will still experience very different waiting times. This is because customers waiting in the queue may have considerably different abandonment rates, depending on the announcements that they received, so that the queue length seen upon arrival is not, by itself, a sufficient indicator of the ensuing wait.

We now make a comparison with a system with no customer response to the announcements, i.e., where $\theta_c(w) = 1$ for all w. In Figures 4 and 5, we plot relative errors for queue-lengths as a function of relative errors for the waiting times under $\theta_0(w)$ and $\theta_c(w)$, respectively. On one hand, Figure 4 shows that small queue-length errors of roughly 10% may correspond to large wait-time errors of about 50%. On the other hand, Figure 5 clearly shows that the relative errors in the waiting times are small if, and only if, the relative errors in the queue lengths are small. Contrasting Figures 4 and 5 illustrates how incorporating customer response to the announcements may drastically change the underlying dynamics of the system.

Stability of the Equilibrium. We now consider a system where there exist multiple equilibria. In particular, we exclude customer balking and consider that the abandonment rate

$$\theta_{00}(w) = \begin{cases} 4, & \text{if } w < 0.1, \\ 7.5 - 35w & \text{if } 0.1 \leq w < 0.2, \\ 0.5, & \text{if } w \geq 0.2; \end{cases} \tag{13}$$

then, it is easy to verify that there exist three equilibria of the system: $\bar{w}_1 = 0.084$, $\bar{w}_2 = 0.15$, and $\bar{w}_3 = 0.67$. In Figure 6, we plot the relative errors in the queue lengths as a function of the relative errors in the waiting times, for $\theta_{00}(w)$ in (13); otherwise, we consider the same modelling assumptions as in Figure 4. Figure 6 shows that some large wait-time errors correspond to small queue-length errors, and vice versa. Once more, comparing Figures 5 and 6 illustrates the changes in system performance due to incorporating customer response to the announcements.
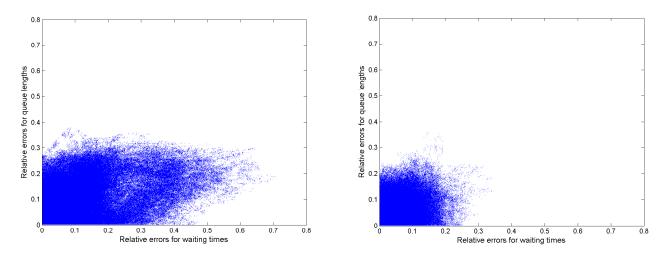
Figure 4 Relative errors for the waiting times and queue lengths for $\theta(w)$ in (12).
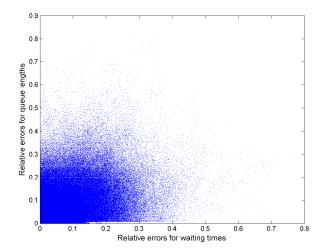

Figure 5 Relative errors for the waiting times and queue lengths for $\theta = 1$.

In Figure 7, we plot sample paths of the LES delays and actual delays observed in the same system as in Figure 6. Figure 7 shows that the system alternates between two equilibria. Indeed, the waiting times first stabilize around $\bar{w}_1 = 0.084$, and then around $\bar{w}_3 = 0.67$. It is interesting to see that $\bar{w}_2 = 0.15$ is an unstable equilibrium of the system, because small stochastic fluctuations drive waiting times away from $\bar{w}_2$. Figure 7 illustrates an important phenomenon which is due to incorporating customer response in the system: With multiple equilibria, if the system is initialized around one equilibrium, then stochastic fluctuations may drive the system away from that equilibrium. In general, our results in Theorem 2 no longer hold. This is why, with customer response, we need to impose additional assumptions on system parameters to guarantee stability, as in §5.2.

6. Additional Numerical Results

In this section, we describe results of a simulation study with the objective to improve our understanding of how customer response affects the accuracy of the LES announcement by going beyond our theoretical results in Theorems 1 and 2. First, we study how changes in model parameters (customer response, congestion level, and arrival process) affect the accuracy of the LES announcement, in steady state (§6.1-§6.3). Then, we derive heuristic adjustments to the LES announcement, and show that they are more accurate than the straightforward announcement in the transient state, albeit at the expense of requiring more information about system parameters (§6.4).

In addition to the ASE, and to get a relative measure of accuracy, we also compute point estimates of the *relative-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. The RASE is useful because it relates the error in the LES announcement to the magnitude of waiting times in the system.
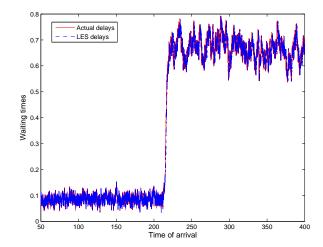
Figure 6   Relative errors for the waiting times and queue lengths in the $M/M/1000 + M$ model for $\rho = 1.4$ and $\theta(w)$ in (13).

Figure 7   LES and actual waiting times in the $M/M/1000 + M$ model for $\rho = 1.4$ and $\theta(w)$ in (13).

6.1.   Congestion Effects

We begin by studying how changes in the system's congestion level affect the accuracy of the LES announcement. We control system congestion in two ways: (i) by varying the magnitude of customer abandonment; and (ii) by altering the traffic intensity in the system. In both cases, we exclude balking from the system, to focus solely on the effect of customer abandonment and traffic intensity, i.e., we let $b(w) = 0$.

Abandonment-Rate Function. In Table 2, we study how changes in the abandonment rate function, $\theta(w)$, affect the accuracy of the LES announcement. We fix $N = 100$ and let $\rho = 1.4$, i.e., we focus on the ED regime. We do so because customer abandonment is then non-negligible. We consider an exponential functional form for $\theta(w)$, and vary its parameters to either speed-up or slow-down customer abandonment; in particular,

$$\theta(w) = k \cdot \left(\frac{3}{4} - \frac{1}{2} e^{-w} \right) \text{ for some } k > 0. \tag{14}$$

We consider the following values for k: 0.2, 1, 2, and 4. (The expression in (5) corresponds to $k = 1$.) The system experiences slower customer abandonment as the value of k decreases, and is then more congested. Therefore, we expect that ASE(LES) will be large for small values of k, but that the LES announcement will be relatively more accurate, i.e., yielding a smaller RASE(LES). Table 2 shows that this is indeed the case; e.g., RASE(LES) ranges from roughly 28% for $k = 2$ to roughly 9% for $k = 0.1$, whereas ASE(LES) is nearly 10 times larger for $k = 0.1$ than for $k = 2$.

Decreasing k leads to both an increase in ASE(LES) and an increase in the average waiting time in the system. Interestingly, Table 2 shows that, for a given decrease in k, the relative increase in

ASE(LES) is equal to the corresponding relative increase in the average waiting time. For example, both ASE(LES) and the average waiting time are multiplied by roughly 1.7 in going from $k = 4$ to $k = 2$. Therefore, RASE(LES) will be smaller for $k = 2$ than for $k = 4$. Indeed, Table 2 shows that RASE(LES) for $k = 1$ is roughly $1/\sqrt{1.7} \approx 0.76 \times$ RASE(LES) for $k = 4$. We observe similar relationships for all other values of k considered in Table 2.

Traffic Intensity. For Table 3, we fix $N = 100$. We consider $\theta(w)$ in (5). As before, we exclude balking from the system. We vary ρ from 1.0 to 1.6. Consistent with intuition, Table 3 shows that an increase in ρ leads to both an increase in ASE(LES) and an increase in the average waiting time. Table 3 also shows that RASE(LES) decreases as ρ increases. For example, RASE(LES) varies from roughly 60% for $\rho = 1$ to roughly 14% for $\rho = 1.6$.

Interestingly, for a given increase in ρ, the relative increase in ASE(LES) is smaller than the relative increase in the average waiting time. For example, ASE(LES) is roughly 3 times larger for $\rho = 1.4$ than for $\rho = 1.2$. In contrast, the average waiting time is roughly 5 times larger for $\rho = 1.4$ than for $\rho = 1.2$. As such, RASE(LES) is smaller for $\rho = 1.4$ than for $\rho = 1.2$. We observe similar relationships for all other values of ρ considered in Table 3.

This section demonstrates an important principle, which should be useful from a managerial perspective: the relative accuracy of LES improves with increased congestion in the system. As such, it is more useful to implement LES delay announcements in more congested systems. In particular, although the absolute magnitudes of the LES errors increase as the congestion in the system increases, the relative accuracy of LES, relative to the increasing average waiting time, improves. Comparing the first and last rows of Table 2, we see that when the average waiting time is multiplied by 10, RASE(LES) is roughly divided by $\sqrt{10} \approx 3$. Similarly, comparing the second and third rows of Table 3, we find that when the average waiting time is multiplied by 1.6, RASE(LES) is roughly divided by $\sqrt{1.6} \approx 1.3$. So, based on our numerical examples, it appears that when the average waiting time in the system is multiplied by $c > 1$, as a result of increased congestion, RASE(LES) is divided in that system by approximately $\sqrt{c}$. This should give some indication to practitioners regarding the relative accuracy of LES in their systems.

6.2. Impact of Customer Response Intensity

We now study the impact of customer response on the asymptotic accuracy of the LES announcement. In particular, we show that the accuracy of the LES announcement degrades with the intensity of customer response. We consider different abandonment-rate functions, and vary their parameters so as to increase the "intensity" of customer response to the announcements. To control for the effect of congestion, we hold the expected waiting time in the system fixed.

In modelling customer abandonment response to the announcements, we draw on the recent literature. In particular, Brown et al. (2005) and Mandelbaum and Zeltyn (2013) quantified the

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

23

k	ASE(LES)	Waiting time	RASE(LES)
4	4.37×10^{-3} $\pm 2.3 \times 10^{-5}$	0.240 $\pm 3.2 \times 10^{-4}$	0.275
2	7.40×10^{-3} $\pm 2.1 \times 10^{-5}$	0.408 $\pm 7.4 \times 10^{-4}$	0.211
1	1.24×10^{-2} $\pm 7.2 \times 10^{-5}$	0.684 $\pm 1.2 \times 10^{-3}$	0.163
0.2	4.17×10^{-2} $\pm 4.4 \times 10^{-4}$	2.40 $\pm 3.6 \times 10^{-3}$	0.0851

Table 2 Accuracy in the $M/M/N + M$ **model** with $\rho = 1.4$ and alternative k in (14).

ρ	ASE(LES)	Waiting time	RASE(LES)
1	2.88×10^{-3} $\pm 1.7 \times 10^{-5}$	8.96×10^{-2} $\pm 6.9 \times 10^{-4}$	0.599 $\pm 3.6 \times 10^{-4}$
1.2	8.22×10^{-3} $\pm 4.4 \times 10^{-5}$	0.429 $\pm 1.5 \times 10^{-3}$	0.211 $\pm 6.9 \times 10^{-5}$
1.4	1.24×10^{-2} $\pm 7.2 \times 10^{-5}$	0.684 $\pm 1.2 \times 10^{-3}$	0.163 $\pm 5.1 \times 10^{-5}$
1.6	1.51×10^{-2} $\pm 1.4 \times 10^{-4}$	0.878 $\pm 1.2 \times 10^{-3}$	0.140 $\pm 8.3 \times 10^{-5}$

Table 3 Accuracy in the $M/M/N + M$ **model** for $\theta(w)$ in (5) and alternative ρ.

effect of the announcements by statistically estimating the hazard-rate of the abandonment-time distribution and showing that customers typically become more impatient as delay announcements increase; e.g., see Figure 5 in Brown et al. (2005) and Figures 13 and 14 in Mandelbaum and Zeltyn (2013). Consistent with this evidence, we assume in §5 that the abandonment rate is an increasing function of the announcement.

Once more, we exclude balking from the system; we also let $\rho = 1.4$ and $N = 100$. To ensure robustness, we consider three functional forms for $\theta(w)$:

$$\theta_1(w) = b - e^{-aw}; \quad \theta_2(w) = aw + b; \quad \theta_3(w) = b + e^{aw} \text{ where } a > 0; \tag{15}$$

for example, letting $k = 1$ in (14) corresponds to $\theta_1(w)$ with $a = 1$ and $b = 1.5$. It is readily seen that the functions θ_1, θ_2, and θ_3 are all continuous and strictly increasing in w. Moreover, even though the sufficient conditions on balking behavior stated in §5.1 do not apply with a constant balking probability (equal to 0), it is readily seen that a unique equilibrium continues to exist in each case. We vary a and b in (15) to consistently have that $\bar{w} = \ln(\rho) = \ln(1.4)$; this is the steady-state fluid waiting time in a system with no customer response to the announcements, and with $\theta(w) = 1$ for all w. Increasing a amounts to increasing the intensity of customer response to the announcements.

In Table 4, we present estimates for ASE(LES) and RASE(LES) for each abandonment-rate function. In each case, $a = 0$ corresponds to a constant, announcement-independent, abandonment rate equal to 1. Table 4 clearly shows that RASE(LES) increases with a. That is, the LES announcement is less accurate with customer response in the system. For example, with θ_3, RASE(LES) increases from roughly 22% for $a = 0$, to roughly 42% for $a = 4$ (we do not increase a further to guarantee that $\theta_3(w) \geq 0$). Similarly, for θ_2, RASE(LES) increases to roughly 33% for $a = 10$.

Table 4 illustrates an interesting phenomenon: The asymptotic accuracy of the LES announcement does not degrade as extremely for function θ_1, as it does for functions θ_2 and θ_3. Indeed, as explained in §5, stochastic fluctuations, particularly around the equilibrium wait-time value,

24

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

	$\theta_1(w)$				$\theta_2(w)$				$\theta_3(w)$		
a	b	ASE(LES)	RASE(LES)	a	b	ASE(LES)	RASE(LES)	a	b	ASE(LES)	RASE(LES)
0	2	5.88×10^{-3} $\pm 2.3 \times 10^{-5}$	0.224	0	1	5.88×10^{-3} $\pm 2.3 \times 10^{-5}$	0.224	0	0	5.88×10^{-3} $\pm 2.3 \times 10^{-5}$	0.224
2	1.51	6.22×10^{-3} $\pm 3.2 \times 10^{-5}$	0.231	2	0.32	6.57×10^{-3} $\pm 2.2 \times 10^{-5}$	0.238	1	-0.4	6.35×10^{-3} $\pm 2.9 \times 10^{-5}$	0.235
4	1.3	6.27×10^{-3} $\pm 2.13 \times 10^{-5}$	0.230	4	-0.35	7.58×10^{-3} $\pm 3.7 \times 10^{-5}$	0.254	1.5	-0.66	6.76×10^{-3} $\pm 3.3 \times 10^{-5}$	0.243
8	1.1	6.15×10^{-3} $\pm 3.2 \times 10^{-5}$	0.227	8	-1.7	1.09×10^{-2} $\pm 6.5 \times 10^{-5}$	0.301	2	-0.96	7.48×10^{-3} $\pm 4.0 \times 10^{-5}$	0.255
10	1.03	6.08×10^{-3} $\pm 2.6 \times 10^{-5}$	0.226	10	-2.4	1.32×10^{-2} $\pm 9.3 \times 10^{-5}$	0.330	4	-2.8	1.99×10^{-2} $\pm 1.1 \times 10^{-4}$	0.421

Table 4 **Effect of varying the intensity of customer response on the asymptotic accuracy of LES in the** $M/M/100 + M$ **model with** $\rho = 1.4$ **and the abandonment-rate functions** θ_1, θ_2, **and** θ_3.

typically impact the accuracy of the LES announcement. Comparing the values of the derivatives of θ_1, θ_2, and θ_3, around $\bar{w}$, reveals that θ_1 changes more slowly than both θ_2 and θ_3. In other words, stochastic fluctuations around $\bar{w}$ have a relatively mild impact under θ_1, which ensures that the system state is relatively stable, and that the LES announcement is relatively accurate.

Our results show that the accuracy of the LES announcement is intimately tied not only to whether or not customers respond to the announcements, but also to how they do so. This substantiates the importance of incorporating customer response in the analysis of the system.

6.3. Time-Varying Arrivals

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

$$\lambda(u) = \bar{\lambda} + \bar{\lambda}\alpha \sin(\gamma u), \text{ for } 0 \leq u < \infty , \tag{16}$$

where $\bar{\lambda}$ is the average arrival rate and α is the relative amplitude. Given an appropriate constant staffing level, this arrival-rate function corresponds to alternating periods of underload and overload in the system. As pointed out by Eick et al. (1993), the parameters of the arrival-rate intensity function, $\lambda(u)$ in (16), should be interpreted relative to the mean service time. Then, we speak of γ as the relative frequency. Small (large) values of γ correspond to slow (fast) time-variability in the arrival process, relative to the service times.

γ	Cycle length	ASE(LES)	RASE(LES)
0		1.14×10^{-2} $\pm 3.8 \times 10^{-5}$	0.174
0.0436	144	1.18×10^{-2} $\pm 8.0 \times 10^{-5}$	0.176
0.0873	72	1.19×10^{-2} $\pm 9.5 \times 10^{-5}$	0.177
0.262	24	1.38×10^{-2} $\pm 6.1 \times 10^{-5}$	0.190
0.524	12	1.8×10^{-2} $\pm 6.6 \times 10^{-5}$	0.219
1.571	4	2.79×10^{-2} $\pm 7.8 \times 10^{-5}$	0.268

Table 5   **Effect of the arrival-rate frequency γ on the accuracy of the LES announcement.**
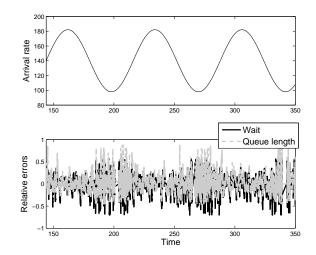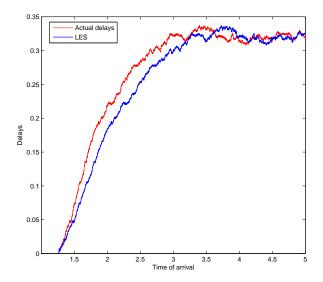

Figure 8   **Relative errors for the waiting times and queue lengths ($\gamma = 0.0873, \alpha = 0.3$).**

Accuracy of the LES Announcement. In Table 5, we study the effect of varying γ on the accuracy of the LES announcement. We also include values of the relative frequency as a function of the mean service time, assuming a 12 hour daily cycle, e.g., from 8:00AM to 8:00PM.

As before, we consider $b(w)$ and $\theta(w)$ in (4) and (5). The first row in Table 5 corresponds to the case with stationary arrivals, which we include here as a benchmark. Consistent with intuition, Table 5 clearly shows that the accuracy of the LES announcement deteriorates as γ increases. Indeed, the LES announcement performs poorly when the arrival rate changes rapidly over time, because delays then vary systematically over time. For example, RASE(LES) ranges from roughly 17% for $\gamma = 0.0436$ (slow time variation) to roughly 27% for $\gamma = 1.57$ (fast time variation). Interestingly, ASE(LES) appears to be roughly constant for different values of γ.

The conclusions that we draw from Table 5 are consistent with those in Ibrahim and Whitt (2011). They showed, in the context of delay announcements with no customer response, that the accuracy of the LES announcement degrades as arrival rates become more time variable. Table 5 shows that the same holds with announcement-dependent balking and abandonment as well.

Relative Errors of the Queue Length and Waiting Times. With a stationary arrival process, in both the QED and ED regimes, we established an important asymptotic result which unified our analysis throughout: The relative error in the LES prediction is small if, and only if, the relative error in the queue length is small. We now investigate whether this main result continues to hold with time-varying arrivals as well. We consider a system with a sinusoidal arrival-rate function as in (16), and with $N = 100$ servers. We let $\gamma = 0.0873$, $\alpha = 0.3$, and $\bar{\lambda} = 140$. As such, the arrival rate fluctuates from a minimum of 98 ($\rho = 0.98$) to a maximum of 182 ($\rho = 1.82$). In the bottom
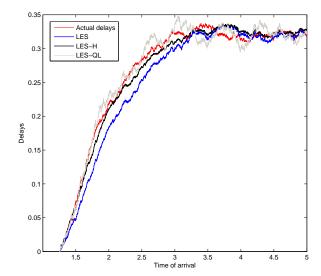
Figure 9   Actual and LES delays in the $M/M/5000 + M$ which is initially empty with $\rho = 1.4$ and $\theta(w) = 1.51 - e^{-2w}$.

Figure 10   Direct and adjusted LES announcements, LES-H and LES-QL, and actual delays in the same model as in **Figure 9**.

subplot of Figure 8, we plot the relative errors for the waiting times and for the queue lengths seen upon arrival in the system. The results in Figure 8 are based on one simulation path, rather than being averaged over multiple simulation replications as before. In the top subplot of Figure 8, we plot the arrival-rate function. Figure 8 shows that our asymptotic result continues to hold with time-varying arrival rates. Indeed, the relative errors in the queue lengths and waiting times increase and decrease in sync, as can be seen by their matching curves in the plots. We remove from the bottom subplot of Figure 8 the top and bottom 0.5% of the relative errors which correspond to dividing by very small values for the queue length and the waiting time.

6.4. Adjustments of the LES Announcement

Delay announcements are typically both noisy and biased. The noise is equal to the variance of the conditional waiting times (conditional on the information about current system state); the bias is the difference between the delay announcement and the expected conditional waiting time in the system. The ASE of the LES announcement captures both the aforementioned noise and bias.

As with time-varying arrivals (Figure 8), Figure 9 shows that the LES announcement is typically biased in the transient state of a given system. In Figure 9, we consider customer abandonment response according to $\theta_1(w)$ in (15) with $a = 2$ and $b = 1.51$; we let $N = 5,000$ to reduce the effect of stochastic noise. For example, in the transient state of a system which is initially empty, LES announcements are systematically biased downwards. In the online supplement, we derive an adjustment of the LES announcement, in a system with no customer response, which exploits

fluid-model dynamics. We show that this adjustment is more accurate (less biased) than the LES announcement in the transient state. The adjusted announcement, LES_a, is given by:

$$\text{LES}_a \text{ announcement} = \frac{1}{\theta} \ln(\rho + 1 - \rho e^{-\theta w_{LES}}) \ , \tag{17}$$

where w_{LES} is the direct LES announcement. In this section, we derive adjustments of the LES announcement in systems with customer response. Since the direct analysis of transient fluid-model dynamics in such systems is difficult, we derive heuristic adjustments instead.

Our adjusted LES announcement in a system with no customer response depends on the constant abandonment rate, θ, in the system, as shown in (17). Here, we derive a heuristic adjustment of the LES announcement by replacing θ (no response) with $\theta(0)$ (at the origin). That is, we propose the following adjustment to the direct LES announcement, w_{LES}:

$$\text{LES-H announcement} = \frac{1}{\theta(0)} \ln(\rho + 1 - \rho e^{-\theta(0)w_{LES}}) \ . \tag{18}$$

We tried replacing $\theta(0)$ in (18) by $\theta_{w_{LES}}$, but this lead to slightly less accurate delay announcements; this is why we exclude such an announcement from consideration here.

We also propose another adjusted LES announcement which exploits the queue-length seen upon arrival by the LES and new customers. Let QL_{LES} denote the queue length seen upon arrival by the LES customer, and let QL_n be the queue length seen by the new customer. Then, we study the accuracy of the following queue-length-based adjustment:

$$\text{LES-QL announcement} = w_{LES} \times \frac{QL_n}{QL_{LES}}. \tag{19}$$

Additionally, we considered announcements based on several past LES delays (either a pre-determined fixed number, or all LES delays occurring within a certain time window) experienced by successive customers in the system. We fit linear, quadratic, and exponential functions to those delays (as a function of the time of arrival to the system), and extrapolated those functions to the time of arrival of the current customer. We did so to obtain adjusted announcements based on additional past delays besides the most recent LES delay. Here, we do not include a separate discussion for those adjusted announcements because numerous simulation experiments indicated that they did not consistently perform better than the LES announcement.

In Table 6, we present estimates for both the ASE and bias of the LES announcement and the heuristic adjustments, LES-H and LES-QL. We let $N = 1,000$ and $\rho = 1.4$. We consider different simulation run lengths, but generally focus on the transient state of the system, which we assume starts empty. We also consider two abandonment-rate functions. Table 6 shows that both LES-H and LES-QL are less biased than the LES announcement and their ASE's are also smaller, irrespective of the abandonment-rate function considered. Based on Table 6, we can also compute the

28

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

$\theta(w) = 4w - 0.35$ (units of 10^{-4} for ASE and Bias)

Run length	ASE(LES)	Bias(LES)	ASE(LES-QL)	Bias(LES-QL)	ASE(LES-H)	Bias(LES-H)	Number Delayed
2000	1.87 ±1.7	95.6 ±34	0.759 ±0.24	-5.0 ±14	0.841 ±0.89	35.9 ±32	164 ±52
3000	12.5 ±4.0	300 ±55	2.48 ±0.71	-45.5 ±17.2	3.09 ±1.29	91.7 ± 53.2	932 ±59
5000	21.2 ±3.5	354 ±44	6.99 ±2.0	-78.3 ±47	11.4 ±2.8	219 ±40	2740 ±84
10,000	12.6 ±1.7	118 ±14	7.94 ±1.2	-42.5 ±13	7.93 ±1.3	84.7 ±16	7749 ±93

$\theta(w) = -2.8 + e^{4w}$ (units of 10^{-4} for ASE and Bias)

Run length	ASE(LES)	Bias(LES)	ASE(LES-QL)	Bias(LES-QL)	ASE(LES-H)	Bias(LES-H)	Number Delayed
2000	1.87 ±1.7	95.6 ±34	0.759 ±0.24	-5.0 ±14	0.841 ±0.89	35.9 ±32	165 ±52
3000	15.5 ±5.7	331 ±68	2.33 ±0.63	-30.4 ±14	4.42 ±2.4	124 ±66	910 ±54
5000	65.6 ±16	228 ±64	30.2 ±11	-223 ±42	45.1 ±13	169 ±55	2834 ±68
10,000	63.5 ±23	103 ±28	36.9 ±11	-160 ±49	45.8 ±17	92.0 ±19	7773 ±96

Table 6 **Accuracy of heuristic adjustments for the LES announcement in the $M/M/1000 + M$ model with $\rho = 1.4$ and alternative abandonment-rate functions.**

noise (i.e., conditional variance) in each of the predictions, and find that it is smaller with LES-QL and LES-H compared to LES. Table 6 also shows that ASE(LES-QL) is generally slightly smaller than ASE(LES-H). Since LES-QL is usually less biased than LES-H, as shown by Table 6, this implies that LES-QL announcements should be slightly more noisy than LES-H announcements. In Figure 10, we plot LES-H, LES, and actual delays for the same system as in Figure 9. Figure 10 nicely illustrates how the LES-H and actual delays closely match, particularly initially, and LES-QL announcements exhibit slightly stronger variations, consistently with Table 6.

In practical terms, selecting which predictor to implement, either LES-H or LES-QL, ultimately depends on the error measure that is of interest in the system. Indeed, if the manager is interested in reducing bias so that the announcements given are, on average, close to actual delays, then our experiments suggest that LES-QL is the better alternative since it reduces that bias. On the other hand, if the manager is interested in reducing the average square of the errors, so as to penalize

against both underestimation and overestimation in the announcements, then our experiments suggest that LES-H is the better alternative.

7. Conclusions

In this paper, we studied the problem of making accurate real-time delay announcements in large service systems. In particular, we focused on the LES delay announcement: This type of announcement is practically appealing because it depends solely on the history of delays in the system, i.e., it does not require any additional information about current system parameters.

There is ample empirical evidence showing that customers typically respond to delay announcements in practice; e.g., see Yu et al. (2014) and Aksin et al. (2015). Nevertheless, to the best of our knowledge, besides Armony et al. (2009) who focus solely on a fluid model of the system, there are no studies of how the customer response impacts the accuracy of the individual announcements. In this paper, we took a step towards filling that gap in the literature. In particular, we established the asymptotic accuracy of the LES announcement in a system with announcement-dependent balking and abandonment. Doing so is complicated mainly because customer response impacts system dynamics which, in turn, impact the future announcements made. For example, customers who are announced a very long delay may become very impatient and abandon rapidly. In consequence to this increase in customer abandonment, delays in the system decrease, which in turn decreases future delay announcements. In response to the decreased announcements, customers abandon less, which causes the delays in the system to increase again. Thus, future announcements will increase as well. The analysis of such a system involves a complex high-dimensional equilibrium since it is necessary to keep track of all customers in queue and their respective announcements.

Our theoretical results showed that the LES announcement is asymptotically accurate, i.e., with a large number of servers. Through our numerical study, we found that the LES announcement performs relatively poorly when the number of servers is very small, but that its accuracy improves rapidly as the number of servers increases. We also found that the relative asymptotic accuracy of LES improves as the system's congestion increases, which suggests that this type of announcement would be particularly useful in busy service contexts.

Our results also illustrated that customer response on one hand, and time-variation in the arrival rates, on the other hand, both lead to a degradation in the asymptotic accuracy of the LES announcement. We numerically investigated how the relative error in the queue length translates into the accuracy of LES announcement, and found that wait-time errors fluctuate less extremely than queue-length errors, and that the LES announcement should be accurate even when the queue-length error is not so small. This provided a practical dimension to our theoretical result establishing the equivalence between the stability of wait-time and queue-length relative errors.

This is particularly useful because real-time information about the queue length is routinely collected, e.g., as in the Automatic Call Distributor (ACD) of call centers.

There are several research directions that remain to be investigated. One such direction is to further analyze systems with time-varying arrival rates, and to provide theoretical support for our observations in §6.3. Another direction for future research is to consider multiple customer classes and multiple customer priorities. Those are often observed in real-life, particularly in hospital emergency departments where patients are often seen according to the severity of their ailments. In that setting, it would be interesting to study the effectiveness of the LES announcement and to develop appropriate adjustments, if need be. Yet another setting which is interesting to consider is that of queueing networks, which is also useful in representing service in a hospital context where patients sequentially go through several units for treatment. One could then think of other types of announcements which would be more appropriate in that setting, given the additional information.

References

Aksin, O.Z., Armony, M. and Mehrotra, V. 2007. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research, *Production and Operations Management*, 16(6), 665 – 688.

Aksin, Z., B. Ata, S. Emadi, and C.-L. Su. 2015. Impact of delay announcements in call centers: An empirical approach. *working paper.*

Allon, G, Bassambo, A. and I. Gurvich. 2012a. We will be right with you: managing customer with vague promises, *Operations Research.* 59(6): 1382-1394.

Allon, G, Bassambo, A. and I. Gurvich. 2012b. The impact of delaying the delay announcements. *Operations Research.* 59(5): 1198-1210.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and G.B. Yom-Tov. 2015. Patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, forthcoming.

Armony, M., C. Maglaras. 2004. Contact centers with a call-back option and real-time delay information, *Operations Research*, 52: 527 – 545.

Armony, M., N. Shimkin and W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonments. *Operations Research.* 57: 66-81.

Bertsimas, D. and D. Nakazato. 1995. The distributional Little's law and its applications. *Operations Research* 43.2: 298–310.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Operations Research*, 52(1): 17-34.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469), 36-50.

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

31

Eick, S., W.A. Massey, W. Whitt. 1993. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science.* 39(2): 241–252.

Fleming P. A., Stolyar B., Simon B. 1994. Proc. 2nd Internat. Conf. Telecomm. Syst. Mod. Anal (Nashville, TN), Heavy traffic limit for a mobile phone system loss model.

Garnett, O., A. Mandelbaum, M.I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Operations Management* 5: 79-141.

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information, *Management Science.* 53: 962-970

Gurvich, I. and W. Whitt. 2009. Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Mathematics of Operations Research*, 34:2, pp. 363-396.

Halfin, S. and Whitt, W. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research.* 29: 567-588.

Huang, J., A. Mandelbaum, H. Zhang, and J. Zhang. 2015. Refined Models for Efficiency-Driven Queues with Applications to Delay Announcements and Staffing, *working paper.*

Ibrahim, R. and W. Whitt. 2009. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science.* 55: 1729-1742.

Ibrahim, R. and W. Whitt. 2011. Real-Time Delay Estimation Based on Delay History in Many-Server Queues with Time-Varying Arrivals. *Production and Operations Management*, 20(5): 654-667.

Jagerman, D.L. 1974. Some properties of the Erlang loss function. *Bell System Tech J.*, 53: 525-551.

Jennings, O., Mandelbaum, A., W., Massey, and W. Whitt. 1996. Server Staffing to Meet Time-Varying Demand, *Management Science*, 42(10): 1383–1394.

Jouini, O., Z. Aksin and Y. Dallery. 2011. Call Centers with Delay Information: Models and Insights. Manufacturing & Service Operations Management, 13:534-548.

Jouini, O., Z. Aksin, F. Karaesmen, M.S. Aguir and Y. Dallery. 2015. Call center Delay Announcement Using a Newsvendor-Like Performance Criterion. *Production & Operations Management*, 24(4), 587-604.

Kang W., K. Ramanan. 2010. Fluid limits of many-server queues with reneging. *Annals of Applied Probability.* 20(6):2204-2260

Little, J. and S. C. Graves. Little's law. *Building Intuition,* 81-100, Springer US.

Mandelbaum, A. and G. Pats. 1995. State-Dependent Queues: Approximations and Applications, *Stochastic Networks*, IMA Volume 71, Editors F. Kelly and R.J. Williams, Springer-Verlag, 239-282.

Mandelbaum, A. and S. Zeltyn. 2013. Data Stories about (Im)Patient Customers in Tele-Queues. *Queueing Systems*, 75 (24), 115-146.

Plambeck, E., Bayati, M., Ang, E., Kwasnick, S., and M. Aratow. 2015. Forecasting Emergency Department Wait Times. *Working paper.*

Pang, G., R. Talreja and W.Whitt. 2007. Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues. *Probability Surveys*, 4, 193–267.

Puhalskii, A. A. 1994. On the invariance principle for the first passage time. *Math of Operations Research*, (19), pp. 946954.

Puhalskii, A. A. and Reiman, M. I. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Advances in Applied Probability.* 32: 564-595.

Reed, J. E, and A. R. Ward. 2008. Approximating the GI/GI/1+GI Queue with a Nonlinear Drift Diffusion, *Mathematics of Operations Research*, 33(3), 606-644.

Reiman, M. I. 1982. The heavy traffic diffusion approximation for Sojourn times in Jackson networks. R. L. Disney, T. J. Ott, eds. Applied ProbabilityComputer Science, *The Interface*, II. Birhauser, Boston, 409-422.

Senderovich A., Weidlich M., Gal A. and A. Mandelbaum. 2015. Queue Mining for Delay Prediction in Multi-Class Service Processes. *Working paper.*

Talreja R. and W.Whitt. 2009. Heavy-Traffic Limits for Waiting Times in Many-Server Queues with Abandonment. *Annals of Applied Probability,* 19(6): 2137-2175.

Reed, J. E, and T. Tezcan. 2012. Hazard Rate Scaling for the $GI/M/n + GI$ Queue, *Operations Research*, 60(4), 981-995.

Whitt, W. 1999a. Predicting queueing delays. *Management Science.* 45: 870–888.

Whitt, W. 1999b. Improving service by informing customers about anticipated delays. *Management Science.* 45: 192–207.

Whitt, W. Stochastic-Process Limits. 2002. Springer-Verlag, New York.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science.* 50: 1449–1461.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Operations Research* 54: 37–54.

Yu, Q., G. Allon, and A. Bassamboo. 2015. How do delay announcements shape customer behavior? An empirical study, *Management Science,* forthcoming.

Zhang, J. 2013. Fluid models of many-server queues with abandonment, *Queueing Systems,* 73(2): 147-193.

TECHNICAL APPENDIX

We present additional analytical results in this appendix to the main paper.

8. Asymptotic Results in the QED Regime: Proof of Theorem 1

We begin with the following lemma which establishes the tightness of $\{\sqrt{N}W^N(\tau_t^N)\}_{N\geq 1}$ and $\{Q^N(\tau_t^N)/\sqrt{N}\}_{N\geq 1}$, under the initial condition in Theorem 1, where $Q^N(s)$ denotes the queue length at time s in the Nth queueing system. We need to establish tightness because τ_t^N is a random variable, and it is not clear whether $\sqrt{N}W^N(\tau_t^N)$ and $Q^N(\tau_t^N)/\sqrt{N}$ converge as $N \to \infty$.

LEMMA 1. *For a fixed time t, $\{\sqrt{N}W^N(\tau_t^N)\}_{N\geq 1}$ and $\{Q^N(\tau_t^N)/\sqrt{N}\}_{N\geq 1}$ are tight.*

PROOF. By assumption, $\bar{Z}^N(0) = Z^N(0)$; thus, $\{(\bar{Z}^N(0) - N)/\sqrt{N}\}_{N\geq 1}$ is tight. Since $\bar{Q}^N(0)/\sqrt{N} = (\bar{Z}^N(0) - N)^+/\sqrt{N} = (Z^N(0) - N)^+/\sqrt{N}$, $\{\bar{Q}^N(0)/\sqrt{N}\}_{N\geq 1}$ is also tight by the continuity of the positive part function. Additionally, we can write the following:

$$\sqrt{N}\bar{W}^N(0) = \sqrt{N}\sum_{i=0}^{\bar{Q}^N(0)} X_i,$$

where X_i are independent and exponentially distributed with rate $N\mu + i\bar{\theta}$. Thus, $\sqrt{N}\bar{W}^N(0)$ is stochastically dominated by $(\bar{Q}^N(0) + 1)Y/\sqrt{N}$, where Y is exponentially distributed with rate μ. By assumption, $(\bar{Q}^N(0) + 1)/\sqrt{N}$ converges in distribution to a finite limit as $N \to \infty$. Thus, $(\bar{Q}^N(0)+1)Y/\sqrt{N}$ converges in distribution as well (assuming that Y is defined on the same probability space as $\bar{Q}^N(0)$), and $\sqrt{N}\bar{W}^N(0)$ is stochastically dominated by a sequence which converges weakly to a finite random variable. Thus, $\{\sqrt{N}\bar{W}^N(0)\}_{N\geq 1}$ is tight. Since tightness on products of separable metric spaces is characterized by the tightness of the individual components, we conclude that $(\bar{Q}^N(0)/\sqrt{N}, \sqrt{N}\bar{W}^N(0))$ is also tight. Given our construction, $\sqrt{N}W^N(\tau_t^N)$ is stochastically dominated by $\sup_{0\leq s\leq t}\sqrt{N}\bar{W}^N(s)$. Additionally, $Q^N(\tau_t^N)/\sqrt{N}$ is stochastically dominated by $\sup_{0\leq s\leq t}\bar{Q}^N(s)/\sqrt{N}$. By Theorem 2 and Theorem 3 of Garnett et al. (2002), we have that both supremum upper bounds converge weakly to finite random variables as $N \to \infty$. Tightness easily follows, and will be used subsequently to establish convergence for the quantities above. ∎

We are now ready to state and prove Proposition 1.

PROPOSITION 1. *For any fixed time point t,*

$$\tau_t^N \Rightarrow t \quad as \quad N \to \infty. \tag{20}$$

PROOF. By the tightness of $\{\sqrt{N}W^N(\tau_t^N)\}_{N\geq 1}$ (Lemma 1), we may conclude that

$$W^N(\tau_t^N) \Rightarrow 0 \quad \text{as} \quad N \to \infty. \tag{21}$$

Thus, to show that τ_t^N converges weakly to t, it is sufficient to establish that

$$\mathbb{P}(t - \tau_t^N - W^N(\tau_t^N) > \epsilon, \ i.o.) = 0 \quad \text{for every } \epsilon > 0, \tag{22}$$

where $t - \tau_t^N - W^N(\tau_t^N$ is the time elapsed since the LES customer entered service until the new arrival epoch t. Fix $\epsilon > 0$, and define the following events:

E_N: $t - \tau_t^N - W^N(\tau_t^N) > \epsilon$,

$E1_N$: At least one service completion occurs in the interval $(\tau_t^N + W^N(\tau_t^N), \tau_t^N + W^N(\tau_t^N) + 0.5\epsilon]$,

$E2_N$: At least one arrival occurs in the interval $[\tau_t^N + W^N(\tau_t^N) + 0.5\epsilon, \tau_t^N + W^N(\tau_t^N) + \epsilon]$,

$E3_N$: At least one arrival occurs in the interval $(\tau_t^N + W^N(\tau_t^N), \tau_t^N + W^N(\tau_t^N) + \epsilon]$,

A_N: All servers are busy at $\tau_t^N + W^N(\tau_t^N)$.

Then, the following relation holds:

$$E_N \subseteq ((E1_N \cap E2_N)^c \cup A_N^c) \cap (E3_N^c \cup A_N) . \tag{23}$$

This leads to:

$$\mathbb{P}(E_N) \leq \mathbb{P}((E1_N \cap E2_N)^c \cap A_N) + \mathbb{P}(E3_N^c \cap A_N^c) + \mathbb{P}((E1_N \cap E2_N)^c \cap E3_N^c) .$$

Further,

$$\mathbb{P}((E1_N \cap E2_N)^c \cap A_N) \leq \mathbb{P}(E1_N^c | A_N) + \mathbb{P}(E2_N^c | A_N) = \mathbb{P}(E1_N^c | A_N) + \mathbb{P}(E2_N^c).$$

Also, using the fact that λ^N satisfies (2), there exists a constant $C_2 > 0$ such that for large N

$$\mathbb{P}(E2_N^c) \leq e^{-C_2 \epsilon N}.$$

Additionally, if all the servers are busy, then the time until the next service completion is exponentially distributed with rate $N\mu$. Thus, we have that for large N

$$\mathbb{P}(E1_N^c | A_N) \leq e^{-C_1 \epsilon N}.$$

Also, note that there exists $C_3 > 0$ such that $\mathbb{P}(E3_N^c \cap A_N^c) + \mathbb{P}((E1_N \cap E2_N)^c \cap E3_N^c) \leq \mathbb{P}(E3_N^c) \leq e^{-C_3 \epsilon N}$. Hence, we have that

$$\sum_{N=1}^{\infty} \mathbb{P}(t - \tau_t^N - W^N(\tau_t^N) > \epsilon) < \infty.$$

Using the Borel-Cantelli lemma, we obtain (22). In conclusion, we obtain, by Theorem 11.4.5 of Whitt (2002), the following joint convergence since limits are deterministic:

$$(W^N(\tau_t^N), t - \tau_t^N - W^N(\tau_t^N) \Rightarrow (0,0) \quad \text{as} \quad N \to \infty;$$

this implies that

$$\tau_t^N \Rightarrow t, \quad \text{as} \quad N \to \infty , \tag{24}$$

as desired. ∎

We now prove, appealing to Proposition 1, that the relative error between the queue lengths seen upon arrival by the LES and new customer is also asymptotically negligible.

PROPOSITION 2. *For any fixed t,*

$$\frac{Q^N(t) - Q^N(\tau_t^N)}{\sqrt{N}} \Rightarrow 0, \ \text{as } N \to \infty.$$

PROOF. For each queueing system indexed by N, we consider two auxiliary queueing systems: (i) System L^N is an Erlang-B pure loss system (customers who cannot be served immediately are lost) with N servers and the same arrival and service rates as the original system; and (ii) System U^N is an Erlang-A system with N servers and the same arrival and service rates as the original system. There is no balking in system U^N and the abandonment rate there is constant and equal to $\underline{\theta}$.

Let $\underline{Q}^N(s)$, $\underline{Z}^N(s)$, $\underline{I}^N(s)$, and $\underline{W}^N(s)$ denote the queue length, number of customers in the system, number of idle servers, and virtual waiting time, at time s, in system L^N. Similarly, let $\overline{Q}^N(s)$, $\overline{Z}^N(s)$, $\overline{I}^N(s)$, and $\overline{W}^N(s)$ denote those same quantities in system U^N. By appropriately coupling the arrival, service, balking, and abandonment times, we can construct all systems on the same probability space such that if, for a given u, $\underline{Z}^N(u) \le Z^N(u) \le \overline{Z}^N(u)$, then the following inequalities hold for all $v \ge u$:

$$\underline{Z}^N(v) \le Z^N(v) \le \overline{Z}^N(v),$$
$$\underline{Q}^N(v) \le Q^N(v) \le \overline{Q}^N(v),$$
$$\underline{W}^N(v) \le W^N(v) \le \overline{W}^N(v),$$
$$\underline{I}^N(v) \ge I^N(v) \ge \overline{I}^N(v).$$

Lastly, we initialize systems L^N and U^N at time 0, for $N \ge 1$, as follows:

$$\underline{Z}^N(0) = \min\{Z^N(0), N\} \text{ and } \overline{Z}^N(0) = Z^N(0). \tag{25}$$

We will establish that $(Z^N(t) - Z^N(\tau_t^N))/\sqrt{N} \Rightarrow 0$ as $N \to \infty$. The stated result then follows immediately by the continuous mapping theorem. Our objective is to find upper and lower bounds which are aligned with $Z^N(\cdot)$ at τ_t^N and are tight in the vicinity of t. Therefore, we need to define two more auxiliary processes, $Z_H^N(s)$ and $Z_L^N(s)$ for $0 \le s \le t$, as follows:

$$Z_H^N(s) \equiv \bar{Z}^N(s) - \left(\bar{Z}^N(\tau_t^N) - Z^N(\tau_t^N)\right) \text{ and } Z_L^N(s) \equiv \underline{Z}^N(s) - \left(\underline{Z}^N(\tau_t^N) - Z^N(\tau_t^N)\right).$$

By construction, it is easy to verify that:

$$Z_L^N(\tau_t^N) = Z_H^N(\tau_t^N) = Z^N(\tau_t^N). \tag{26}$$

First, we need to prove that:

$$Z_L^N(s^N) - o(\sqrt{N}) \leq Z^N(s^N) \leq Z_H^N(s^N) + o(\sqrt{N}) \text{ for } s^N \in \{\tau_t^N, t\}. \tag{27}$$

We begin by proving the left-hand side of (27), and then restrict attention to proving the right-hand side of that equation here. To prove the left-hand side of (27): By definition,

$$Z_L^N(s) - Z^N(s) = \left(Z^N(\tau_t^N) - \underline{Z}^N(\tau_t^N) \right) - \left(Z^N(s) - \underline{Z}^N(s) \right). \tag{28}$$

By construction, for $s \in [\tau_t^N, t]$, the right-hand side of (28) is upper bounded by the difference in the number of departures (service completions, abandonment, and balking) between the original system and system L^N in the interval $[\tau_t^N, t]$. In particular,

$$\frac{1}{\sqrt{N}} \left(Z_L^N(s) - Z^N(s) \right) \leq \frac{1}{\sqrt{N}} \mathcal{L}_{\text{aban}}^N + \frac{1}{\sqrt{N}} \mathcal{L}_{\text{comp}}^N + \frac{1}{\sqrt{N}} \mathcal{L}_{\text{balk}}^N, \tag{29}$$

where, with a slight abuse of notation,

$$\mathcal{L}_{\text{aban}}^N \sim \text{Poisson}\left((t - \tau_t^N) \cdot \bar{\theta} \cdot \sup_{\tau_t^N \leq u \leq t} \bar{Q}^N(u) \right),$$

$$\mathcal{L}_{\text{comp}}^N \sim \text{Poisson}\left((t - \tau_t^N) \cdot \mu \cdot \sup_{\tau_t^N \leq u \leq t} \underline{I}^N(u) \right),$$

and

$$\mathcal{L}_{\text{balk}}^N \sim \text{Poisson}\left((t - \tau_t^N) \cdot \lambda^N \cdot b\left(\sup_{\tau_t^N \leq u \leq t} \bar{W}^N(u) \right) \right),$$

where "aban" stands for abandonment and "comp" for service completions. In particular, if Y is a non-negative random variable, then we take $X \sim \text{Poisson}(Y)$ to mean that $(X|Y = y)$ is Poisson distributed with mean y. We also need the following lemma.

LEMMA 2. *Let X^N, Y^N be two sequences of non-negative random variables such that $X^N \sim Poisson(Y^N)$ and $Y^N/\sqrt{N} \Rightarrow 0$ as $N \to \infty$. Then, $X^N/\sqrt{N} \Rightarrow 0$ as $N \to \infty$.*

PROOF. Let $\epsilon > 0$. We will establish that $\mathbb{P}\left(\frac{X^N}{\sqrt{N}} > \epsilon \right) \to 0$, as $N \to \infty$. Let $0 < \delta < \epsilon$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\frac{X^N}{\sqrt{N}} > \epsilon \right) \\
&= \mathbb{P}\left(\frac{X^N}{\sqrt{N}} > \epsilon \ \Big| \ \frac{Y^N}{\sqrt{N}} > \delta \right) \mathbb{P}\left(\frac{Y^N}{\sqrt{N}} > \delta \right) + \mathbb{P}\left(\frac{X^N}{\sqrt{N}} > \epsilon \ \Big| \ \frac{Y^N}{\sqrt{N}} \leq \delta \right) \mathbb{P}\left(\frac{Y^N}{\sqrt{N}} \leq \delta \right) \\
&\leq \mathbb{P}\left(\frac{Y^N}{\sqrt{N}} > \delta \right) + \mathbb{P}\left(\frac{X^N}{\sqrt{N}} > \epsilon \ \Big| \ \frac{Y^N}{\sqrt{N}} \leq \delta \right) \\
&\leq \mathbb{P}\left(\frac{Y^N}{\sqrt{N}} > \delta \right) + \mathbb{P}\left(\frac{Z^N}{\sqrt{N}} > \epsilon \right),
\end{aligned} \tag{30}$$

where $Z^N \sim \text{Poisson}(\delta\sqrt{N})$. Recall that $\delta < \epsilon$. Then by the law of large numbers and by the assumption that $\frac{Y^N}{\sqrt{N}} \Rightarrow 0$, we have that the right hand side of (30) converges to 0, as $N \to \infty$. $\blacksquare$

It follows from Lemma 2, Garnett et al. (2002), Proposition 1, and Jagerman (1974) who establishes diffusion limits of a pure loss system in the QED regime, that

$$\frac{1}{\sqrt{N}}\mathcal{L}^N_{\text{aban}} + \frac{1}{\sqrt{N}}\mathcal{L}^N_{\text{comp}} \Rightarrow 0 \text{ as } N \to \infty.$$

Thus, there remains to establish that $b\left(\sup_{\tau^N_t \leq u \leq t} \bar{W}^N(u)\right)$ is $O_p(1/\sqrt{N})$. The latter follows from Garnett et al. (2002) and the Lipschitz continuity of the balking probability function $b(\cdot)$. This completes the proof of the left-hand side of (27).

We now prove the right-hand side of (27). By definition,

$$Z^N(s) - Z^N_H(s) = \left(\bar{Z}^N(\tau^N_t) - Z^N(\tau^N_t)\right) - \left(\bar{Z}^N(s) - Z^N(s)\right). \tag{31}$$

By construction, for $s \in [\tau^N_t, t]$, the right-hand-side of (31) is upper bounded by the difference in the number of departures (service completions and abandonment) between system U^N and the original system in the interval $[\tau^N_t, t]$. In particular,

$$\frac{1}{\sqrt{N}}\left(Z^N(s) - Z^N_H(s)\right) \leq \frac{1}{\sqrt{N}}\mathcal{M}^N_{\text{aban}} + \frac{1}{\sqrt{N}}\mathcal{M}^N_{\text{comp}}, \tag{32}$$

where, with a slight abuse of notation, we define:

$$\mathcal{M}^N_{\text{aban}} \sim \text{Poisson}\left((t - \tau^N_t) \cdot \bar{\theta} \cdot \sup_{\tau^N_t \leq u \leq t} \bar{Q}^N(u)\right),$$

and

$$\mathcal{M}^N_{\text{comp}} \sim \text{Poisson}\left((t - \tau^N_t) \cdot \mu \cdot \sup_{\tau^N_t \leq u \leq t} \underline{I}^N(u)\right).$$

It follows from Lemma 2, Garnett et al. (2002), Proposition 1, and Jagerman (1974) establishing diffusion limits of a pure loss system in the QED regime, that:

$$\frac{1}{\sqrt{N}}\left(Z^N(s^N) - \bar{\bar{Z}}^N(s^N)\right) \Rightarrow 0, \quad s^N = \tau^N_t, t.$$

By (26) and (27), we have that:

$$Z^N_L(t) - Z^N_L(\tau^N_t) - o(\sqrt{N}) \leq Z^N(t) - Z^N(\tau^N_t) \leq Z^N_H(t) - Z^N_H(\tau^N_t) + o(\sqrt{N}),$$

Hence, by the definition of the processes Z^N_H and Z^N_L, we have that:

$$\underline{Z}^N(t) - \underline{Z}^N(\tau^N_t) - o(\sqrt{N}) \leq Z^N(t) - Z^N(\tau^N_t) \leq \bar{Z}^N(t) - \bar{Z}^N(\tau^N_t) + o(\sqrt{N}). \tag{33}$$

38

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

By Garnett et al. (2002), Proposition 1, Jagerman (1974), and a time change argument, we have that both sides of (33) converge weakly to 0 as $N \to \infty$, when divided by $\sqrt{N}$. That is, we established that the snapshot principle holds for both the lower and upper bound systems, L^N and U^N. Consequently, it must hold for our original system as well, as desired. ■

We complete the proof of Theorem 1 by stating and proving Proposition 3 which establishes an asymptotic relation between the waiting time and queue length seen upon arrival for both the LES and new customer. Appealing to Propositions 2 and 3, we can then establish that the relative error in the LES announcement is asymptotically negligible, under the initial condition in Theorem 1, as desired.

PROPOSITION 3. *For $s^N \in \{t, \tau_t^N\}$,*

$$\sqrt{N}\left(W^N(s^N) - \frac{Q^N(s^N) + 1}{N\mu}\right) \Rightarrow 0 \text{ as } N \to \infty.$$

PROOF. Let Y_i be a sequence of i.i.d. random variables which are exponentially distributed with expected value $\mathbb{E}Y = 1$ (we omit subscripts when the specific index is not important). Then, given that $\overline{\theta}$ and $\underline{\theta}$ bound the abandonment rate, the following holds:

$$\sum_{i=0}^{Q^N(s)} \frac{Y_i}{N\mu + Q^N(s)\overline{\theta}} \leq^{\mathcal{D}} \sum_{i=0}^{Q^N(s)} \frac{Y_i}{N\mu + i\overline{\theta}} \leq^{\mathcal{D}} W^N(s) \leq^{\mathcal{D}} \sum_{i=0}^{Q^N(s)} \frac{Y_i}{N\mu + i\underline{\theta}}.$$

Upper bound. We begin by establishing convergence for an upper bound of the difference in Proposition 3.

$$\sqrt{N}\left(W^N(s) - \frac{Q^N(s) + 1}{N\mu}\right) \leq^{\mathcal{D}} \sqrt{N}\left(\sum_{i=0}^{Q^N(s)} \frac{Y_i}{N\mu} - \frac{Q^N(s) + 1}{N\mu}\right) = \sqrt{N}\sum_{i=0}^{Q^N(s)} \frac{Y_i - 1}{N\mu} =: V^N.$$

Note that $\mathbb{E}V^N = 0$ since $\mathbb{E}Y^N = 1$. Also, using the conditional variance formula:

$$Var(V^N) = Var\left(\mathbb{E}\left(V^N \mid Q^N(s)\right)\right) + \mathbb{E}\left(Var\left(V^N \mid Q^N(s)\right)\right) = \frac{1}{N\mu^2}\mathbb{E}[Q^N(s) + 1],$$

where $Var(X)$ denotes the variance of random variable X. Additionally,

$$Q^N(s) \leq^{\mathcal{D}} \tilde{Q}^N(s), \tag{34}$$

where $\tilde{Q}^N(s)$ is the queue length, at time s, in an $M/M/N + M$ system with arrival rate λ^N and identical service and abandonment rates, both equal to $\min\{\mu, \underline{\theta}\}$. We let the initial state in this system be the same as in our original system. We need the following lemma.

LEMMA 3. *Let $X^N(s)$ be the number of customers in an $M/M/\infty$ system with arrival rate λ^N and service rate μ, at time s. Assume that $\frac{\mathbb{E}X^N(0) - N}{\sqrt{N}}$ is bounded. Then, there exists $c > 0$ such that $\mathbb{E}[(X^N(s) - N)^+] < c\sqrt{N}$.*

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

39

PROOF. Let $\lambda^N = N\rho^N$, where $\rho^N \to 1$ as $N \to \infty$, be the arrival rate to the system. Let $X^N(0)$ be the number of customers in the system at time 0. Let $s > 0$. Then, the total number of customers in the system at time s, $X^N(s)$, can be written as

$$X^N(s) = Y_1^N(s) + Y_2^N(s),$$

where $Y_1^N(s)$ is the number of customers who were present at time 0 and remain in the system at time s, and $Y_2^N(s)$ is the number of customers who have arrived after time 0 and remain in the system at time s. Thus, $Y_1^N(s)$ is binomial with parameters $X^N(0)$ and success probability $e^{-\mu s}$.

As in Eick et al. (1993), $Y_2^N(s)$ is Poisson distributed with rate $(\lambda^N/\mu)(1-e^{-\mu s}) = (N\rho^N/\mu)(1-e^{-\mu s})$. We can show the following.

LEMMA 4. *Let Y^N is a random variable that has Poisson distribution with mean $N\nu$. Then, there exist a constant C such that*

$$\mathbb{E}[(Y^N - N\nu)^+] \le (1+\nu)\sqrt{N}.$$

PROOF. The proof follows by noting that the following

$$
\begin{aligned}
\mathbb{E}[(Y^N - N\nu)^+] &= \int_0^\infty \mathbb{P}((Y^N - N\nu)^+ > x)dx \\
&= \int_0^{\sqrt{N}} \mathbb{P}((Y^N - N\nu)^+ > x)dx + \int_{\sqrt{N}}^\infty \mathbb{P}((Y^N - N\nu)^+ > x)dx \\
&\overset{(a)}{\le} \sqrt{N} + \int_{\sqrt{N}}^\infty \frac{\mathbb{E}[(Y^N - N\nu)^2]}{x^2}dx \\
&= \sqrt{N} + \int_{\sqrt{N}}^\infty \frac{N\nu}{x^2} \\
&= \sqrt{N} + \sqrt{N}\nu = \sqrt{N}(1+\nu),
\end{aligned}
$$

where inequality (a) follows from Markov's inequality. This completes the proof. ∎

Combining the above lemma with the fact that $X^N(0)$ is assumed to be bounded establishes the desired. ∎

By Lemma 3, we have that there exists $C > 0$ such that

$$\mathbb{E}\tilde{Q}^N(s) < C\sqrt{N}. \tag{35}$$

Finally we have that, for all $\epsilon > 0$,

$$
\begin{aligned}
\mathbb{P}(V^N > \epsilon) &\le \mathbb{P}(|V|^N > \epsilon) \\
&\le \frac{Var(V^N)}{\epsilon^2} \quad \text{(by Chebyshev's inequality)} \\
&\le \frac{\mathbb{E}Q^N(t)+1}{N\mu^2\epsilon^2} \\
&\le \frac{C\sqrt{N}}{N\mu^2\epsilon^2} \to 0, \quad \text{as } N \to \infty .
\end{aligned}
\tag{36}
$$

This implies that $V^N \Rightarrow 0$ as $N \to \infty$, which completes our proof of convergence for the upper bound. We are now ready to establish convergence for the lower bound.

Lower bound.

$$\sqrt{N}\left(W^N(s) - \frac{Q^N(s)+1}{N\mu}\right) \geq^{\mathcal{D}} \sqrt{N}\left(\sum_{i=0}^{Q^N(s)} \frac{Y_i}{N\mu+Q^N(s)\overline{\theta}} - \frac{Q^N(s)+1}{N\mu}\right)$$

$$= \sqrt{N}\left(\sum_{i=0}^{Q^N(s)} \frac{(Y_i-1)N\mu - Q^N(s)\overline{\theta}}{N\mu(N\mu+Q^N(s)\overline{\theta})}\right) \qquad (37)$$

$$= \frac{1}{1+\frac{Q^N(s)\overline{\theta}}{N\mu}} V^N - \sqrt{N}R^N,$$

where $V^N := \sqrt{N}\sum_{i=0}^{Q^N(s)} \frac{Y_i-1}{N\mu}$ as before, and $R^N := \frac{\overline{\theta}Q^N(s)(Q^N(s)+1)}{(N\mu+\overline{\theta}Q^N(s))N\mu}$. By (34) and (35) we have that

$$\frac{1}{1+\frac{Q^N(s)\overline{\theta}}{N\mu}} \Rightarrow 1, \text{ as } N \to \infty,$$

and that

$$-\sqrt{N}R^N \geq \frac{-\overline{\theta}\left(Q^N(s)(Q^N(s)+1)\right)\sqrt{N}}{(N\mu)^2} \Rightarrow 0 \quad \text{as} \quad N \to \infty.$$

Thus, $\sqrt{N}\left(W^N(s) - \frac{Q^N(s)}{N\mu}\right) \Rightarrow 0$, as desired. ∎

Combining the above results yields the proof for Theorem 1.

9. QED Regime with a General Abandonment Distribution

Consider the system where, depending on the announcement $w^N(\tau_t^N)$, the new arrival abandons according to a general distribution whose hazard rate is given by $h_{w^N(\tau_t^N)}(\cdot)$. Further, assume that $h_w(\cdot)$ is bounded from above and below, i.e.,

$$\underline{\theta} \leq h_w(y) \leq \overline{\theta} \text{ for all } y \geq 0.$$

We begin by establishing the following lemma, which will be used in the proof of our main theorem below.

LEMMA 5. *Consider a random variable U with hazard rate $h(\cdot)$ such that $\underline{\theta} \leq h(x) \leq \overline{\theta}$. Then,*

$$e^{-\overline{\theta}s} \leq \mathbb{P}(U > t+s | U > t) \leq e^{-\underline{\theta}s} .$$

PROOF. Using the definition of hazard rate, one can express the cumulative probability function for U, denoted by F_U, as follows:

$$F_U(t) = 1 - e^{-\int_0^t h(s)ds}.$$

Thus, we have that $\mathbb{P}(U > t) = e^{-\int_0^t h(s)ds}$. Based on this, we obtain

$$\mathbb{P}(U > t+s | U > t) = e^{-\int_t^{t+s} h(u)du}.$$

Using the fact that $\underline{\theta} \leq h(u) \leq \overline{\theta}$, for all u, completes the proof. ∎

THEOREM 3. *If $(Z^N(0) - N)/\sqrt{N}$ is tight, then*

$$\sqrt{N}|w^N(t) - w^N(\tau_t^N)| \Rightarrow 0,$$

as $N \to \infty$.

PROOF. The proof follows along the lines of to the Theorem 1.

Tightness of $Q^N(\tau_t^N)/\sqrt{N}$. By noting that the abandonment times in the given system are stochastically larger than for an exponential distribution with rate $\underline{\theta}$, we can proceed as in Lemma 1 in the main paper.

The time elapsed between the arrival of LES customer τ_t^N is asymptotically close to t. Again, using Lemma 5 along with the arguments in Theorem 1 we can prove that τ_t^N is close to t by establishing that:

$$\mathbb{P}(t - \tau_t^N - w^N(\tau_t^N) > \epsilon \ i.o.) = 0 \quad \text{for all } \epsilon > 0. \tag{38}$$

Two Systems. For the construction of the two systems, as in Proposition 2, we use the result of Lemma 5 which states that the excess distribution of the abandonment time is stochastically bounded above and below by exponential random variables with rate $\bar{\theta}$ and $\underline{\theta}$, respectively. Thus, the two systems constructed in the proof of Theorem 1 will also bound our system. Following the arguments of the proof of Theorem 1 completes the proof. ∎

10. Asymptotic Results in the ED Regime: Proof of Theorem 2
10.1. Stopping time.

Let $\epsilon > 0$. For the Nth system, we define σ^N and α^N as follows:

$$\sigma^N = \inf\{s : |\bar{Z}^N(s) - \bar{z}| > \epsilon\} \ , \quad \text{and} \tag{39}$$

$$\alpha^N = \min\{\sigma^N, T\} \ , \tag{40}$$

where $\bar{Z}^N(s) = Z^N(s)/N$. Then, for all N, we must have that on the interval $[0, \alpha^N)$:

$$|\bar{Z}^N(s) - \bar{z}| \le \epsilon. \tag{41}$$

10.2. Stochastic boundedness of the stopped waiting times.

We now establish that if (41) holds, then $W^N(s)$ will be stochastically bounded on $[0, \alpha^N)$ as well. To this aim, we define the function $\Gamma(\cdot)$ as follows, for all $x \ge 1$:

$$\Gamma(x) = \frac{1}{\theta} \ln\left(\frac{\mu + \theta(x-1)}{\mu}\right). \tag{42}$$

We will need this function $\Gamma(\cdot)$ to establish a relationship between the waiting time and the scaled number of customers in the system. For a customer arriving at a time $s \in [0, \alpha^N)$:

$$W^N(s) = \sum_{j=N}^{Z^N(s)} \frac{Y_j}{\mu N + \theta(j - N)}, \tag{43}$$

where Y_j is exponentially distributed with mean 1. Let $\delta > 0$, and define

$$U_j \equiv \frac{Y_j - 1}{\mu N + \theta(j - N)} \quad \text{for } j \ge N.$$

Then, subtracting $E[W^N(s)]$, we can write:

$$\mathbb{P}\left(\left|W^N(s) - \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j-N)}\right| > \delta\right) = \mathbb{P}\left(\left|\sum_{j=N}^{Z^N(s)} U_j\right| > \delta\right).$$

We can also establish the following lemma.

LEMMA 6. *There exists $C_1 > 0$ such that for sufficiently large N,*

$$\mathbb{P}\left(\left|\sum_{j=N}^{Z^N(s)} U_j\right| > \delta\right) < e^{-C_1\delta N}. \tag{44}$$

PROOF.

Note that

$$\mathbb{P}\left(\left|\sum_{j=N}^{Z^N(s)} U_j\right| > \delta\right) = \mathbb{P}\left(\sum_{j=N}^{Z^N(s)} U_j > \delta\right) + \mathbb{P}\left(\sum_{j=N}^{Z^N(s)} U_j < -\delta\right) \tag{45}$$

We shall show the bound on the first term; the bound on the second term proceeds similarly and will therefore be omitted. Let $C > 0$. Then by Chebyshev's inequality,

$$\mathbb{P}\left(\sum_{j=N}^{Z^N(s)} U_j > \delta\right) \leq \mathbb{P}\left(\sum_{j=N}^{Z^N(s)} CU_j - C\delta > 0\right) \leq \mathbb{E}\left[\exp\left(C\sum_{j=N}^{Z^N(s)} U_j - C\delta\right)\right]. \tag{46}$$

Note that for $N \leq j$:

$$\mathbb{E}[\exp(CU_j)] = \mathbb{E}\left[\exp\left(\frac{C(Y_j - 1)}{\mu N + \theta(j-N)}\right)\right] \tag{47}$$

$$= \frac{\mu N + \theta(j-N)}{\mu N + \theta(j-N) - C}\exp\left(-\frac{C}{\mu N + \theta(j-N)}\right). \tag{48}$$

Taking logarithms on both sides and choosing $C = C_0 N$ where $0 < C_0 < \mu$ we obtain

$$\log\mathbb{E}[\exp(C_0 U_j)] = \log\left(\frac{\mu + \theta(\frac{j}{N} - 1)}{\mu + \theta(\frac{j}{N} - 1) - C_0}\right) - \frac{C_0}{\mu + \theta(\frac{j}{N} - 1)} \tag{49}$$

$$< \frac{C_0}{\mu + \theta(\frac{j}{N} - 1) - C_0} - \frac{C_0}{\mu + \theta(\frac{j}{N} - 1)}, \tag{50}$$

$$= \frac{C_0^2}{(\mu + \theta(\frac{j}{N} - 1) - C_0)(\mu + \theta(\frac{j}{N} - 1))} \tag{51}$$

$$\leq \frac{C_0^2}{(\mu - C_0)\mu}, \tag{52}$$

where the first inequality follows from the fact that $\log(1 + x) < x$ and second by noting that $N \leq j$. Hence, we obtain

$$\log\mathbb{P}\left(\sum_{j=N}^{Z^N(s)} U_j > \delta\right) < \frac{C_0^2}{(\mu - C_0)\mu}\mathbb{E}[Z^N - N] - C_0 N\delta \tag{53}$$

$$< -N\left[C_0\delta - \frac{C_0^2}{(\mu - C_0)\mu}\bar{z}\right] \tag{54}$$

where the last inequality holds using (41) from the paper. Note that one can choose C_0 small enough that makes

$$\left[C_0\delta - \frac{C_0^2}{(\mu - C_0)\mu}\bar{z}\right] > 0,$$

Thus, we have that there exist $C' > 0$ such that

$$\mathbb{P}\left(\sum_{j=N}^{Z^N(s)} U_j > \delta\right) < \exp(-C'N\delta).$$

Similarly, we can show that there exist $C'' > 0$ such that

$$\mathbb{P}\left(\sum_{j=N}^{Z^N(s)} U_j < -\delta\right) < \exp(-C''N\delta).$$

Combining both inequalities, we obtain that there exists $C_1 > 0$ such that:

$$P\left(\left|\sum_{j=N}^{Z^N(s)} U_j\right| > \delta\right) < e^{-C_1\delta N}.$$

■

From (44), we deduce that for any $\delta > 0$, some integer N_0, and some $M < \infty$:

$$\sum_{N=1}^{\infty} \mathbb{P}\left(\left|W^N(s) - \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)}\right| > \delta\right) \leq M + \sum_{N=N_0}^{\infty} \mathbb{P}\left(\left|W^N(s) - \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)}\right| > \delta\right)$$

$$< M + \sum_{N=N_0}^{\infty} e^{-C_1\delta N} < \infty.$$

By the Borel-Cantelli lemma, we obtain that for all $s \in [0, \alpha^N)$:

$$\left|W^N(s) - \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)}\right| \to 0 \text{ almost surely as } N \to \infty.$$

Now, let's write:

$$\left|W^N(s) - \Gamma(\bar{Z}^N(s))\right| = \left|W^N(s) - \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)} + \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)} - \Gamma(\bar{Z}^N(s))\right|$$

$$\leq \left|W^N(s) - \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)}\right| + \left|\sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j - N)} - \Gamma(\bar{Z}^N(s))\right|$$

We have just shown that the first part on the right-hand side converges to 0 almost surely as $N \to \infty$; there remains to show that the same holds for the second part. For this, we need the following lemma:

44

Ibrahim, Armony, and Bassamboo: *Asymptotic Accuracy of the LES Predictor*
Article submitted to ; manuscript no. (Please, provide the mansucript number!)

LEMMA 7. *For $n \geq 0$, and any $\alpha \in \mathbb{R}$:*

$$\sum_{j=0}^{n} \frac{1}{n\alpha + j} = \ln\left(\frac{1+\alpha}{\alpha}\right) + O(1/n).$$

PROOF. For this, first note that

$$\sum_{j=0}^{n} \frac{1}{\lceil n\alpha \rceil + j} \leq \sum_{j=0}^{n} \frac{1}{n\alpha + j} \leq \sum_{j=0}^{n} \frac{1}{\lfloor n\alpha \rfloor + j}.$$

There remains to show that both upper and lower bounds converge as desired, which can be done as follows.

$$\sum_{j=0}^{n} \frac{1}{\lfloor n\alpha \rfloor + j} = \sum_{k=\lfloor n\alpha \rfloor}^{\lfloor n\alpha \rfloor + n} \frac{1}{k} = \sum_{k=1}^{\lfloor n\alpha \rfloor + n} \frac{1}{k} - \sum_{k=1}^{\lfloor n\alpha \rfloor - 1} \frac{1}{k},$$

$$= \ln(\lfloor n\alpha \rfloor + n) - \ln(\lfloor n\alpha \rfloor - 1) + O(1/n)$$

$$= \ln\left(\frac{\lfloor n\alpha \rfloor + n}{\lfloor n\alpha \rfloor - 1}\right) + O(1/n).$$

Proceeding similarly for the lower bound, we obtain that:

$$\sum_{j=0}^{n} \frac{1}{\lceil n\alpha \rceil + j} = \ln\left(\frac{\lceil n\alpha \rceil + n}{\lceil n\alpha \rceil - 1}\right) + O(1/n).$$

It's not hard to see that, letting $n \to \infty$ both bounds converge to $\ln\left((1+\alpha)/\alpha\right)$ as desired. ∎

Based on Lemma 7, we can deduce that:

$$\sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j-N)} = \sum_{k=0}^{Z^N(s)-N} \frac{1}{\mu N + \theta k} = \frac{1}{\theta} \sum_{k=0}^{Z^N(s)-N} \frac{1}{(\mu/\theta)N + k}.$$

$$= \frac{1}{\theta} \ln\left(\frac{\bar{Z}^N(s) - 1 + \mu/\theta}{\mu/\theta}\right) + O(1/N)$$

$$= \frac{1}{\theta} \ln\left(\frac{\mu + \theta(\bar{Z}^N(s) - 1)}{\mu}\right) + O(1/N)$$

$$= \Gamma(\bar{Z}^N(s)) + O(1/N).$$

Based on the above, we get that:

$$\left| \sum_{j=N}^{Z^N(s)} \frac{1}{\mu N + \theta(j-N)} - \Gamma(\bar{Z}^N(s)) \right| \to 0 \text{ almost surely as } N \to \infty.$$

That is,

$$\left| W^N(s) - \Gamma(\bar{Z}^N(s)) \right| \to 0 \text{ almost surely as } N \to \infty.$$

Since the null limit above has (trivially) continuous sample paths on $[0, \alpha^N)$, we also have almost sure convergence in the uniform topology. That is, we have that:

$$\lim_{N \to \infty} \|W^N(s) - \Gamma(\bar{Z}^N(s))\| = 0 \text{ almost surely over } [0, \alpha^N), \tag{55}$$

where we use the notation $||\cdot||$ to represent sup norm over the time interval $[0, \alpha^N)$. Noting that $\Gamma(\bar{z}) = \bar{w}$, and due to (41), we obtain by using a Taylor expansion argument that:

$$\lim_{N \to \infty} ||W^N(s) - \bar{w}|| \leq \Gamma'(\bar{z})\epsilon + O(\epsilon^2) \text{ almost surely over } [0, \alpha^N). \tag{56}$$

10.3. Divergence of the stopping time.

Since the LES announcement is the waiting time for some customer, and since $\bar{b}(\cdot)$ is a monotone decreasing function, we obtain, based on (56), that the arrival rate to the Nth system is bounded above and below by $N\lambda\bar{b}(\bar{w} + \Gamma'(\bar{z})\epsilon + O(\epsilon^2))$ and $N\lambda\bar{b}(\bar{w} - \Gamma'(\bar{z})\epsilon - O(\epsilon^2))$, respectively on $[0, \alpha^N)$ (we consider that $O(\epsilon^2)$ denotes a positive quantity). Using Mandelbaum and Pats (1995), we know that the scaled number in the system $\bar{Z}^N(s)$, for $s \in [0, \alpha^N)$, satisfies

$$\lim_{N \to \infty} \bar{Z}^N(s) \leq \frac{1}{\theta}(\lambda\bar{b}(\bar{w} + \Gamma'(\bar{z})\epsilon + O(\epsilon^2)) - \mu) + 1 \tag{57}$$

$$= \frac{1}{\theta}(\lambda\bar{b}(\bar{w}) - \mu) + 1 + \frac{\lambda\bar{b}'(\bar{w})\Gamma'(\bar{z})\epsilon}{\theta} + O(\epsilon^2) \tag{58}$$

$$= \bar{z} + \frac{\lambda\bar{b}'(\bar{w})\Gamma'(\bar{z})\epsilon}{\theta} + O(\epsilon^2) \text{ almost surely.} \tag{59}$$

Similarly, we can obtain a corresponding lower bound. Consequently, we have that:

$$\lim_{N \to \infty} ||\bar{Z}^N(s) - \bar{z}|| \leq \frac{\lambda\bar{b}'(\bar{w})\Gamma'(\bar{z})\epsilon}{\theta} + O(\epsilon^2) \text{ almost surely.}$$

Noting that $\Gamma'(\bar{z}) = 1/\lambda\bar{b}(\bar{w})$, we obtain that

$$\lim_{N \to \infty} ||\bar{Z}^N(s) - \bar{z}|| \leq \frac{\bar{b}'(\bar{w})\epsilon}{\bar{b}(\bar{w})\theta} + O(\epsilon^2) \text{ almost surely.}$$

Now, assume that (10) holds. Combining this with the fact that $\bar{Z}^N(s)$ cannot jump by more than $1/N$, we obtain for large N and small $\epsilon > 0$, the following:

$$\left|\bar{Z}^N(\alpha^N) - \bar{z}\right| < \epsilon. \tag{60}$$

Thus, we must have that $\alpha^N < \sigma^N$ for large N, and that $T < \sigma^N$. Since the above holds for all small enough $\epsilon > 0$ and any $T > 0$, we obtain the following:

$$\lim_{N \to \infty} ||\bar{Z}^N(s) - \bar{z}||_{[0,T]} = 0 \text{ almost surely }, \tag{61}$$

as desired. This establishes that the relative error in the number of customers in the system (or, equivalently, the queue length) is asymptotically negligible uniformly over compact sets. The asymptotic accuracy of LES follows by applying the continuous mapping theorem using $\Gamma(\cdot)$:

$$\lim_{N \to \infty} ||W^N(s) - \bar{w}||_{[0,T]} = 0 \text{ almost surely }, \tag{62}$$

which completes the proof.

11. ED Regime with Announcement-Dependent Abandonment

We consider the $M/M/N+M$ system in the ED limiting regime. We assume that customers respond to the announcements via announcement-dependent balking and abandonment. We establish the asymptotic accuracy of LES in that case. The main theorem and its proof are largely similar to Theorem 2. First, we characterize the corresponding equilibrium fluid behavior, as in §5.2 of the main paper.

11.1. Fluid Steady-State Equilibrium

Let $\bar{z}$ denote an equilibrium fluid content in the system. Then, $\bar{w}$ and $\bar{z}$ must satisfy the two following simple equations:

$$\lambda \bar{b}(\bar{w}) = \mu + \theta(\bar{w})(\bar{z} - 1) , \tag{63}$$

$$\bar{w} = \frac{1}{\theta(\bar{w})} \ln\left(1 + \frac{\theta(\bar{w})(\bar{z} - 1)}{\mu}\right) . \tag{64}$$

Sufficient conditions for the existence and uniqueness of this equilibrium were stated in §5.1.

THEOREM 4. *For the $M/M/N + M$ model in the ED heavy-traffic limiting regime with announcement-dependent balking and abandonment,*

If

$$\frac{Z^N(0)}{N} \Rightarrow \bar{z} \quad in \quad (63) \quad and \quad (64) \quad as \quad N \to \infty , \tag{65}$$

then

$$||W^N(t) - W^N(\tau_t^N)||_{[0,T]} \to 0 \quad as \quad N \to \infty \text{ almost surely} , \tag{66}$$

under the condition that

$$\lambda \bar{b}'(\bar{w}) + \frac{|\theta'(\bar{w})|}{\theta(\bar{w})} \lambda \bar{b}(\bar{w}) - \frac{\mu|\theta'(\bar{w})|}{\theta(\bar{w})} < \theta(\bar{w})\lambda \bar{b}(\bar{w})(1 - \theta'(\bar{w})K) , \tag{67}$$

and

$$K|\theta'(\bar{w})| < 1 . \tag{68}$$

PROOF. The proof is similar to that for Theorem 2, so we will be brief. We define the three stopping times, for $\epsilon > 0$, $\delta > 0$, and $T > 0$:

$$\sigma^N = \inf\{s : |\bar{Z}^N(s) - \bar{z}| > \epsilon\} , \tag{69}$$

$$\nu^N = \inf\{s : |W^N(s) - \bar{w}| > \delta\} , \tag{70}$$

$$\alpha^N = \min\{\sigma^N, T\} . \tag{71}$$

Define the following function, paralleling (42), for $x \geq 0$ and $y \geq 0$:

$$\Gamma(x, y) = \frac{1}{y} \ln\left(\frac{\mu + y(x - 1)^+}{\mu}\right) .$$

We note that on the interval $[0, \alpha^N)$,

$$|\theta(W^N(s)) - \theta(\bar{w})| \le |\theta'(\bar{w})|\delta + O(\delta^2) .$$

Using Taylor series expansion, we can show that:

$$\limsup_{N \to \infty} ||W^N(s) - \bar{w}|| \le \epsilon \left. \frac{\partial \Gamma}{\partial x} \right|_{(\bar{z}, \bar{w})} + \theta'(\bar{w})\delta \left. \frac{\partial \Gamma}{\partial y} \right|_{(\bar{z}, \bar{w})} .$$

We know that

$$\left. \frac{\partial \Gamma}{\partial x} \right|_{(\bar{z}, \bar{w})} = \frac{1}{\lambda \bar{b}(\bar{w})} ,$$

and

$$\left. \frac{\partial \Gamma}{\partial y} \right|_{(\bar{z}, \bar{w})} = -\frac{\bar{w}}{\theta(\bar{w})} + \frac{1}{\theta(\bar{w})} \frac{\bar{z} - 1}{\lambda \bar{b}(\bar{w})} .$$

Now, let

$$K = \left| \frac{-\bar{w}}{\theta(\bar{w})} + \frac{1}{\theta(\bar{w})} \frac{\bar{z} - 1}{\lambda \bar{b}(\bar{w})} \right| .$$

Further,

$$\lim_{n \to \infty} \bar{Z}^N(s) \le \frac{1}{\theta(\bar{w})(1 - \frac{|\theta'(\bar{w})|\delta}{\theta(\bar{w})})} (\lambda \bar{b}(\bar{w}) + \lambda |\bar{b}'(\bar{w})|\delta + O(\delta^2) - \mu) + 1 , \tag{72}$$

$$= \frac{1}{\theta(\bar{w})} \left(1 + \frac{|\theta'(\bar{w})|\delta}{\theta(\bar{w})} + O(\delta^2) \right) \left(\lambda \bar{b}(\bar{w}) + \lambda |\bar{b}'(\bar{w})|\delta + O(\delta^2) - \mu \right) + 1 \tag{73}$$

$$= \bar{z} + \delta \left(\frac{1}{\theta(\bar{w})} (\lambda |\bar{b}'(\bar{w})| + \frac{|\theta'(\bar{w})|}{\theta(\bar{w})} \lambda \bar{b}(\bar{w}) - \mu \frac{|\theta'(\bar{w})|}{\theta(\bar{w})}) \right) + O(\delta^2) . \tag{74}$$

To ensure that $\alpha^N \le \sigma^N$ and $\alpha^N < \nu^N$ for large N, we need the following conditions to hold:

$$\delta \left(\frac{1}{\theta(\bar{w})} (\lambda |\bar{b}'(\bar{w})| + \frac{|\theta'(\bar{w})|}{\theta(\bar{w})} \lambda \bar{b}(\bar{w}) - \mu \frac{|\theta'(\bar{w})|}{\theta(\bar{w})}) \right) < \epsilon , \tag{75}$$

and

$$\frac{1}{\lambda \bar{b}(\bar{w})} \epsilon + \theta'(\bar{w}) K \delta < \delta . \tag{76}$$

Equations (75) and (76) will be satisfied if:

$$|\theta'(\bar{w}) K| < 1 ,$$

and

$$\left(\frac{1}{\theta(\bar{w})} \left(\lambda |\bar{b}'(\bar{w})| + \frac{|\theta'(\bar{w})|}{\theta(\bar{w})} \lambda \bar{b}(\bar{w}) - \mu \frac{|\theta'(\bar{w})|}{\theta(\bar{w})} \right) \right) \frac{1}{\lambda \bar{b}(\bar{w})(1 - \theta'(\bar{w}) K)} < 1 .$$

The above inequalities can be restated as:

$$\lambda \bar{b}'(\bar{w}) + \frac{|\theta'(\bar{w})|}{\theta(\bar{w})} \lambda \bar{b}(\bar{w}) - \frac{\mu |\theta'(\bar{w})|}{\theta(\bar{w})} < \theta(\bar{w}) \lambda \bar{b}(\bar{w})(1 - \theta'(\bar{w})K) \ , \tag{77}$$

and

$$K|\theta'(\bar{w})| < 1 \ . \tag{78}$$

∎