Received: 9 May 2020 | Revised: 28 August 2020 | Accepted: 14 October 2020

DOI: 10.1002/nav.21960

RESEARCH ARTICLE

WILEY

Staffing many-server queues with autoregressive inputs

Xu Sun[1] | **Yunan Liu**[2]

[1]Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida, USA

[2]Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina, USA

Correspondence

Yunan Liu, Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695.

Email: yliu48@ncsu.edu

Abstract

Recent studies reveal significant overdispersion and autocorrelation in arrival data at service systems such as call centers and hospital emergency departments. These findings stimulate the needs for more practical non-Poisson customer arrival models, and more importantly, new staffing formulas to account for the autocorrelative features in the arrival model. For this purpose, we study a multiserver queueing system where customer arrivals follow a doubly stochastic Poisson point process whose intensities are driven by a Cox–Ingersoll–Ross (CIR) process. The nonnegativity and autoregressive feature of the CIR process makes it a good candidate for modeling temporary dips and surges in arrivals. First, we devise an effective statistical procedure to calibrate our new arrival model to data which can be seen as a specification of the celebrated expectation–maximization algorithm. Second, we establish functional limit theorems for the CIR process, which in turn facilitate the derivation of functional limit theorems for our queueing model under suitable heavy-traffic regimes. Third, using the corresponding heavy traffic limits, we asymptotically solve an optimal staffing problem subject to delay-based constraints on the service levels. We find that, in order to achieve the designated service level, such an autoregressive feature in the arrival model translates into notable adjustment in the staffing formula, and such an adjustment can be fully characterized by the parameters of our new arrival model. In this respect, the staffing formulas acknowledge the presence of autoregressive structure in arrivals. Finally, we extend our analysis to queues having customer abandonment and conduct simulation experiments to provide engineering confirmations of our new staffing rules.

KEYWORDS

autocorrelation, heavy-traffic approximations, many-server queues, mean-reverting process, non-Poisson arrivals, optimal staffing, parameter uncertainty, queues with customer abandonment

1 | INTRODUCTION

Poisson arrival is one of the most prevalent assumptions in queueing theory. Evidence that supports its validity is provided by Brown et al. (2005) and Kim and Whitt (2014). A Poisson input facilitates the mathematical analysis and produces insights on capacity planning for service systems. For example, the performance of an $M_t/G/\infty$ model with a nonhomogeneous Poisson arrival process and infinite service capacity has a simple expression which gives rise to the celebrated square-root-staffing rule. Although natural and convenient from a mathematical point of view, the Poisson assumption does not always align well with real-life data. Indeed, a growing body of empirical research has shown that the variance of the arrival count over a fixed time period tends to dominate its mean, a common feature known as overdispersion (see e.g., Jongbloed & Koole, 2001; Mathijsen et al., 2018). This phenomenon violates the fundamental property that underpins the Poisson input assumption and can potentially affect performance evaluation and choice of staffing rules. Indeed, it has been shown by He et al. (2016) that, in order to achieve a desired quality-of-service level for

queues with overdispersed arrivals, one must make significant adjustments on the staffing level.

It is found that the overdispersion can sometimes be explained by the strong autocorrelation of arrival counts over successive periods during the day (Ibrahim et al., 2016; Ibrahim & L'Ecuyer, 2013; Shen & Huang, 2008; Ye et al., 2019). Such intraday dependencies may arise due to a variety of reasons. For example, when a firm sends notification letters/emails to its group of customers, or runs a promotion on social media, there could be a temporary surge in the number of calls for the purpose of inquiry; in call centers designated for emergency services, a common event (e.g., a wide-scale power outage) may trigger bursts of calls within a short period of time, resulting in a much higher call volumes over that time duration. Redials and reconnects may also give rise lead to a certain degree of dependency in call arrival counts (see Ding et al., 2015). These features of dependencies observed in practice serve as the primary motivation for this work.

The central theme of this paper is to investigate how the autocorrelation structure of an arrival process can affect the performance of a stochastic service system. Specifically, we model customer arrival process $A(t)$ as a doubly stochastic Poisson process (DSPP) of which the intensities are driven by a possibly time-dependent Cox–Ingersoll–Ross (CIR) process. Specifically, we assume

$$A(t) \equiv \Pi_a\left(\int_0^t \lambda(u)du\right) \quad t \geq 0, \tag{1}$$

where $\Pi_a(\cdot)$ is a unit-rate Poisson point process and $\lambda(t)$ is a stochastic process satisfying the following stochastic differential equation (SDE)

$$d\lambda(t) = \kappa(\alpha - \lambda(t))dt + \sigma\sqrt{\lambda(t)}d\mathcal{B}(t), \tag{2}$$

where κ, σ and α are model parameters and $\mathcal{B}(\cdot)$ is a standard Brownian motion that is independent of the point process $\Pi(\cdot)$. In fact, (2) represents the dynamics of a CIR process with mean-reversion speed κ, mean-reversion level α, and volatility rate parameter σ. For brevity, hereinafter we refer to the above arrival process simply as M_{CIR}.

To visualize the impact of the autoregressive feature of the arrivals on the queueing performance, we consider an $M_{\text{CIR}}/M/s$ example with different volatility parameters. See Figure 1 for the simulation results of the waiting-time distributions. It is apparent from the plot that the degree of variability can exert considerable influence on the system performance. In particular, higher variability in arrival times leads to heavier tails and hence longer waits. We will later show that these parameters also play an important role in the staffing decision.

The CIR process is typically used to model instantaneous interest rates (see, e.g., Dias & Shackleton, 2011; Moreno & Platania, 2015). Because of its nice properties, the CIR process can be leveraged to model the randomly varying intensities of a counting process. First, the CIR process will never become negative under appropriate regularity condition (known as the Feller condition). Second, the process periodically spikes but has a tendency to return to its mean level. Lastly, the process is very amenable to statistical inference. For example, the method of maximum likelihood estimation can be easily implemented for this process (see, e.g., Overbeck & Rydén, 1997). These properties make the CIR process a natural candidate for modeling the unpredictable and temporary fluctuations observed in arrival data at modern call centers.

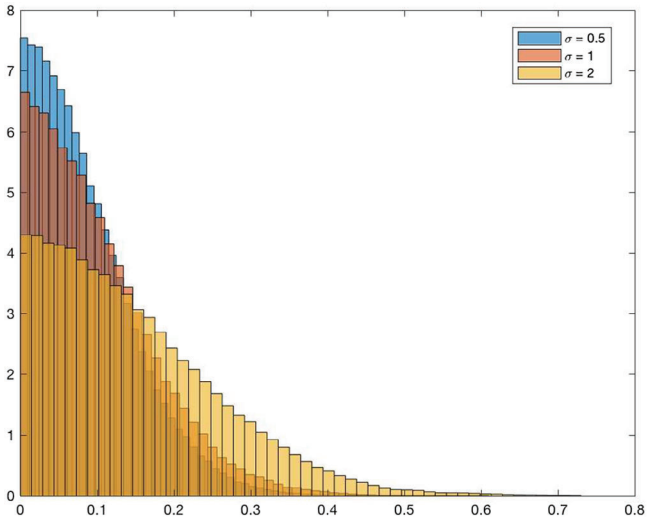

FIGURE 1 Estimated waiting-time distributions of an $M_{\text{CIR}}/M/s$ queue from computer simulation with $\alpha = 100$, $\kappa = 1$, $s = 100$, service rate $\mu = 1$ [Colour figure can be viewed at wileyonlinelibrary.com]

We are by no means the first to use the CIR process to model the intensities of a counting process. The use of a stationary CIR process to model the intensity of call arrivals has been proposed by Zhang et al. (2013) and Zhang et al. (2014). In Zhang et al. (2014) the authors demonstrate that the proposed traffic model can faithfully reproduce the behavior of interest observed in practice and derive the scaling limit for the arrival process concerning "convergence in marginal distribution." By contrast, we show that the arrival process (with proper scaling) converges weakly to a limit in the Skorokhod topology, a stronger version of Theorem 2 in Zhang et al. (2014); in addition, we propose practical staffing rules based on the resulting heavy-traffic approximation.

Our paper relates to a growing body of research that considers random arrival rate. Brown et al. (2001) develop an autoregressive model for the arrival rate that can capture the correlation across successive time periods. Built upon the work of Harrison and Zeevi (2005) and Whitt (2006) uses fluid-model analysis to derive staffing solutions for a call center with uncertain arrival rate and employee absenteeism. Bassamboo et al. (2010) too consider a capacity sizing problem for a call center facing significant uncertainty in call volume. Our paper differs from theirs in three key aspects. First, Bassamboo et al. focus on an "uncertainty-dominated" regime meaning that the noise in arrival rate is in the same order as the mean, whereas we assume the noise in arrival

rate to be of smaller magnitude relative to the mean value. Second, the arrival models are different—despite assuming a random arrival intensity, they consider the value of the arrival rate to be fixed once it is realized; in contrast, we view the arrival rate itself as a stochastic process possessing an autoregressive feature as encoded in the CIR model assumption. Third, the mathematical treatments are different—they solve the staffing problem by adopting a newsvendor formulation whereas we seek to directly adjust the extant square-root staffing rule to account for added variability via the approach of heavy-traffic approximations. More recently, an economic model to aid staffing decisions in the presence of random arrival rates (with a co-sourcing option) has been developed by Koçağa et al. (2015). In contrast, we do not consider economic models; instead, we work directly with performance measures associated with customer delays and abandonments. Infinite-server models with Hawkes arrival process have been studied in Gao and Zhu (2018), Daw and Pender (2018) and Koops et al. (2017) analyze an infinite-server queue fed by a Cox process of which the intensities are driven by a short-rate process. Unlike Hawkes or short-rate process which exhibits jumps and decays exponentially over time, our intensity process is continuous, positive and mean-reverting. These properties are especially suitable for modeling traffic sources with smoothly changing intensity. Moreover, the CIR process is more parsimonious as the main features are characterized by fewer parameters, each having clear physical interpretation. It is worth noting that both Hawkes and CIR processes are special cases of affined point processes as considered by Zhang et al. (2015).

Moreover, instead of studying infinite-server models, we deal exclusively with constrained systems in which customer delay and abandonment can actually occur. Admittedly, an infinite-server queue can sometimes be used to approximate queueing systems with many servers, yet a model with infinite capacity provides little indication of how the temporary dips and surges in arrivals can affect delay-related metrics. On the other hand, real-world service systems tend to operate in a resource-constrained environment. This is especially true for modern call centers and healthcare settings where the challenge is to translate service-quality metrics (often expressed in terms of delay statistics) into concrete staffing decisions. An analysis of large-scale constrained systems, as we do here, sheds light on how the autocorrelation structure of the arrival process affects key performance indicators such as queue length and customer waiting time. Ultimately, we hope that our findings can help service providers make informed staffing decisions in the presence of stochastic demand fluctuations.

Typically, arrival models that possess autoregressive features are amenable to short-term forecasting, thereby being useful for service systems where the staffing level can be easily adjusted in a real-time fashion to handle unexpected shifts in demand. In the present study, we do not intend to incorporate forecasting into short-term capacity planning decisions; we do, however, perceive this as an interesting area for future research. In this respect, we feel that our proposed solutions are more suitable for systems with inflexible staffing.

1.1 | Our contributions

- First, we establish what we believe the first functional weak law of large numbers (FWLLN) and functional central limit theorem (FCTL) for the DSPP arrival process with a CIR rate (see Lemmas 3.1 and 3.2). Because CIR processes are widely used in many application domains, these functional limit theorems can be of independent interest.
- Second, to investigate the impact of the DSPP arrivals with CIR rate on the system performance and optimal staffing decisions, we first provide the *many-server heavy-traffic* (MSHT) FCLT limit for a critically loaded $M_{\mathrm{CIR}}/M/s$ model. In particular, we show that the limit of the headcount process is a piecewise-linear Gaussian process driven by a superposition of a Brownian motion and an integrated Ornstein–Uhlenbeck (OU) process. Hence the limit of the headcount process is not a diffusion process, which stands in contrast to He et al. (2016) where despite a general arrival process the limit of the headcount process is indeed a diffusion process.
- Third, using the FCLT result of the $M_{\mathrm{CIR}}/M/s$ queue, we solve an optimal staffing problem for the $M_{\mathrm{CIR}}/M/s$ queue subject to a delay-based constraint, asymptotically as the system scale increases. In particular, we show that the traditional square-root staffing is ineffective in achieving desired performance target in the presence of autoregressive arrivals; accordingly, we propose a refined staffing formula and illustrate through extensive numerical experiments the effectiveness of the our new staffing rule. We then extend the framework to an $M_{\mathrm{CIR}}/M/s+M$ with customer abandonment and discuss the implications of the M_{CIR} arrival on staffing decisions.

1.2 | Organization

In Section 2.1, we formally introduce the DSPP with CIR-driven intensities and describe the corresponding $M_{\mathrm{CIR}}/M/s$ model. In Section 2.2, we discuss possible ways to simulate and calibrate DSPPs with CIR-driven intensities. In Section 3 we introduce the asymptotic framework and establish the MSHT FCLT. In Section 4 we propose novel staffing rules in the presence of autoregressive inputs based on the FCLT and we confirm the effectiveness of our staffing formulas by conducting numerical studies. In Section 5 we extend our analysis to the $M_{\mathrm{CIR}}/M/s+M$ model having customer abandonment. All proofs are given in Section 6. We give concluding remarks in Section 7.

1.3 | Notations

We denote by $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{N}$, respectively, the sets of all real numbers, nonnegative reals and nonnegative integers. We use

$\lceil a \rceil$ to denote the least integer that is greater than or equal to a and z_α denote the quantile value from a standard normal distribution at α. For a real-value function f, we write $f[x_1, x_2]$ as shorthand for $f(x_2) - f(x_1)$. We use $\mathbf{e}$ to denote the constant function of one. Let $(\mathcal{D}([0,\infty), \mathbb{R}), J_1)$ denote the space of càdlàg (right continuous with left limits) functions equipped with the Skorokhod J_1 topology, and write "$\Rightarrow$" for weak convergence. All random entities introduced in this paper are supported by a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

2 | A QUEUEING MODEL WITH AUTOREGRESSIVE ARRIVALS

In this section we formally introduce the $M_{\text{CIR}}/M/s$ model fed by arrivals according to a DSPP with CIR-driven intensities. We also discuss algorithms to simulation sample paths of a DSPP with CIR intensity and explain how its model parameters can be calibrated from arrival data.

2.1 | The $M_{\text{CIR}}/M/s$ model

We study the $M_{\text{CIR}}/M/s$ queueing model having arrivals according to a DSPP described by (1) with CIR-driven intensity process $\lambda(\cdot)$ given by (2), and i.i.d. service times following an exponential distribution with rate μ. Throughout, we assume that the following Feller condition is satisfied so that $\lambda(\cdot)$ is always positive.

Assumption 1 *The model parameters κ, σ and α satisfy $2\kappa\alpha \geq \sigma^2$.*

To facilitate the presentation, we summarize below some key properties of the DSPP that will prove useful in the subsequent analysis.

1. (Markov property) The process $(\lambda(t), A(t))$ is Markovian with respect to the natural filtration $(\mathcal{F}_t)_{t\geq 0}$, and the intensity process $\lambda(t)$ itself is also Markovian. For all time intervals $(t_1, t_2]$,
$$\mathbb{E}[A[t_1,t_2]|\mathcal{F}_{t_1}] \overset{a.s.}{=} \mathbb{E}\left[\int_{t_1}^{t_2} \lambda(u)du|\mathcal{F}_{t_1}\right],$$
where we have defined $A[t_1, t_2] \equiv A(t_2) - A(t_1)$.
2. (Martingale property) By the definition of the intensity process $\lambda(\cdot)$ in (2), we have that $A(t) - \int_0^t \lambda(u)du$ is a square integrable martingale with quadratic variable given by $\int_0^t \lambda(u)du$, so that
$$\left(A(t) - \int_0^t \lambda(u)du\right)^2 - \int_0^t \lambda(u)du$$
is also a martingale.

To proceed, we stipulate that the system adopts a work-conserving policy; that is, no customers wait in queue if there is an available server. Let $Q(t)$ denote the number of customers in queue at time t, Furthermore, we use $E(t)$ to represent the number of customers that have entered service, all up to time t. By flow conservation,

$$Q(t) = Q(0) + A(t) - E(t). \tag{3}$$

In addition, let $B(t)$ be the number of busy servers at time t and $D(t)$ be the cumulative number of customer that have departed *due to service completion* up to time t. Again by flow conservation, we have

$$B(t) = B(0) + E(t) - D(t). \tag{4}$$

Finally, let $X(t)$ denote the head-count process recording the total number of customers in the system (both in queue and in service). Adding up (3) and (4) yields

$$X(t) = Q(t) + B(t) = X(0) + A(t) - D(t). \tag{5}$$

Alternatively, one can derive (5) directly from flow conservation.

It remains to specify the staffing levels (number of servers). In practice, staffing levels are selected to trade off operational efficiency and service quality. Here we follow a constraint-satisfaction approach; that is, the system operator or service provider specifies a performance metric and then assigns the least staffing level that satisfies the target. Of particular interest is a constraint on the probability of delay

$$\mathbb{P}(V(t) > 0) \leq \varrho, \tag{6}$$

where $V(t)$ represent the potential waiting time at time t, that is, the waiting time of an arriving customer at time t assuming the customer has infinite patience. It has long been known that to stabilize the delay probability, the system would have to be *critically loaded* and has negligible delay ($V(t) \approx 0$)(see, e.g., Feldman et al., 2008; Garnett et al., 2002). For our purpose, we propose a modified version of the classical square-root staffing (SRS) rule as given below

$$s \equiv \lceil \alpha/\mu + \sqrt{\alpha} c^\star \rceil,$$

where $c^\star$ is some constant traditionally referred to as the safety-staffing coefficient.

2.2 | Simulation and calibration of DSPP with CIR intensities

The CIR process is very easy to simulate. Indeed, there is an exact simulation algorithm by sampling chi-square random variables (see, e.g., Section 3.4 in Glasserman, 2013). To develop an intuitive understanding of the CIR-driven arrival process, in Figure 2 we depict the simulated sample paths of three CIR intensity processes, one for each scenario as covered in Figure 1. The sample path is obtained by adopting the exact simulation algorithm as detailed in Glasserman (2013, p. 124). Then arrivals are generated by using the

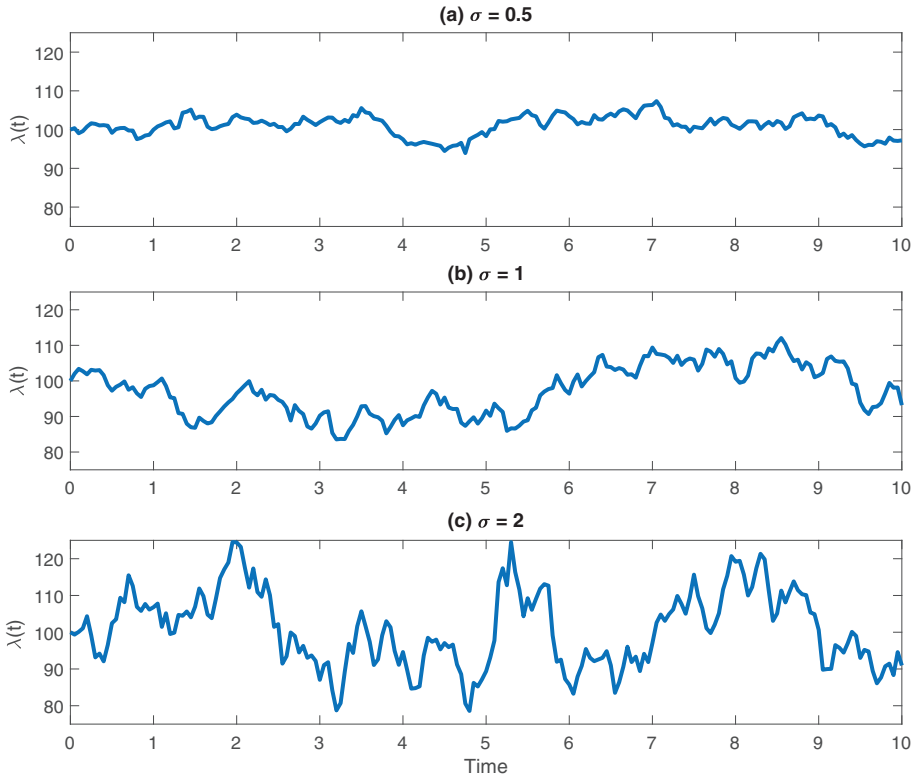

FIGURE 2 Sample paths of CIR-driven intensity processes with $\alpha = 100$, $\kappa = 1$ and $\sigma = 0.5,1,2$ [Colour figure can be viewed at wileyonlinelibrary.com]

thinning method for nonhomogeneous Poisson process (see, e.g., Ross, 1996, Chap. 2). We observe enlarging the parameter value of σ would lead to an increase in the magnitude of the random fluctuations, thereby bringing in more variability into the system. Moreover, by increasing (decreasing) the parameter value of κ, one can raise (reduce) the speed at which the intensity reverts to the long-run mean. Therefore it follows that the CIR process not only possesses analytical tractability but also permits versatile correlation structure.

As mentioned in the introduction, estimation of parameters for a CIR process based on realized trajectories can be done fairly easily. Parameter estimation for a DSPP with CIR intensity based on observed arrival counts is, however, not so straightforward, due to the underlying intensity process being not directly observable. To address this, one may resort to various techniques dealing with missing data. One typical example is the expectation–maximization (EM) algorithm designed to infer parameters in statistical models, where the model output depends on unobserved data (see, e.g., Lange, 2010, Chap. 13). The algorithm iterates between an expectation (E) step, which produces the expectation of the log-likelihood evaluated using the current parameter estimates, and a maximization (M) step, which seeks parameters maximizing the expected log-likelihood found in the E step.

To apply the EM in the present setting, we may regard the CIR intensity as roughly constant on a sufficiently small time interval. As a result, we can divide the time horizon into L smaller intervals and act as if the intensity equals some (unknown) constant λ_l on interval l, for $l = 1, \ldots, L$. Let N_l denote the number of arrivals observed on the lth interval and write $\mathbf{N} \equiv \{N_1, \ldots, N_L\}$. Similarly, we define $\lambda \equiv \{\lambda_1, \ldots, \lambda_L\}$. Now, for a set of unknown parameters $\Theta \equiv (\kappa, \sigma, \alpha)$, the likelihood function can be expressed as

$$\mathcal{L}(\Theta; \lambda, \mathbf{N}) = p(\lambda|\Theta)p(\mathbf{N}|\lambda).$$

Note that the above likelihood function admits an analytical expression, as both densities on the right-hand side, namely, $p(\lambda|\,\Theta)$ and $p(\mathbf{N}|\,\lambda)$, are in closed form. Given the current estimates of the parameters $\Theta^{(k)}$, the E step then calculates the expected value of the log-likelihood function $q(\Theta|\,\Theta^{(k)})$with respect to the distribution of λ given $\Theta^{(k)}$; that is,

$$q(\Theta|\Theta^{(k)}) = \mathbb{E}_{\lambda|\Theta^{(k)}}[\log \mathcal{L}(\Theta; \lambda, \mathbf{N}].$$

Next, the M step seeks the parameters that maximize this quantity; that is,

$$\Theta^{(k+1)} \equiv \arg\max_{\Theta} q(\Theta|\Theta^{(k)}).$$

Once this iterative procedure converges, the final output Θ^* will be used as estimates for the unknown model parameters.

3 | MANY-SERVER HEAVY-TRAFFIC ANALYSIS

The presence of a stochastic arrival rate makes an exact analysis of the queueing system extremely difficult. This leads us to apply fairly standard approximation techniques used in the extant literature. In particular, we assume that the system is facing high demand volume and has a large number of servers. Below, we formally introduce our asymptomatic framework

and perform some preliminary analysis in Section 3.1. The main results are presented in Section 3.2.

3.1 | Asymptotic framework

We consider an asymptotic framework in which the long-run average demand volume grows to infinity, that is, $\alpha \to \infty$ for α given as in (2). Following the convention in the literature, we will use n in place of α as the scaling parameter. More precisely, we let

$$\alpha_n \equiv n \tag{7}$$

be the mean-reversion level of the intensity process in the nth stochastic model. Accordingly, we subscript all relevant notation with n to capture the dependence on this scaling parameter n. For example, A_n denotes a DSPP with intensity process λ_n satisfying

$$d\lambda_n(t) = \kappa(\alpha_n - \lambda_n(t))dt + \sigma\sqrt{\lambda_n(t)}d\mathcal{B}(t). \tag{8}$$

Note that κ and σ are fixed—this sequence of CIR processes indexed by the scaling parameter n shares common mean-reversion speed and volatility rate. It is readily checked that with this scaling, the mean value of the intensity process $\lambda_n(\cdot)$ and the number of arrivals over any fixed time period blow up linearly by a factor of n. The remainder of this section is devoted to showing that a sequence of properly scaled intensity processes converges weakly to a Gaussian process. For that purpose, we define

$$\overline{\lambda}_n(t) \equiv \lambda_n(t)/n \quad \text{and} \quad \widehat{\lambda}_n(t) \equiv \sqrt{n}(\overline{\lambda}_n(t) - 1).$$

The result below establishes the FCLT for the sequence of intensity processes.

Lemma 3.1 (FWLLN and FCLT for the arrival intensity process). *Suppose that the intensity process for the n*th *model follows* (8). *If, in addition, there is convergence of the initial distribution at time* 0, *that is, if*

$$\widehat{\lambda}_n(0) \Rightarrow \widehat{\lambda}(0) \quad in \quad \mathbb{R} \quad as \quad n \to \infty, \tag{9}$$

then we have the joint convergence

$$(\overline{\lambda}_n(t), \widehat{\lambda}_n(t)) \Rightarrow (\mathbf{e}, \widehat{\lambda}(t)) \quad in \quad \mathcal{D}^2 \quad as \quad n \to \infty, \tag{10}$$

where $\widehat{\lambda}(\cdot)$ *satisfies the stochastic integral equation*

$$\widehat{\lambda}(t) = \widehat{\lambda}(0) - \kappa\int_0^t \widehat{\lambda}(u)\mathrm{d}u + \sigma\mathcal{B}(t). \tag{11}$$

Hence the limit of the arrival-rate process is an OU process whose solution admits a closed-form expression:

$$\widehat{\lambda}(t) = e^{-\kappa t}\widehat{\lambda}(0) + \sigma\int_0^t e^{-\kappa(t-u)}d\mathcal{B}(u). \tag{12}$$

To proceed, we define the scaled versions of the arrival process:

$$\overline{A}_n(t) \equiv A_n(t)/n \quad \text{and} \quad \widehat{A}_n(t) \equiv n^{-1/2}(A_n(t) - nt). \tag{13}$$

Per our previous discussion, it is natural to center $A_n(t)$ around nt. We will show in Lemma 3.2 that this centering indeed gives rise to meaningful limit.

Lemma 3.2 (FWLLN and FCLT for the arrival process). *The centered and normalized version of the arrival process* $\widehat{A}_n$ *satisfies a FCLT*:

$$\widehat{A}_n(t) \Rightarrow \widehat{A}(t) \equiv \mathcal{B}_\lambda(t) + \mathcal{K}(t) \quad in \quad \mathcal{D} \quad as \quad n \to \infty \tag{14}$$

for

$$\mathcal{K}(t) \equiv \int_0^t \widehat{\lambda}(u)\mathrm{d}u, \tag{15}$$

where $\widehat{\lambda}(\cdot)$ *is given in* (11) *and* $\mathcal{B}_\lambda(\cdot)$ *is a standard Brownian motion independent of* $\mathcal{K}(\cdot)$. *As an immediate consequence, we have the FWLLN*

$$\overline{A}_n(t) \Rightarrow t \quad in \quad \mathcal{D} \quad as \quad n \to \infty$$

jointly with (14).

According to Lemma 3.2, the diffusion-scaled arrival process $\widehat{A}_n(\cdot)$ converges to a Gaussian process which is characterized by two independent terms. The first term is a Brownian motion that arises from the inherent variability in the Poisson process Π_a, while the second term is an integrated OU process that stems from the stochasticity of the intensity process. For $s \le t$ the covariance between $\widehat{A}(s)$ and $\widehat{A}(t)$ can be computed using the following formula

$$\mathrm{Cov}(\widehat{A}(s), \widehat{A}(t)) = \frac{\sigma^2}{\kappa^3}(\kappa s - 1 + e^{-\kappa s} + e^{-\kappa t}) - \frac{\sigma^2}{2\kappa^3}(e^{-\kappa(t-s)} + e^{-\kappa(t+s)}) + s.$$

In particular, the formula for the variance is given by

$$\mathrm{Var}(\widehat{A}(t)) = \left(1 + \frac{\sigma^2}{\kappa^2}\right)t - \frac{3\sigma^2}{2\kappa^3} + \frac{2\sigma^2}{\kappa^3}e^{-\kappa t} - \frac{\sigma^2}{2\kappa^3}e^{-2\kappa t}. \tag{16}$$

When $\sigma = 0$, we have $\mathrm{Var}(A_n(t)) = nt$. So with deterministic arrival rate, the variance of the number of arrivals up to time t is equal to its mean. This is the level of variability that staffing levels are typically chosen to handle. If, however, $\sigma = 2$ and $\kappa = 1$, then from (16) it follows that the variance can be roughly five times the mean.

3.2 | Many-server heavy-traffic limits

To proceed, it is convenient and natural to consider a sequence of queueing systems, indexed by the scaling parameter n. In the nth model, customers arrive to the system according to A_n, that is, $A_n(t)$ represents the number of arrivals over $[0, t]$. The service rate μ is held fixed, but the staffing process is allowed to grow linear with n, so that the corresponding staffing function satisfies

$$s_n = \lceil n/\mu + \sqrt{n}c \rceil, \tag{17}$$

where c is a design parameter to be determined to meet the performance target (6).

3.3 | Heavy-traffic scalings

For the headcount and queue-length processes, define their centered and normalized versions as follows:

$$\widehat{X}_n(\cdot) \equiv n^{-1/2}(X_n(\cdot) - s_n), \quad \widehat{Q}_n(\cdot) \equiv n^{-1/2}Q_n(\cdot) \quad \text{and} \quad \widehat{V}_n(\cdot) \equiv \sqrt{n}V_n(\cdot).$$

The theorem below establishes the FCLT results showing that the above diffusion-scaled processes converge weakly to their corresponding limits.

Theorem 3.1 (FCLT for the critically-loaded $M_{\text{CIR}}/M/s$ model). *Suppose customers arrive according to the DSPP $A_n(\cdot)$ with intensity process $\lambda_n(\cdot)$ given by* (8)*, the system is staffed according to* (17)*. If in addition,*

$$\widehat{X}_n(0) \Rightarrow \widehat{X}(0)$$

jointly with (9)*, then*

$$(\widehat{X}_n, \widehat{Q}_n, \widehat{V}_n) \Rightarrow (\widehat{X}, \widehat{Q}, \widehat{V}) \quad \text{in} \quad \mathcal{D}^3 \quad \text{as} \quad n \to \infty,$$

jointly with (10)*, in which $\widehat{V}(t) = \widehat{Q}(t) = [\widehat{X}(t)]^+$, and the limiting process $\widehat{X}$ satisfies*

$$\widehat{X}(t) = \widehat{X}(0) - \mu c t + \mu \int_0^t [\widehat{X}(u)]^- \mathrm{d}u + \mathcal{K}(t) + \sqrt{2}\mathcal{B}_0(t), \quad (18)$$

where $\mathcal{K}$ is given by (15)*, and $\mathcal{B}_0$ is a standard Brownian motion, independent of $\mathcal{K}$.*

Proof Theorem 3.1 is in agreement with Theorem 2 of Halfin and Whitt (1981) except for the additional term $\mathcal{K}(\cdot)$ that arises naturally from the autoregressive assumption of the arrival-rate process. The proof of FCLT in Halfin and Whitt (1981) uses Stones criteria. Here our proof of Theorem 3.1 is more complex. Here Stone's criteria does not apply because $\widehat{X}_n$ is not a Markov process. For this reason we resort to the martingale approach as employed by Puhalskii (2013). The proof starts by showing that the properly scaled arrival, departure, and abandonment processes are all stochastically bounded. The stochastic boundedness then implies required fluid limits or FWLLN for random-time-changed stochastic processes, needed for an application of the continuous mapping theorem with the composition map. One major difference is that, unlike the proof for an $M/M/n+M$ system wherein the scaled arrival processes converge weakly to a Brownian motion, our scaled arrival processes converge to a Gaussian process having two independent terms as characterized by Lemma 3.2.

It is worth mentioning that the randomly varying arrival rates introduce additional variability that propagates as time progresses. The additional source of randomness requires proper handling in making staffing decisions. This issue will be further explored in Section 4. ■

4 | OPTIMAL STAFFING SUBJECT TO DELAY-BASED CONSTRAINT

A common approach for achieving prescribed performance target is to use the square root staffing law to estimate the amount of capacity needed assuming that call arrivals follow a Poisson process with a fixed rate parameter (see Mandelbaum & Zeltyn, 2009; also Liu & Whitt, 2017) for the case with customer feedback. As we will demonstrate below, when the arrival rates themselves are modeled as a random process, a naive application of the square-root staffing rule can fail to achieve the desired levels of service quality. Thus, it would be worthwhile to investigate techniques for selecting staffing levels in the context of stochastic arrival rates.

The proposition below, characterizing the stationary distribution of the pair $(\widehat{X}, \widehat{\lambda})$, follows from the general theory of diffusion processes.

Proposition 4.1 *The pair of limiting processes $(\widehat{X}, \widehat{\lambda})$ has a stationary distribution whose density $p(x, y; c)$ is continuous on the boundary $x = 0$ and twice differentiable elsewhere, and satisfy the following Fokker–Planck equation (FPE)*:

$$-\frac{\partial}{\partial x}[(a(x)+y)p(x,y)] + \kappa\frac{\partial}{\partial y}[yp(x,y)] + \frac{\partial^2}{\partial x^2}p(x,y) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial y^2}p(x,y) = 0, \quad (19)$$

where

$$a(x) \equiv \begin{cases} -\mu(c+x) & \text{if} \quad x \le 0, \\ -\mu c & \text{if} \quad x > 0. \end{cases}$$

Remark 4.1 (Computational considerations). Being a partial differential equation, the FPE in general does not admit an analytical solution. On the other hand, since the FPE involves just two variables, it can solved very efficiently via numerical schemes such as finite-difference and finite-element methods.

Proposition 4.1 suggests that in order to stabilize the *probability of delay* (PoD) at the target value ϱ for a many-server $M_{\text{CIR}}/M/s$ system, it suffices to adopt the SRS formula

$$s \equiv \lceil \alpha/\mu + \sqrt{\alpha}c \rceil. \quad (20)$$

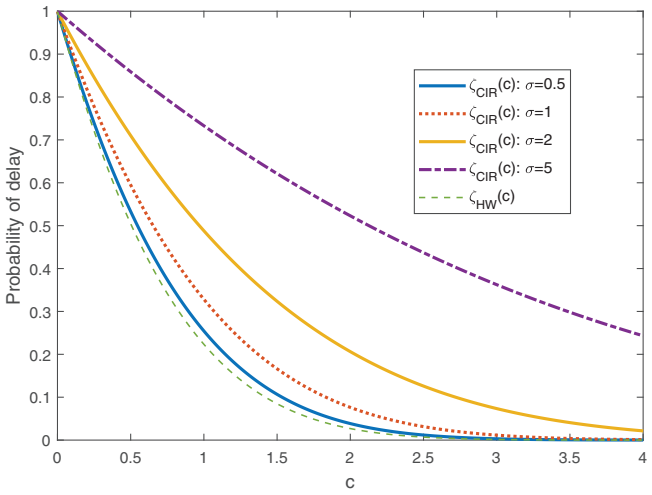

FIGURE 3 Comparing the two PoD functions $\zeta_{\mathrm{CIR}}(c)$ and $\zeta_{\mathrm{HW}}(c)$, with $\mu = 1$, $\kappa = 0.5$, $\sigma = 0.5,1,2,5$ [Colour figure can be viewed at wileyonlinelibrary.com]

with the constant c solving the equation

$$\begin{aligned}\varrho = \zeta_{\mathrm{CIR}}(c) &\equiv \int_0^\infty \int_{\mathbb{R}} p(x,y;c)\mathrm{d}y\mathrm{d}x \\ &= \mathbb{P}(\widehat{X}(\infty) > 0) \approx \mathbb{P}(X_n(\infty) \geq s_n). \end{aligned} \tag{21}$$

When $\sigma = 0$, the FPE given in (19) reduces to an ordinary differential equation (ODE)

$$-\frac{\partial}{\partial x}[a(x)p(x)] + \frac{\partial^2}{\partial x^2}p(x) = 0,$$

from which we recover the Halfin–Whitt formula for the limiting density

$$p(x) = \begin{cases} c\sqrt{\mu}\exp\{-c\sqrt{\mu}x\}\zeta_{\mathrm{HW}}(c), & \text{if } x \geq 0, \\ \frac{\phi(c\sqrt{\mu}+x)}{\Phi(c\sqrt{\mu})}(1-\zeta_{\mathrm{HW}}(c)), & \text{if } x < 0, \end{cases}$$

where $\zeta_{\mathrm{HW}}(c)$ is the well-known Halfin–Whitt function, defined as

$$\zeta_{\mathrm{HW}}(c) \equiv \left(1 + \frac{c\sqrt{\mu}\Phi(c\sqrt{\mu})}{\phi(c\sqrt{\mu})}\right)^{-1}; \tag{22}$$

(see Halfin & Whitt, 1981). Before we test the performance of the staffing formula (17) with c given by (21), we first compare the two formulas. In Figure 3, we visualize that the limiting PoD increases as σ increases by graphing $\zeta_{\mathrm{CIR}}(c)$ and $\zeta_{\mathrm{HW}}(c)$, with $\mu = 1$, $\kappa = 0.5$, $\sigma = 0.5,1,2,5$. Indeed, in order to asymptotically achieve the same PoD target, we need a bigger c for a higher system variability. Our results here provide qualitative insights on how many extra servers would be needed to account for the CIR feature in the arrival process.

We next conduct a computer simulation to test the staffing formula using Equation (21). In Figure 4, we show the delay probabilities obtained from computer simulation with targets $\varrho = 0.25,0.5,0.75$. We estimate these delay probabilities by performing 2000 independent replications with the staffing function specified by (20) and (21). We observe that for each of the three cases the delay probabilities are remarkably accurate and stable.

Finally, we quantify the consequence of neglecting the CIR feature in the arrival process when making staffing decisions. Specifically, we report estimations of the PoD for an $M_{\mathrm{CIR}}/M/s$ model with the staffing level determined as if it were a corresponding $M/M/s$ model (according to the staffing formula (17) with c determined by the inversion of the function (22)). In Table 1, we consider PoD target $\varrho = 0.3,0.5,0.7$, $\kappa = \mu = 1$, $\sigma = 0.2,2$, and α ranging from 20 to 500. For each PoD target ϱ, we use the simulated PoD $\widehat{\varrho}$ to quantify the effectiveness of the staffing levels by giving the relative error: $(\widehat{\varrho} - \varrho)/\varrho$ ("rel. err." in the table). First, we observe that the simulated PoD under the correct (CIR) staffing is effective when the system scale α is large, and it becomes inaccurate as α decreases (see column "CIR"). On the other hand, we show that simply treating an $M_{\mathrm{CIR}}/M/s$ model as an $M/M/s$ model leads to poor results; indeed, the simulated PoD are higher than the corresponding target ϱ (so the system is understaffed), especially when σ is not close to 0.

5 | MODEL WITH CUSTOMER ABANDONMENT

5.1 | The $M_{\mathrm{CIR}}/M/s + M$ model

We now consider an $M_{\mathrm{CIR}}/M/s + M$ queueing which allows waiting customers to abandon the queue, and assume the abandonment times of successive arrivals to be i.i.d. exponential random variables with rate θ. Moreover, we stipulate that service times and abandonment times are mutually independent, independent of the arrival processes.

We use $R(t)$ to represent the number of customers that have entered service and the number of abandonments from the queue, up to time t. Due to the inclusion of customer abandonment, Equations (3) and (5) are modified as

$$Q(t) = Q(0) + A(t) - E(t) - R(t) \tag{23}$$

and

$$X(t) = Q(t) + B(t) = X(0) + A(t) - D(t) - R(t).$$

In addition, letting $H(t)$ denote the head-of-line waiting time at time t, that is, the waiting time of the customer who has been waiting the longest (if there is any). Following Aras et al. (2018) and Liu and Whitt (2014) we depict $E(t)$ and $Q(t)$ as

$$E(t) = \sum_{i=1}^{A(t-H(t))} \mathbb{1}_{\{\gamma_i > V(\tau_i)\}} \quad \text{and} \tag{24}$$

$$Q(t) = \sum_{i=A(t-H(t))}^{A(t)} \mathbb{1}_{\{\tau_i+\gamma_i > t\}} \quad \text{for} \quad t \geq 0, \tag{25}$$

where the random variables $0 \leq \tau_1 \leq \tau_2 \leq \cdots$ denote arrival epochs, and $\gamma_1, \gamma_2, \ldots$ represent the abandonment times of successive customers that arrived to the system.

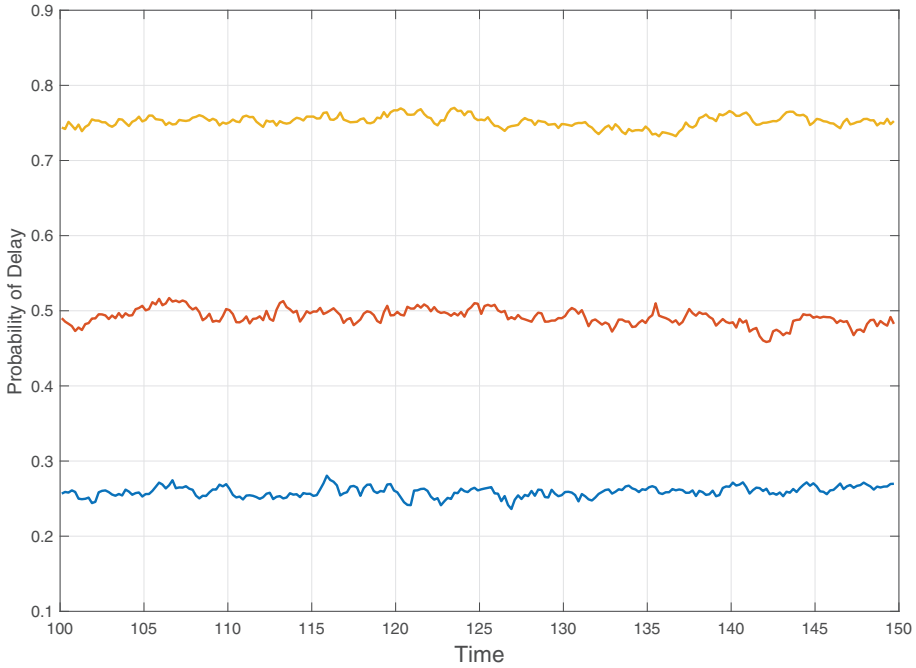

FIGURE 4 Probabilities of delay for $\rho = 0.25,0.5,0.75$, with $\alpha = 100$, $\kappa = 1$, $\sigma = 2$, exponential services with rate $\mu = 1$ [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Simulations of the PoD and relative error for the $M_{\text{CIR}}/M/s$ queue under the (a) traditional SRS staffing (column "SRS") and (b) new CIR staffing (column "CIR"), with $\alpha = 20,100,500$, $\sigma = 0.2,2$, $\rho = 0.3,0.5,0.7$

			$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
α	σ		SRS	CIR	SRS	CIR	SRS	CIR
500	0.2	Staffing	519	519	512	512	507	507
		PoD	$0.303 \pm 1.0\text{E}{-2}$	$0.303 \pm 1.0\text{E}{-2}$	$0.480 \pm 1.2\text{E}{-2}$	$0.480 \pm 1.2\text{E}{-3}$	$0.655 \pm 1.5\text{E}{-3}$	$0.655 \pm 1.5\text{E}{-3}$
		rel. err	1.14%	1.14%	−4.00%	−4.00%	−6.37%	−6.37%
	2	Staffing	519	527	512	517	507	509
		PoD	$0.456 \pm 1.2\text{E}{-2}$	$0.321 \pm 1.4\text{E}{-2}$	$0.610 \pm 1.0\text{E}{-2}$	$0.497 \pm 1.5\text{E}{-3}$	$0.739 \pm 1.8\text{E}{-2}$	$0.698 \pm 1.7\text{E}{-2}$
		rel. err	52.08%	7.13%	22.09%	−0.70%	5.61%	−0.25%
100	0.2	Staffing	109	109	106	106	103	103
		PoD	$0.279 \pm 8.0\text{E}{-3}$	$0.279 \pm 8.0\text{E}{-3}$	$0.447 \pm 2.0\text{E}{-3}$	$0.447 \pm 2.0\text{E}{-3}$	$0.690 \pm 9.0\text{E}{-3}$	$0.690 \pm 9.0\text{E}{-3}$
		rel. err	−6.97%	−6.97%	−10.68%	−10.68%	−1.46%	−1.46%
	2	Staffing	109	112	106	108	103	104
		PoD	$0.454 \pm 8.3\text{E}{-3}$	$0.338 \pm 7.1\text{E}{-3}$	$0.595 \pm 9.6\text{E}{-3}$	$0.494 \pm 2.6\text{E}{-3}$	$0.781 \pm 1.3\text{E}{-2}$	$0.709 \pm 1.3\text{E}{-2}$
		rel. errr	51.23%	12.53%	18.89%	−1.22%	11.55%	1.27%
20	0.2	Staffing	24	24	23	23	22	22
		PoD	$0.298 \pm 6.4\text{E}{-3}$	$0.298 \pm 6.4\text{E}{-3}$	$0.418 \pm 1.6\text{E}{-3}$	$0.418 \pm 1.6\text{E}{-3}$	$0.575 \pm 1.1\text{E}{-2}$	$0.575 \pm 1.1\text{E}{-2}$
		rel. errr	−0.6%	−0.6%	−16.44%	−16.44%	−17.89%	−17.89%
	2	Staffing	24	26	23	24	22	22
		PoD	$0.481 \pm 1.3\text{E}{-2}$	$0.324 \pm 9.9\text{E}{-3}$	$0.589 \pm 8.0\text{E}{-3}$	$0.479 \pm 2.1\text{E}{-3}$	$0.700 \pm 1.5\text{E}{-3}$	$0.700 \pm 1.5\text{E}{-3}$
		rel. err	60.28%	8.03%	17.70%	−4.18%	0.1%	0.1%

5.2 | Many-server heavy-traffic analysis

Following the asymptotic framework and staffing formula in Section 3.1, we will next give the MSHT FCLT for the $M_{\text{CIR}}/M/s+M$ queue.

Theorem 5.1 (FCLT for the critically loaded $M_{CIR}/M/s+M$ model). *Suppose customers arrive according to the DSPP $A_n(\cdot)$ with intensity process $\lambda_n(\cdot)$ given by* (8)*, the system is staffed according to* (17)*. If in addition*

$$\widehat{X}_n(0) \Rightarrow \widehat{X}(0)$$

jointly with (9)*, then*

$$(\widehat{X}_n, \widehat{Q}_n, \widehat{V}_n) \Rightarrow (\widehat{X}, \widehat{Q}, \widehat{V}) \quad in \quad \mathcal{D}^3 \quad as \quad n \to \infty$$

jointly with (10)*, in which* $\widehat{V}(t) = \widehat{Q}(t) = [\widehat{X}(t)]^+$*, and the limiting process* $\widehat{X}$ *satisfies*

$$\widehat{X}(t) = \widehat{X}(0) - \mu ct - \int_0^t m(\widehat{X}(u))\mathrm{d}u + \mathcal{K}(t) + \sqrt{2}\mathcal{B}_1(t),$$

where $m(x) \equiv -\mu x^- + \theta x^+$ *and* $\mathcal{B}_1$ *is a standard Brownian motion, independent of* $\mathcal{K}$.

Theorem 5.1 is in agreement with Theorem 2 of Garnett et al. (2002) except for the additional term $\mathcal{K}(\cdot)$ due to the autoregressive assumption of the arrival-rate process. The randomly varying arrival rates not only introduce additional variability but also leads to non-Markovian heavy-traffic limits as demonstrated in the preceding theorem.

Paralleling Proposition 4.1, we can formally characterize the stationary distribution of $(\widehat{X}, \widehat{\lambda})$, for $\widehat{X}$ given in (10), through the solution to a partial differential equation.

Proposition 5.1 *The pair of limiting processes* $(\widehat{X}, \widehat{\lambda})$ *given in Theorem 5.1 has a stationary distribution whose density* $p(x, y; c)$ *is continuous on the boundary* $x = 0$ *and twice differentiable elsewhere, and satisfy the following FPE*:

$$-\frac{\partial}{\partial x}[(\widetilde{a}(x) + y)p(x, y)] + \kappa\frac{\partial}{\partial y}[yp(x, y)] + \frac{\partial^2}{\partial x^2}p(x, y) + \frac{\sigma^2}{2}\frac{\partial^2}{\partial y^2}p(x, y) = 0, \tag{26}$$

where

$$\widetilde{a}(x) \equiv \begin{cases} -\mu c - \mu x & \text{if} \quad x \le 0, \\ -\mu c - \theta x & \text{if} \quad x > 0. \end{cases}$$

5.3 | Optimal staffing

Similar to the no-abandonment case, the staffing problem requires solving the FPE as specified by (26). Unfortunately, the FPE does not admit no closed-form solution except for very special cases. To gain greater simplicity and tractability, we first focus attention on a special case where service rate and the abandonment rate are equal, that is, $\theta = \mu$. We next provide a heuristic formula for the more general case $\theta \neq \mu$.

5.3.1 | The special case: $\theta = \mu$

If $\theta = \mu$, we would get $m(x) \equiv \mu x^- + \theta x^+ = \mu x$, and the resulting FPE can be solved explicitly, yielding

$$p(x, y; c) = (2\pi)(\det D)^{-1/2} \times \exp\left(-\frac{1}{2}(x - c, y)D^{-1}(x - c, y)^\top\right), \tag{27}$$

where D is a square matrix, given as

$$D \equiv \begin{pmatrix} \frac{1}{\mu} + \frac{\sigma^2}{2\mu\kappa(\mu+k)} & \frac{\sigma^2}{2\kappa(\mu+k)} \\ \frac{\sigma^2}{2\kappa(\mu+k)} & \frac{\sigma^2}{2\kappa} \end{pmatrix}.$$

Upon integrating out the variable y in (27), we arrive at the following result.

Proposition 5.2 *Suppose* $\mu = \theta$*. Then, as* $t \to \infty$*, the sequence of random variables* $\widehat{X}(t)$ *in* (29) *converges weakly to* $\widehat{X}(\infty)$ *which is a Gaussian random variable with mean* $-c$ *and variance* k *where* k *is given by*

$$k \equiv k(\mu, \kappa, \sigma) \equiv \sqrt{\frac{1}{\mu} + \frac{\sigma^2}{2\mu\kappa(\mu + \kappa)}}. \tag{28}$$

To explain why Equation (26) leads to an explicit solution for $\mu = \theta$, we see that the limiting process $\widehat{X}$ can be solved explicitly and expressed as

$$\widehat{X}(t) = e^{-\mu t}\left(\widehat{X}(0) - c\mu\int_0^t e^{\mu u}\mathrm{d}u + \int_0^t e^{\mu u}\widehat{\lambda}(u)\mathrm{d}u + \sqrt{2}\int_0^t e^{\mu u}\mathrm{d}\mathcal{B}_2(u)\right). \tag{29}$$

From the above expression, it is easy to see that the mean value $\mathbb{E}[\widehat{X}(t)]$ converges to $-c$ as t goes to infinity. For the variance, we have

$$\mathrm{Var}(\widehat{X}(t)) = \vartheta_1(t) + \vartheta_2(t),$$

where

$$\begin{aligned}\vartheta_1(t) &\equiv \mathrm{Var}\left(\int_0^t e^{-\mu(t-u)}\widehat{\lambda}(u)\mathrm{d}u\right) \\ &= \frac{\sigma^2}{(\mu-\kappa)^2}\int_0^t (e^{-2\kappa(t-s)} - 2e^{-(\kappa+\mu)(t-s)} + e^{-2\mu(t-s)})\mathrm{d}s \\ &= \frac{\sigma^2}{(\mu-\kappa)^2}\left(\frac{1}{2\kappa}(1 - e^{-2\kappa t}) - \frac{2}{\kappa+\mu}(1 - e^{-(\kappa+\mu)t}) + \frac{1}{2\mu}(1 - e^{-2\mu t})\right),\end{aligned}$$

and

$$\begin{aligned}\vartheta_2(t) &\equiv 2\mathrm{Var}\left(\int_0^t e^{-\mu(t-u)}\mathrm{d}\mathcal{B}_2(u)\right) \\ &= 2\int_0^t e^{-2\mu(t-u)}\mathrm{d}u = \frac{1}{\mu}(1 - e^{-2\mu t}).\end{aligned}$$

Combining the above expressions and then sending $t \to \infty$ leads us to the same result as stated in Proposition 5.2.

Continuing our discussion on optimal staffing, we recall that the number of servers was selected according to the square-root staffing formula, as in (20). Then Proposition 5.2 states that when the mean-reversion level α is large, we can heuristically approximate the steady-state distribution of the number of customers in the $M_{\mathrm{CIR}}/M/s + M$ model *having equal service and abandonment rates* as follows:

$$X(\infty) = \alpha/\mu + \sqrt{\alpha}k\mathcal{N}, \tag{30}$$

where $\mathcal{N}$ denotes a standard normal random variable. Combining the square-root staffing formula and (30) yields a simple *normal approximation*:

$$\mathbb{P}(V(\infty) > 0) = \mathbb{P}(X(\infty) > s) \approx \mathbb{P}(\mathcal{N} > c/k).$$

From this normal approximation we immediately understand that, in order to stabilize the delay probability at the target value ρ, one ought to choose

$$c = z_{1-\rho} \cdot k \equiv z_{1-\rho} \cdot \sqrt{\frac{1}{\mu} + \frac{\sigma^2}{2\mu\kappa(\mu+\kappa)}} \tag{31}$$

in the square-root staffing formula (20).

Relative to the no-abandonment case as discussed in Section 4, the expression in (31) is fairly explicit and therefore generate clear-cut operational insights; it demonstrates that ignoring a randomly varying arrival rate in making staffing decisions can result in severe under-staffing. In other words, in the presence of stochastic arrival rates, the use of square-root staffing rule would lead to a higher safety staffing level compared to the case where the arrival rate is deterministic. This happens because of a mismatch between the realized arrival rate and the number of servers available to handle those demands. As a result, the service provider needs to hire additional staff (corresponding to the second term under the square root) to ensure that the system can handle a larger-than-foreseen demand volume without jeopardizing the quality of service.

5.3.2 | The general case: $\theta \neq \mu$

It is widely recognized that a normal approximation may not be appropriate when $\theta \neq \mu$; see, for example, the discussion in Section 6 of Feldman et al. (2008). As previously alluded to, the more general scenario ($\theta \neq \mu$) requires numerically solving a two-dimensional partial differential equation. As a rule of thumb, we propose the following staffing formula

$$s \equiv \lceil \alpha/\mu + \sqrt{\alpha}c \rceil$$

with the constant c solving the equation

$$\varrho = \widetilde{\zeta}(c) \equiv \left[1 + \frac{h((c/k)\sqrt{\mu/\theta})}{\sqrt{\mu/\theta}h(-c/k)}\right]^{-1}, \tag{32}$$

where k is as in (28). To see that this heuristic approximation can deliver a good performance, consider the special scenario with $\sigma = 0$, in which case the expression in (33) reduces to

$$\left[1 + \frac{h((c\sqrt{\mu})\sqrt{\mu/\theta})}{\sqrt{\mu/\theta}h(-c\sqrt{\mu})}\right]^{-1}. \tag{33}$$

The careful reader would notice that this is exactly the delay-probability formula for $M/M/n+M$ systems as established in Garnett et al. (2002). The above formula can be readily extended to $GI/M/n+M$ systems by replacing c in (33) with $c' \equiv c/\sqrt{(1+c_a^2)/2}$ for c_a being the coefficient of variation of the time between arrivals. This is because a greater value of c_a would increase system variability by a factor of $\sqrt{(1+c_a^2)/2}$.

On the other hand, the expression in (28) suggests that the additional variability in the CIR intensity tends to inflate system variability by a factor of $\sqrt{1+\sigma^2/(\kappa(\kappa+\mu))}$. Thus, thinking of the $M_{\mathrm{CIR}}/M/s+M$ model as a $GI/M/n+M$ system, we should replace c in (33) with $c'' \equiv c/\sqrt{1+\sigma^2/(\kappa(\kappa+\mu))}$ to achieve effective staffing, which leads up to the formula given in (32). At this point we conjecture that the right-hand side of (32) is equal to the integral $\int_0^\infty \int_{\mathbb{R}} p(x,y)\mathrm{d}y\mathrm{d}x$ with $p(x,y)$ solving the FPE (26). Finally, as in Garnett et al. (2002), it is easy to verify that by passing to the limit $\theta \to 0$, the function $\widetilde{\zeta}(c)$ reduces to a modification of (22) given as

$$\zeta'_{\mathrm{CIR}}(c) \equiv \left(1 + \frac{c/k\Phi(c/k)}{\phi(c/k)}\right)^{-1},$$

which serves as an approximation for $\zeta_{CIR}(c)$ given in (21).

To demonstrate that the approximating formula (32) can achieve the desired time-stable performance. In Figure 5, we plot the delay probabilities obtained from computer simulation with targets $\varrho = 0.1,0.3,0.5,0.7,0.9$. These delay probabilities are estimated by performing 2,000 independent replications with the staffing function specified by (20) and (32). We observe that delay probabilities fluctuate around the target in each case., i.e., the probabilities of delay are stabilized remarkably well.

To test the robustness of our staffing algorithm against system scales, we also perform simulation experiments on two (much) smaller systems, using the same set of delay probability targets. Figure 6 illustrates the delay probabilities estimated from 2,000 simulation runs with $\alpha = 10, 20$. We see that our proposed staffing algorithm remains effective when $\alpha = 20$ and becomes less efficient when $\alpha = 10$ (our formula seems to be understaffing the $M_{\mathrm{CIR}}/M/s+M$ model when $\alpha = 10$).

6 | PROOFS

In this section, we provide the proofs for Lemma 3.1, Lemma 3.2, Proposition 4.1 and Theorem 5.1. Since Theorem 3.1 is a special case of Theorem 5.1, its proof is omitted.

Proof of Lemma 3.1 Dividing both sides of (8) by n and in view of (7), we obtain

$$\bar{\lambda}_n(t) = \bar{\lambda}_n(0) + \kappa \int_0^t (1 - \bar{\lambda}_n(u))\mathrm{d}u + \sigma_n \int_0^t \sqrt{\bar{\lambda}_n(u)}d\mathcal{B}(u), \tag{34}$$

where we have defined $\sigma_n \equiv \sigma/n$. We prove the FWLLN by arguing that the volatility term vanishes as $n \to \infty$. Because $\sigma_n \to 0$ as $n \to \infty$, it suffices to argue that the sequence $\{\bar{\lambda}_n; n \in \mathbb{N}\}$ is stochastically bounded. For this purpose we appeal to Lemma 3.9 of Whitt (2007). In particular, if the sequence has continuous sample paths, then the proof of stochastic boundedness amounts to verifying the modulus of continuity condition (MCC). (We refer the reader to Theorem 3.2 in Whitt (2007) or Theorem 16.8 in Billingsley (Billingsley, 2013) for a formal

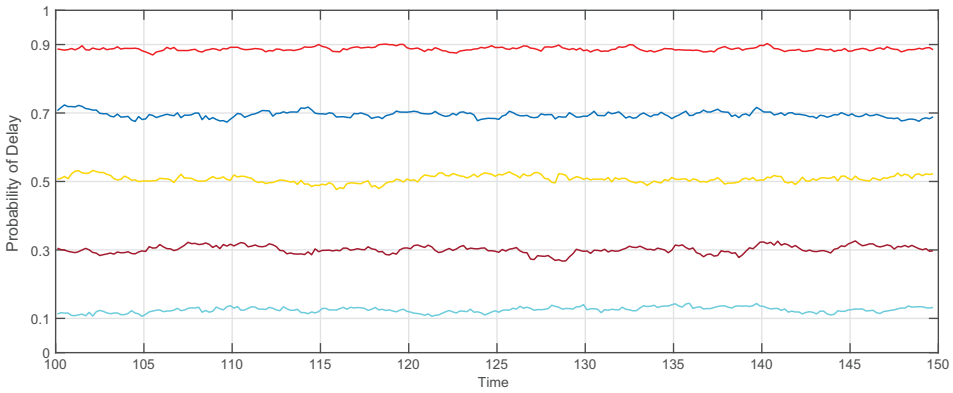

FIGURE 5 Probabilities of delay for five delay-probability targets $\rho = 0.1,0.3,0.5,0.7,0.9$, with $\alpha = 100$, $\kappa = 1$, $\sigma = 2$, exponential services and abandonments with rates $\mu = 1$ and $\theta = 0.5$, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

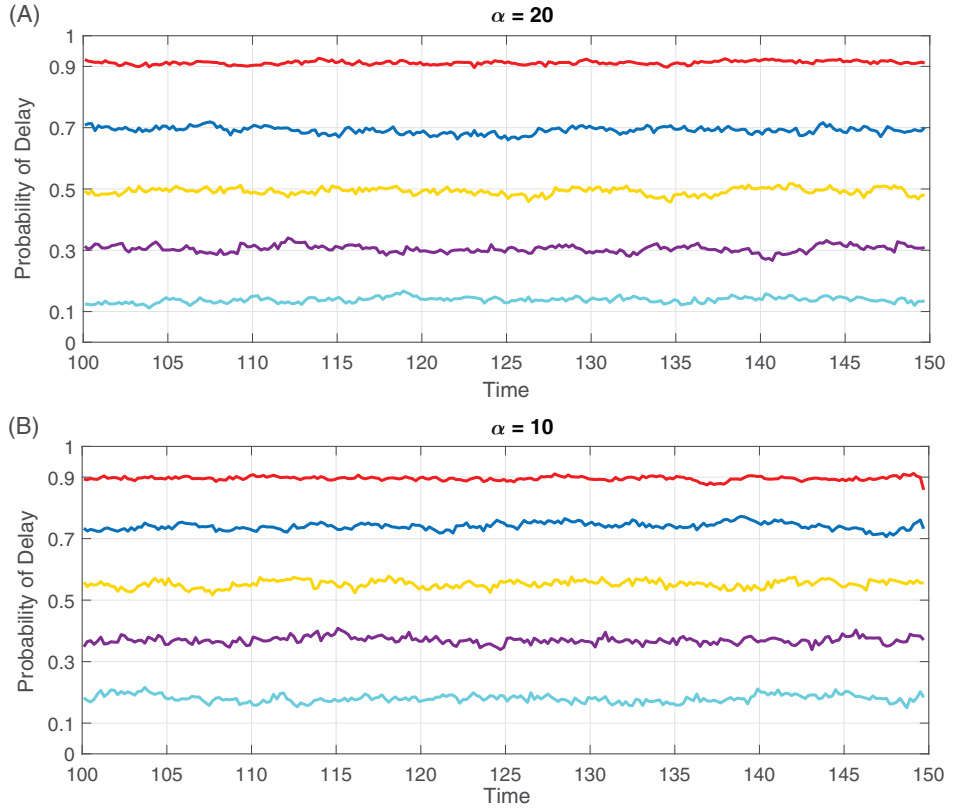

FIGURE 6 Probabilities of delay for five delay-probability targets $\rho = 0.1,0.3,0.5,0.7,0.9$, with $\alpha = 20$, 10, $\kappa = 1$, $\sigma = 2$, exponential services and abandonments with rates $\mu = 1$ and $\theta = 0.5$, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

definition of MCC.) Towards that end, we verify the (sufficient) moment condition laid out in Lemma 3.11 (ii.b) of Whitt (2007). Note that

$$
\begin{aligned}
&\mathbb{E}[(\overline{\lambda}_n(t+u) - \overline{\lambda}_n(t))^2 | \mathcal{F}_t] \\
&\le 2\mathbb{E}\left[\left(\int_t^{t+u} \kappa(1 - \overline{\lambda}_n(s))\mathrm{d}s\right)^2\right] \\
&\quad + 2\sigma_n^2 \mathbb{E}\left[\left(\int_t^{t+u} \sqrt{\overline{\lambda}_n(s)}d\mathcal{B}(s)\right)^2\right] \\
&\le 2u\mathbb{E}\left[\int_t^{t+u} \kappa^2(1 - \overline{\lambda}_n(s))^2\mathrm{d}s\right] \\
&\quad + 2\sigma_n^2 C_p \mathbb{E}\left[\int_t^{t+u} \overline{\lambda}_n(s)\mathrm{d}s\right] \\
&\le 2u\int_0^T \kappa^2((1 + 2\mathbb{E}[\overline{\lambda}_n(s)] + \mathbb{E}[\overline{\lambda}_n^2(s)])\mathrm{d}s \\
&\quad + \sigma_n^2 C_p u^{1/4}\left(1 + \int_0^T \mathbb{E}[\overline{\lambda}_n^2(s)]\mathrm{d}s\right),
\end{aligned}
$$

where the first inequality follows from (34), the second inequality follows by applying the Cauchy–Schwartz inequality to the first term and the Burkholder–Davis–Gundy inequality to the second term, and the third inequality follows by another application of the Cauchy–Schwartz inquality. By Lemma 6.1 below we conclude that both integrals on the right-hand side approach zero as $u \to 0$, uniformly overall t and

n. This shows that the MCC stated in Lemma 3.11 (ii.b) of Whitt (2007) is indeed satisfied. Then by Lemma 3.9 in Whitt (2007), we obtain the stochastic boundedness of $\{\overline{\lambda}_n; n \in \mathbb{N}\}$ from which the desired WFLLN follows, namely,

$$\overline{\lambda}_n \Rightarrow \mathfrak{e} \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty. \tag{35}$$

Next subtract α_n from both sides of (34) and scale up both sides of the resulting equation by $\sqrt{n}$ to get

$$\widehat{\lambda}_n(t) = \widehat{\lambda}_n(0) - \kappa \int_0^t \widehat{\lambda}_n(u)\mathrm{d}u + \sigma \int_0^t \sqrt{\overline{\lambda}_n(u)} d\mathcal{B}(u). \tag{36}$$

Note that the mapping $g : \mathbb{R} \times \mathcal{D} \to \mathcal{D}$ taking (b, x) into y determined by the integral representation

$$y(t) = b - \kappa \int_0^t y(u)\mathrm{d}u + \sigma \int_0^t \sqrt{x(u)} d\mathcal{B}(u) \quad \text{for} \quad t \geq 0$$

is continuous. We can therefore invoke the continuous mapping theorem with the established weak convergence in (35) to obtain the desired FCLT for the sequence $\{\widehat{\lambda}_n; n \in \mathbb{N}\}$. ∎

Lemma 6.1 *If $\overline{\lambda}_n \equiv \lambda_n/n$, then for any finite $T > 0$ and $k \in \mathbb{N}$, we have*

$$\sup_n \sup_{0 \leq t \leq T} \mathbb{E}[\overline{\lambda}_n^k(t)] < \infty.$$

Proof of Lemma 6.1 Recall that

$$\mathrm{d}\overline{\lambda}_n(t) = \kappa(1 - \overline{\lambda}_n(t))\mathrm{d}t + \sigma_n \sqrt{\overline{\lambda}_n(t)} d\mathcal{B}(t),$$

where $\sigma_n \equiv \sigma/n$. Applying Itô's formula to the smooth function $f(x) = x^k$, we obtain

$$\begin{aligned}\overline{\lambda}_n^k(t) = \overline{\lambda}_n^k(0) + k\kappa \int_0^t \overline{\lambda}_n^{k-1}(u)\mathrm{d}u - k\kappa \int_0^t \overline{\lambda}_n^k(u)\mathrm{d}u \\ + k\sigma_n \int_0^t \overline{\lambda}_n^{k-1/2}(u) d\mathcal{B}(u).\end{aligned} \tag{37}$$

An application of Young's inequality yields

$$\overline{\lambda}_n^{k-1}(u) \leq (k-1)\overline{\lambda}_n^k(u)/k + 1/k.$$

Substituting the above into (37) gives

$$\begin{aligned}\overline{\lambda}_n^k(t) \leq \overline{\lambda}_n^k(0) + \kappa \int_0^t [(k-1) - k]\overline{\lambda}_n^k(u)\mathrm{d}u \\ + \kappa t + k\sigma_n \int_0^t \overline{\lambda}_n^{k-1/2}(u) d\mathcal{B}(u).\end{aligned}$$

Taking expectation on both sides, we get

$$\mathbb{E}[\overline{\lambda}_n^k(t)] \leq \mathbb{E}[\overline{\lambda}_n^k(0)] + \kappa \int_0^t [(k-1) - k]\mathbb{E}[\overline{\lambda}_n^k(u)]\mathrm{d}u + \kappa t.$$

An application of the Gronwall's inequality allows us to conclude

$$\mathbb{E}[\overline{\lambda}_n^k(t)] \leq C(k, T)e^{C(k,T)T}.$$

The result immediately follows due to fact that the bound on the right hand side is independent of both t and n. ∎

Proof of Lemma 3.2 First use (13) to write

$$\widehat{A}_n(t) = \widehat{A}_{n,1}(t) + \widehat{A}_{n,2}(t)$$

where we defined

$$\widehat{A}_{n,1}(t) \equiv n^{-1/2}\left(A_n(t) - \int_0^t \lambda_n(u)\mathrm{d}u\right) \quad \text{and}$$

$$\widehat{A}_{n,2}(t) \equiv \int_0^t \widehat{\lambda}_n(u)\mathrm{d}u. \tag{38}$$

The first term is a square integrable martingale with quadratic variation $\int_0^t \overline{\lambda}_n(u)\mathrm{d}u$ converging to t as $n \to \infty$. Appealing to the martingale FCLT, we obtain

$$\widehat{A}_{n,1}(t) \Rightarrow \mathcal{B}_\lambda(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.$$

For the second term, we obtain the convergence $\widehat{A}_{n,2}(t) \Rightarrow \mathcal{K}(t)$ by applying the continuous mapping theorem with (10). To establish the desired result, one would need to strengthen the above individual convergence to joint convergence. The joint convergence of multiple random elements is equivalent to individual convergence if they are independent. Here $\widehat{A}_{n,1}$ and $\widehat{A}_{n,2}$ are not independent because both involves the arrival-rate process. But they are conditionally independent given the arrival intensity. Hence one can establish the required joint convergence by first conditioning and then unconditioning. In this way, the desired result follows from yet another application of the continuous mapping theorem. ∎

Proof of Theorem 5.1 As the proof is relatively standard, we outline the key components only. To start, subtract (5) by s_n and divide both sides by $\sqrt{n}$ to get the stochastic integral equation satisfied by the diffusion-scaled process $\widehat{X}_n$

$$\begin{aligned}\widehat{X}_n(t) = \widehat{X}_n(0) - \mu ct - \int_0^t m(\widehat{X}_n(u))\mathrm{d}u \\ + \widehat{A}_n(t) - \widehat{D}_n(t) - \widehat{R}_n(t),\end{aligned} \tag{39}$$

where $m(x) \equiv -\mu x^- + \theta x^+$,

$$\widehat{D}_n(t) \equiv n^{-1/2}\left(D_n(t) - \mu \int_0^t B_n(u)\mathrm{d}u\right),$$

$$\widehat{R}_n(t) \equiv n^{-1/2}\left(R_n(t) - \theta \int_0^t Q_n(u)\mathrm{d}u\right),$$

and $\widehat{A}_n(t)$ is defined by (13). It follows easily that both $\widehat{D}_n$ and $\widehat{R}_n$ are square-integrable martingales with respect to proper filtration. In

particular, both $\{\widehat{D}_n; n \in \mathbb{N}\}$ and $\{\widehat{R}_n; n \in \mathbb{N}\}$ are processes that are stochastically bounded. Upon applying the Gronwall's inequality to (39), we conclude that the sequence of processes $\{\widehat{X}_n; n \in \mathbb{N}\}$ is stochastically bounded; as an immediate consequence, the quadratic variations of $\widehat{D}_n$ and $\widehat{R}_n$ converge to the corresponding limits, which in turn implies

$$\widehat{D}_n(t) \Rightarrow \mathcal{B}_\mu(t) \quad \text{in} \quad \mathcal{D} \quad \text{and} \quad \widehat{R}_n(t) \Rightarrow 0 \quad \text{in} \quad \mathcal{D}, \tag{40}$$

where $\mathcal{B}_\mu$ is a standard Brownian motion and 0 denotes the zero function. Appealing to Theorem 4.1 in Pang et al. (2007) with (14) and (40) yields the desired FCLT result for $\{\widehat{X}_n; n \in \mathbb{N}\}$. The assertion $\widehat{Q}_n \Rightarrow \widehat{Q}$ is fairly straightforward and follows from the simple relation between $\widehat{X}_n$ and $\widehat{Q}_n$. Finally, the FCLT for $\{\widehat{V}_n; n \in \mathbb{N}\}$ follows from the well-known snap-shot principle. ∎

7 | CONCLUSIONS

Motivated by recent empirical findings on the autocorrelative nature of the arrival data in service systems, we study the $M_{\text{CIR}}/M/s$ model, a queueing model where customers arrive according to a doubly stochastic Poisson point process with a random arrival rate driven by a CIR process. We first establish functional limit theorems for the CIR-driven arrival process, based on which we develop functional limit theorems for the queue length process of the $M_{\text{CIR}}/M/s$ model. These theoretical developments serve as a cornerstone of an optimal staffing problem for the $M_{\text{CIR}}/M/s$ queue subject to delay-based constraints on the service levels. Our analysis acknowledges the presence of autoregressive structure in arrivals and lead to novel staffing rules. In addition, we extend our results to the $M_{\text{CIR}}/M/s + M$ queue having customer abandonment.

There are several venues for future research. One nature extension is to consider queueing models having M_{CIR} arrivals subject to service-level constraints based on the more practical *tail probability of delay* (TPoD), which is defined as the probability that the customer delay exceeds a customary delay target, that is,

$$\mathbb{P}(V(t) > w) \leq \rho, \tag{41}$$

where w is a pre-specified delay target. TPoD is widely used as a performance metric in many real-world service systems. Examples include the 80–20 rule in customer contact centers and the Canadian triage and acuity scale (CTAS) guideline that classifies patients in the emergency department into five acuity levels (Liu, 2018; Liu et al., 2020). Another future direction is to extend to models having multiple customer classes where the manager has to concurrently determine the required staffing level and scheduling policy (assigning newly idle servers to a waiting customer from one of the classes). It is especially interesting to investigate necessary changes to conventional scheduling rules due to the M_{CIR} arrivals (as opposed to models with Poisson arrivals, e.g., Liu et al., 2020).

ACKNOWLEDGMENT

We thank the editors and two anonymous referees for their constructive comments.

ORCID

Xu Sun https://orcid.org/0000-0003-2560-7370
Yunan Liu https://orcid.org/0000-0001-9961-2610

REFERENCES

Aras, A. K., Chen, X., & Liu, Y. (2018). Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment. *Queueing Systems*, *89*(1–2), 81–125.

Bassamboo, A., Randhawa, R. S., & Zeevi, A. (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, *56*(10), 1668–1686.

Billingsley, P. (2013). Convergence of probability measures. New York, NY: John Wiley & Sons.

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

Brown, L. D., Zhang, R., &Zhao, L. H. (2001). *Root-unroot methods for nonparametric density estimation and Poisson random-effects models* (Tech. rep.), Department of Statistics, University of Pennsylvania.

Daw, A., & Pender, J. (2018). Queues driven by Hawkes processes. *Stochastic Systems*, *8*(3), 192–229.

Dias, J. C., & Shackleton, M. B. (2011). Hysteresis effects under cir interest rates. *European Journal of Operational Research*, *211*(3), 594–600.

Ding, S., Koole, G., & van der Mei, R. D. (2015). On the estimation of the true demand in call centers with redials and reconnects. *European Journal of Operational Research*, *246*(1), 250–262.

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

Gao, X., & Zhu, L. (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems*, *90*(1), 1–46.

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

Glasserman, P. (2013). Monte Carlo methods in financial engineering. Springer Science & Business Media. Berlin, Germany, Springer.

Halfin, S., & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, *29*(3), 567–588.

Harrison, J. M., & Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management*, *7*(1), 20–36.

He, B., Liu, Y., & Whitt, W. (2016). Staffing a service system with non-Poisson non-stationary arrivals. *Probability in the Engineering and Informational Sciences*, *30*(4), 593–621.

Ibrahim, R., & L'Ecuyer, P. (2013). Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management*, *15*(1), 72–85.

Ibrahim, R., Ye, H., LEcuyer, P., & Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, *32*(3), 865–874.

Jongbloed, G., & Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, *17*(4), 307–318.

Kim, S.-H., & Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management*, *16*(3), 464–480.

Koçağa, Y. L., Armony, M., & Ward, A. R. (2015). Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management*, *24*(7), 1101–1117.

Koops, D., Boxma, O., & Mandjes, M. (2017). Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems*, *86*(3–4), 301–325.

Lange, K. (2010). Numerical analysis for statisticians. Berlin, Germany: Springer Science & Business Media.

Liu, Y. (2018). Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research*, *66*(2), 514–534.

Liu, Y., Sun, X., & Hovey, K. (2020). Scheduling to differentiate service in a multiclass service system. *Operations Research* (in press).

Liu, Y., & Whitt, W. (2014). Many-server heavy-traffic limit for queues with time-varying parameters. *The Annals of Applied Probability*, *24*(1), 378–421.

Liu, Y., & Whitt, W. (2017). Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research*, *256*(2), 473–486.

Mandelbaum, A., & Zeltyn, S. (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, *57*(5), 1189–1205.

Mathijsen, B. W., Janssen, A., van Leeuwaarden, J. S., & Zwart, B. (2018). Robust heavy-traffic approximations for service systems facing overdispersed demand. *Queueing Systems*, *90*(3–4), 257–289.

Moreno, M., & Platania, F. (2015). A cyclical square-root model for the term structure of interest rates. *European Journal of Operational Research*, *241*(1), 109–121.

Overbeck, L., & Rydén, T. (1997). Estimation in the Cox–Ingersoll–Ross model. *Econometric Theory*, *13*(3), 430–461.

Pang, G., Talreja, R., & Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, *4*, 193–267.

Puhalskii, A. A. (2013). On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Mathematical Methods of Operations Research*, *78*(1), 119–148.

Ross, S. M. (1996). Stochastic processes. Cambridge, MA: Academic Press.

Shen, H., & Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, *10*(3), 391–410.

Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, *15*(1), 88–102.

Whitt, W. (2007). Proofs of the martingale FCLT. *Probability Surveys*, *4*, 268–302.

Ye, H., Luedtke, J., & Shen, H. (2019). Call center arrivals: When to jointly forecast multiple streams? *Production and Operations Management*, *28*(1), 27–42.

Zhang, X., Blanchet, J., Giesecke, K., & Glynn, P. W. (2015). Affine point processes: Approximation and efficient simulation. *Mathematics of Operations Research*, *40*(4), 797–819.

Zhang, X., Hong, L. J., & Glynn, P. W. (2013). *Timescales in modeling call center arrivals* (Working paper).

Zhang, X., Hong, L. J., & Zhang, J. (2014). *Scaling and modeling of call center arrivals*. In Simulation Conference, 2014 Winter (pp. 476–485). New York, NY: IEEE.

How to cite this article: Sun X, Liu Y. Staffing many-server queues with autoregressive inputs. *Naval Research Logistics* 2020;1–15. https://doi.org/10.1002/nav.21960