Many-Server Queues with Autoregressive Inputs

Xu Sun

*Department of Industrial and Systems Engineering, University of Florida, Gainesville, USA

Abstract

Recent studies reveal significant overdispersion and autocorrelation in arrival data at large call centers. Motivated by these findings, we study a queueing model where customers arrive according to 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. We first prove a functional weak law of large numbers and a functional central limit theorem for the CIR process which we believe can be of independent interest. We then establish functional limit theorems for our queueing model under suitable heavy-traffic regimes. The results acknowledge the presence of autoregressive structure in arrivals and lead to novel staffing rules.

Keywords: queues, autocorrelation, mean-reverting process, heavy-traffic approximations, parameter uncertainty

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); 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 and Koole (2001); Zhang et al. (2014). This phenomenon violates the fundamental property that underpins the Poisson input assumption and can potentially affect performance evaluation and choice of staffing rules. It

Email address: xusun@ufl.edu (Xu Sun)
is found that the overdispersion can be partially explained by the strong autocorrelation of arrival counts over successive periods during the day (Ibrahim and L’Ecuyer (2013); Shen and Huang (2008)). These non-Poisson features 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 arrivals as a doubly stochastic Poisson process (DSPP) of which the intensities are driven by a possibly time-dependent Cox-Ingersoll-Ross (CIR) process. The CIR process is typically used to model instantaneous interest rates; see, e.g., Dias and Shackleton (2011); Moreno and 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 and 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.

We are by no means the first to use the CIR process to model the intensities of a counting process. Zhang et al. (2014) propose to use a stationary CIR process to model the intensity of call arrivals and show that the proposed traffic model can faithfully reproduce the behavior of interest observed in practice. They derive the scaling limit for the arrival process concerning “convergence in marginal distribution”. Here, we consider a more general CIR model by allowing the mean-reversion level to be time-dependent and 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). We then rely on the resulting approximation to study many-server queues under suitable heavy-traffic regimes.

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), Whitt (2006) uses fluid-model analysis to derive staffing solutions for a call center with uncertain arrival rate and employee absenteeism. Koçağa et al. (2015) develop an economic model to aid staffing decisions in the presence of random arrival rates with a co-sourcing option. We too consider staffing problems in the context of call centers, but we do not consider economic models; instead, we work directly
with performance measures associated with customer delays and abandonments.

The present study is perhaps most closely related to Koops et al. (2017), Gao and Zhu (2018), and Daw and Pender (2018). Koops et al. analyze an infinite-server queue fed by a Cox process of which the intensities are driven by a short-rate process. They illustrate via both exact and asymptotic analysis that their traffic model can indeed capture the overdispersed demand observed in practice. Gao and Zhu model customer arrivals as a stationary Hawkes process and derive heavy-traffic approximations for an infinite-server queue fed by a Hawkes process. Their model is especially attractive when demand exhibits clustering and self-exciting features. Daw and Pender study an infinite-server queueing model in which customer arrivals follow a self-exciting Hawkes process and service times are phase-type or deterministic. There is one key difference between the modeling approaches of the aforementioned papers and the one adopted in the present paper. The intensity process in these papers has jumps that decay exponentially over time whereas 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, unlike the aforementioned papers that study infinite-server models, we deal 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 call center managers make informed staffing decisions in the presence of stochastic demand fluctuations.

The contribution and organization of this paper are as follows. We formally introduce the DSPP with CIR-driven intensities and summarize some of its key properties in §2.2. In §2.3 we describe the corresponding $M_{CIR}/M/s_t + G$ model and specify the associated staffing problem that gives rise to a critically loaded system. Our first main result is Lemma 2, stated in §3.1, whose proof
relies on Lemma 1. In Lemma 1, we establish what we believe the first functional weak law of large numbers (FWLLN) and functional central limit theorem (FCTL) for CIR processes. Because CIR processes are widely used in many applications domains, these functional limit theorems can be of independent interest. In §3.2 we present the FCLT result for our $M_{CIR}/M/s_t + G$ 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 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. In §4 we propose novel staffing rules in the presence of autoregressive inputs based on the FCLT limits and present extensive numerical studies. In §5 we establish functional limit theorems for the $M_{CIR}/M/s_t + G$ system under a different scaling motived by a more general performance metric. The main result, Theorem 2, is essentially an FCLT for the overloaded $M_{CIR}/M/s_t + G$ system. The result extends the work of Liu et al. (2014) to allow for autoregressive inputs. However, the techniques employed in the proof are innovative. Specifically, unlike Liu et al. (2014) in which the proof of the FCLT relies on a separate FWLLN result, we establish the FCLT directly by applying a novel change of variables in a stochastic setting. The FWLLN then follows as an immediate consequence.

2. Preliminaries

2.1. Notation and conventions

We denote by $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{N}$, respectively, the sets of all real numbers, non-negative reals and nonnegative integers. We use $\lfloor a \rfloor$ to denote the least integer that is greater than or equal to $a$ and $z_\alpha$ denote the quantile value from a standard normal distribution at $\alpha$. For a real-value function $f$, we write $f[x_1, x_2]$ as shorthand for $f(x_2) - f(x_1)$. We use $\mathbf{1}$ 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 $(\Omega, \mathcal{F}, \mathbb{P})$.

2.2. Doubly stochastic Poisson process with CIR intensities

We model the arrival process $A(t)$ as a DSPP with CIR-driven intensities, i.e.,

$$A(t) \equiv \Pi_\alpha \left( \int_0^t \lambda(u)du \right), \quad t \geq 0,$$  \hspace{1cm} (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)

$$d\lambda(t) = \kappa(\alpha(t) - \lambda(t))dt + \sigma\sqrt{\lambda(t)}dB(t) \quad (2)$$

where $\kappa, \sigma$ and $\alpha(\cdot)$ are model parameters and $B(\cdot)$ 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 $\kappa$, mean-reversion level $\alpha(t)$ at time $t$, and volatility rate parameter $\sigma$. In the context of call centers, the function $\alpha(\cdot)$ can be used to model the predictable time-varying patterns such as the time-of-day effect in arrivals. Throughout, we assume that the following Feller condition is satisfied so that $\lambda(\cdot)$ is always positive.

**Assumption 1.** The model parameters $\kappa, \sigma, \alpha(\cdot)$ satisfy $2\kappa\alpha_\downarrow \geq \sigma^2$ where $\alpha_\downarrow \equiv \inf_t \alpha(t)$.

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

a (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 $(t_1, t_2)$,

$$E[A[t_1, t_2]|\mathcal{F}_{t_1}] \overset{a.s.}{=} E\left[\int_{t_1}^{t_2} \lambda(u)du \bigg| \mathcal{F}_{t_1}\right],$$

where we have defined $A[t_1, t_2] \equiv A(t_2) - A(t_1)$.

b (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. We will apply this martingale property in the proofs of Lemma 2 and the main theorems.

c (First-order behavior) Let $\bar{\lambda}(t) \equiv E[\lambda(t)]$ for each $t \geq 0$. Taking expectation on both sides of (2), we obtain

$$\bar{\lambda}(t) = \bar{\lambda}(0) + \int_0^t \kappa(\alpha(u) - \bar{\lambda}(u))du,$$

whereupon we see

$$\bar{\lambda}(t) = e^{-\kappa t}\bar{\lambda}(0) + \int_0^t \kappa e^{-\kappa(t-u)}\alpha(u)du.$$

Below, we will be using the function $\bar{\lambda}(\cdot)$ in characterizing the process limits of the arrival process.
As alluded to earlier, the CIR process permits versatile correlation structure while maintaining analytical tractability. In addition, it is very easy to simulate. Indeed, when the mean-reversion level is time-varying as in (2), a sample path of $\lambda$ can be produced by Euler discretization. If the mean-reversion level is constant, then there is an exact simulation algorithm by sampling chi-square random variables, see, e.g., §3.4 in Glasserman (2013). To develop an intuitive understanding of the CIR-driven arrival process, we depict a simulated sample path of a time-stationary CIR process and the corresponding arrival times in Fig. 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 thinning method for nonhomogeneous Poisson process; see Chapter 2 of Ross (1996). As expected, we observe that arrivals occur more frequently as the intensity ramps up.

2.3. The $M_{CIR}/M/s_t + G$ model

We are now ready to move to the corresponding queueing model. To start, let customers arrive to the system according to a DSPP with CIR-driven intensity process $\lambda(\cdot)$ given by (2). Service times are assumed to be i.i.d. exponential random variables with rate $\mu$. We allow waiting customers to abandon the queue, and assume the abandonment times of successive arrivals to be i.i.d. random variables with cumulative distribution function (cdf) $F$ and probability density function (pdf) $f$. Moreover, we stipulate that service times and abandonment times are mutually independent, independent of the arrival processes.

At this point we make a key assumption which will be useful in obtaining the FCLT for the
overloaded $M_{CIR}/M/s_t + G$ model.

**Assumption 2.** The pdf $f$ of the abandonment time is bounded and the complementary cdf $F^c(x) \equiv 1 - F(x)$ is bounded away from zero on any compact interval.

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 $E(t)$ and $R(t)$ to represent the number of customers that have entered service and the number of abandonments from the queue, all up to time $t$, respectively. By flow conservation,

$$Q(t) = Q(0) + A(t) - E(t) - R(t). \quad (4)$$

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

$$B(t) = B(0) + E(t) - D(t). \quad (5)$$

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 (4) and (5) yields

$$X(t) = Q(t) + B(t) = X(0) + A(t) - D(t) - R(t). \quad (6)$$

Alternatively, one can derive (6) directly from flow conservation.

We now introduce two waiting-time processes that we will exploit heavily in the subsequent analysis. Let $H(t)$ denote the head-of-line waiting time (HWT) at time $t$, i.e., the waiting time of the customer who has been waiting the longest (if there is any); $H(t) = 0$ if there is no customer waiting in queue. Let $V(t)$ represent the potential waiting time (PWT) at time $t$, i.e., the waiting time of an arriving customer at time $t$ assuming the customer has infinite patience. With the newly introduced processes, namely, $H(t)$ and $V(t)$, we can conveniently express the enter-service and queue-length processes in the following way:

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

$$Q(t) = \sum_{i=A(t-H(t))}^{A(t)} 1\{\tau_i + \gamma_i > t\} \quad \text{for } t \geq 0, \quad (7)$$

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. As will become clear
later on, these representations are especially useful in deriving FCLT results for the overloaded $M_{\text{CIR}}/M/s_{t} + G$ system.

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 (PoD)

$$\mathbb{P}(V(t) > 0) \leq \rho, \quad (9)$$

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., Garnett et al. (2002); Feldman et al. (2008).

Let $\alpha^{*}$ be the long-run average mean-reversion level, namely,

$$\alpha^{*} = \lim_{T \to \infty} T^{-1} \int_{0}^{T} \alpha(u)du. \quad (10)$$

We can then spell out our staffing formula for the $M_{\text{CIR}}/M/s_{t} + G$ model

$$s(t) = \lceil m(t) + \sqrt{\alpha^{*}c(t)} \rceil, \quad (11)$$

where the so-called offered load function $m$ solves the differential equation

$$\dot{m}(t) = \bar{\lambda}(t) - \mu m(t) \quad \text{or} \quad m(t) - m(0) = \int_{0}^{t} \bar{\lambda}(u)du - \mu \int_{0}^{t} m(u)du, \quad (12)$$

and $c(\cdot)$ is a design function to be specified to meet the performance target. It is readily checked that the function $m(\cdot)$ coincides with the offered load process for an $M_{t}/M/\infty$ model with arrival-rate function $\bar{\lambda}(\cdot)$.

If the shift function $\alpha(\cdot)$ is constant, i.e., $\alpha(t) \equiv \alpha$ for some $\alpha \in \mathbb{R}$, and further $c(t) \equiv c^{*}/\sqrt{\mu}$ for some $c^{*} \in \mathbb{R}$, then by (12), we have $m(t) = m(\infty) = \alpha/\mu$ for all $t \geq 0$, and so

$$s(t) = s \equiv \lceil \alpha/\mu + c^{*}\sqrt{\alpha/\mu} \rceil.$$ 

We thus recover the classical square-root staffing rule for the time-stationary model.

Figure 2 shows the histograms of customer waiting times for an $M_{\text{CIR}}/M/s + M$ model with different volatility parameters. 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. It is thus imperative to develop tractable approximations for relevant performance measures and derive rule-of-thumb staffing algorithms in the presence of autoregressive inputs; that is the theme of the remainder of the the article.
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 §3.1. The main results are presented in §3.2.

3.1. Asymptotic framework

We consider an asymptotic framework in which the long-run average demand volume grows to infinity, i.e., $\alpha^* \to \infty$ for $\alpha^*$ given by (10). Following the convention in the literature, we will use $n$ in place of $\alpha^*$ as the scaling parameter. More precisely, we apply a linear scaling to the function $\alpha$

$$\alpha_n(t) \equiv n\alpha(t) \quad \text{for} \quad t \geq 0,$$

where, by slight abuse of notation, we used $\alpha(t)$ to denote the baseline mean-reversion level at time $t$. 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 $\lambda_n$ satisfying

$$d\lambda_n(t) = \kappa(\alpha_n(t) - \lambda_n(t))dt + \sigma\sqrt{\lambda_n(t)}dB(t).$$

Figure 2: Estimated waiting-time distributions from computer simulation with $\alpha = 100, \kappa = 1, s = 100$, exponential services and abandonments with rates $\mu = 1$ and $\theta = 1$, respectively.
Note that $\kappa$ and $\sigma$ are fixed, i.e., 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
\[
\tilde{\lambda}_n(t) \equiv \frac{\lambda_n(t)}{n} \quad \text{and} \quad \hat{\lambda}_n(t) \equiv \sqrt{n} \left( \tilde{\lambda}_n(t) - \bar{\lambda}(t) \right).
\]

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

**Lemma 1** (FWLLN and FCLT for the arrival intensity process). Suppose that the intensity process for the $n$-th model follows (14). If, in addition, there is convergence of the initial distribution at time $0$, i.e., if
\[
(\tilde{\lambda}_n(0), \hat{\lambda}_n(0)) \Rightarrow (\bar{\lambda}(0), \hat{\lambda}(0)) \quad \text{in} \quad \mathbb{R}^2 \quad \text{as} \quad n \to \infty,
\]
then we have the joint convergence
\[
(\tilde{\lambda}_n(t), \hat{\lambda}_n(t)) \Rightarrow (\bar{\lambda}(t), \hat{\lambda}(t)) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty,
\]
where $\bar{\lambda}(\cdot)$ follows (3) and $\hat{\lambda}(\cdot)$ satisfies the stochastic integral equation
\[
\hat{\lambda}(t) = \hat{\lambda}(0) - \kappa \int_0^t \hat{\lambda}(u)du + \sigma \int_0^t \sqrt{\bar{\lambda}(u)}dB(u).
\]
Hence the limit of the arrival-rate process is an Ornstein-Uhlenbeck process whose solution admits a closed-form expression:
\[
\hat{\lambda}(t) = e^{-\kappa t} \hat{\lambda}(0) + \sigma \int_0^t e^{-\kappa(t-u)} \sqrt{\bar{\lambda}(u)}dB(u).
\]

To proceed, we define $\Lambda(t) \equiv \int_0^t \bar{\lambda}(u)du$ and the scaled versions of the arrival process:
\[
\bar{A}_n(t) \equiv A_n(t)/n, \quad \hat{A}_n(t) \equiv n^{-1/2} (A_n(t) - n\Lambda(t)).
\]

Per our previous discussion, it is natural to center $A_n(t)$ around $n\Lambda(t)$. We will show in Lemma 2 below that this centering indeed gives rise to meaningful limit.

**Lemma 2** (FWLLN and FCLT for the arrival process). The centered and normalized version of the arrival process $\hat{A}_n$ satisfies an FCLT:
\[
\hat{A}_n(t) \Rightarrow \hat{A}(t) \equiv \mathcal{B}_\lambda \left( \int_0^t \bar{\lambda}(u)du \right) + \mathcal{K}(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty.
\]
for

$$K(t) \equiv \int_0^t \hat{\lambda}(u) \, du,$$

where $\hat{\lambda}(\cdot)$ is given in (17) and $B_\lambda(\cdot)$ is a standard Brownian motion independent of $K(\cdot)$. As an immediate consequence, we have the WFLLN

$$\tilde{A}_n(t) \Rightarrow \Lambda(t) \text{ in } D \text{ as } n \to \infty,$$

jointly with (20).

According to Lemma 2, the diffusion-scaled arrival process $\tilde{A}_n(\cdot)$ converges to a Gaussian process which is characterized by two independent terms. The first term is a time-changed Brownian motion that arises from the inherent variability in the Poisson process $\Pi_n$, while the second term is an integrated Ornstein-Uhlenbeck process that stems from the stochasticity of the intensity process. If the mean-reversion level of the CIR process is constant, i.e., $\alpha_n(t) \equiv n$, then Lemma 2 applies to the time-stationary model as an important special case.

**Corollary 1** (the stationary case). If the CIR-driven intensity $\lambda_n$ mean-reverts to a constant $n$ and $\hat{\lambda}(0) = 0$, then the scaled arrival process $\tilde{A}_n$ satisfies an FCLT:

$$\tilde{A}_n(t) \Rightarrow \hat{A} \equiv B_\lambda(t) + K(t),$$

where $B_\lambda$ is a standard Brownian motion and

$$K(t) \equiv \frac{\sigma^2}{\kappa^3} \int_0^t \left(1 - e^{-\kappa(t-u)}\right) d\mathcal{B}(u),$$

(22)

where $\mathcal{B}$ is a standard Brownian motion independent of $B_\lambda$.

For $s \leq t$ the covariance between $\hat{A}(s)$ and $\hat{A}(t)$ can be computed using the following formula

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

In particular, the formula for the variance is given by

$$\text{Var}(\hat{A}(t)) = \left(1 + \frac{\sigma^2}{\kappa^2}\right) t - \frac{3\sigma^2}{2\kappa^3} + 2\sigma^2 \kappa^2 e^{-\kappa t} - \frac{\sigma^2}{2\kappa^3} e^{-2\kappa t}.$$

(23)

When $\sigma = 0$, we have $\text{Var}(A_n(t)) = nt$. 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 (23) it follows that the variance can be 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 $n$-th model, customers arrive to the system according to $A_n$, i.e., $A_n(t)$ represents the number of arrivals over $[0,t]$. The service rate $\mu$ and the abandonment-time distribution $F$ are held fixed, but the staffing process is allowed to grow linear with $n$, so that the corresponding staffing function satisfies

$$s_n(t) = \lceil nm(t) + \sqrt{n}c(t) \rceil,$$

where the base-line offered load function $m$ is in the form of (12) and $c$ is a control function to be determined to meet the performance target (9).

**Heavy-traffic scalings.** For the headcount and queue-length processes, define their centered and normalized versions as follows:

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

It is well known that for a many-server queue having customer abandonment and operating in the rationalized regime, only the density of the abandonment-time distribution at the origin plays a role in the corresponding diffusion limit; cf. Zeltyn and Mandelbaum (2005). Therefore, without loss of generality we may assume the abandonment times to be exponentially distributed with rate $\theta$, in which case Assumption 2 is trivially satisfied.

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

**Theorem 1** (FCLT for the critically-loaded $M_{\text{cir}}/M/s_t+M$ model). Suppose customers arrive according to the DSPP $A_n(\cdot)$ with intensity process $\lambda_n(\cdot)$ given by (14), the system is staffed according to (24). If in addition we have

$$\hat{X}_n(0) \Rightarrow \hat{X}(0) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \rightarrow \infty,$$

jointly with (15), then

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

jointly with (16), where $\hat{Q}(t) = [\hat{X}(t)]^+, \hat{V}(t) = \hat{Q}(t)/\hat{\lambda}(t)$, and the limiting process $\hat{X}$ satisfies

$$\hat{X}(t) = \hat{X}(0) - c[0,t] - \mu \int_0^t c(u)du - \int_0^t \Psi(\hat{X}(u))du + \mathcal{K}(t) + \int_0^t \vartheta(u)d\mathcal{B}_2(u),$$

for $\Psi(x) \equiv \mu x^- + \theta x^+$, $\mathcal{K}$ given by (21), $\vartheta(t) \equiv (\hat{\lambda}(t) + \mu m(t))^{1/2}$, and $\mathcal{B}_2$ being a standard Brownian motion, independent of the process $\mathcal{K}$. 

12
Theorem 1 applies to the stationary model as an important special case when the mean-reversion level is constant. In particular, the mean-reversion level is equal to $n$ for the $n$-th stochastic model.

**Corollary 2** (the stationary case). Suppose that the conditions in Theorem 1 are satisfied. In addition, $\alpha_n \equiv n$ in equation (14) and $c(t) \equiv c$ for some constant $c$. Then

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

where, $\hat{Q}(t) = [\hat{X}(t)]^+, \hat{V}(t) = \hat{Q}(t)/\hat{\lambda}(t)$, and the limiting process $\hat{X}$ satisfies

$$\hat{X}(t) = \hat{X}(0) - \mu c t - \int_0^t \Psi(\hat{X}(u))du + K(t) + \sqrt{2B_2(t)},$$

for $K$ specified by (22) and $B_2$ being a standard Brownian motion, independent of $K$.

Corollary 2 is in agreement with Theorem 2 of Garnett et al. (2002) except for the additional term $K(\cdot)$ that arises naturally from the autoregressive assumption of the arrival-rate process. Note 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. We will explore this issue further in §4.

4. Implications on Call Center Staffing

One of the main challenges in operating a telephone call center is finding correct staffing levels to meet desired performance target. 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 and Zeltyn (2009); see also Liu and Whitt (2017) for the case with customer feedback. We will argue that, 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.

For clarity of exposition, we take the mean-reversion level of the CIR-driven intensity process to be a constant so that $\alpha(t) \equiv \alpha$. To gain greater simplicity and tractability, we first consider a special case where service rate and the abandonment rate are equal, in which scenario we would get $\Psi(x) \equiv \mu x^- + \theta x^+ = \mu x$. By Corollary 2, the limiting headcount process $\hat{X}$ satisfies

$$\hat{X}(t) = e^{-\mu t} \hat{X}(0) - c \int_0^t \mu e^{-\mu (t-u)}du + \int_0^t e^{-\mu (t-u)}\hat{\lambda}(u)du + \sqrt{2} \int_0^t e^{-\mu (t-u)}dB_2(u).$$  \hspace{1cm} (25)
From (25) we can directly compute the mean

$$E\left[\hat{X}(t) \big| \hat{X}(0)\right] = e^{-\mu t} \hat{X}(0) - c \int_0^t \mu e^{-\mu(t-u)} du \to -c \quad \text{as} \quad t \to \infty.$$ 

To derive the variance formula for $\hat{X}(t)$, note that

$$\text{Var}(X(t)) = \text{Var} \left( \int_0^t e^{-\mu(t-u)} \lambda(u) du \right) + 2\text{Var} \left( \int_0^t e^{-\mu(t-u)} dB_2(u) \right).$$

For the first term on the right-hand side, we apply Fubini’s theorem and Itô isometry to receive

$$Y_1(t) = \text{Var} \left( \sigma \left( \frac{\alpha}{\mu} + \sqrt{\alpha c} \right) \right) = \frac{\sigma^2}{(\mu - \kappa)^2} \int_0^t \left( e^{-\kappa(t+s)} - e^{-\mu(t+s)} \right)^2 ds$$

$$= \frac{\sigma^2}{(\mu - \kappa)^2} \left[ \frac{1}{2\kappa} \left( 1 - e^{-2\mu t} \right) - \frac{2}{\kappa + \mu} \left( 1 - e^{-(\kappa + \mu)t} \right) + \frac{1}{2\mu} \left( 1 - e^{-2\mu t} \right) \right]$$

Applying Itô isometry again gives us

$$Y_2(t) = 2 \int_0^t e^{-2\mu(t-u)} du = \frac{1}{\mu} \left( 1 - e^{-2\mu t} \right).$$

Combining above and sending $t \to \infty$ gives the following result about the long-term behavior of the limiting process $\hat{X}$.

**Proposition 1.** As $t \to \infty$, the sequence of random variables $\hat{X}(t)$ in (25) converges weakly to $\hat{X}(\infty)$ which is a Gaussian random variable with mean $-c$ and variance

$$\text{Var}(\hat{X}(\infty)) = \frac{1}{\mu} + \frac{\sigma^2}{2\kappa\mu(\kappa + \mu)}.$$

Recall the number of servers was selected according to the square-root staffing formula, which we reproduce here for convenience

$$s \equiv \left\lceil \frac{\alpha}{\mu} + \sqrt{\alpha c} \right\rceil.$$  \hspace{1cm} (26)

Then Proposition 1 states that when the mean-reversion level $\alpha$ is large, we can heuristically approximate the steady-state distribution of the number of customers in the $M_{\text{CIR}}/M/s + M$ model (having equal service and abandonment rates) as follows:

$$X(\infty) = \frac{\alpha}{\mu} + \sqrt{\alpha k N} \quad \text{for} \quad k \equiv k(\mu, \kappa, \sigma) \equiv \sqrt{\frac{1}{\mu} + \frac{\sigma^2}{2\kappa\mu(\kappa + \mu)}},$$

where $N$ denotes a standard normal random variable. Combining (26) and (27) yields a simple normal approximation:

$$P(V(\infty) > 0) = P(X(\infty) > s) \approx P(N > c/k).$$
From this normal approximation we immediately understand that, in order to stabilize the delay probability at the target value \( \varrho \), one ought to choose

\[
c \equiv z_1 \cdot \sqrt{\frac{1}{\mu} + \frac{\sigma^2}{2\kappa\mu(\kappa + \mu)}}
\]  

in the square-root staffing formula (26).

The result of (28) 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.

![Figure 3: Probabilities of delay for nine delay-probability targets with \( \alpha = 100, \kappa = 1, \sigma = 2 \), exponential services and abandonments with rates \( \mu = 1 \) and \( \theta = 1 \), respectively.](image)
We now show that the propose formula achieves the desired time-stable performance. In Figure 3, we plot the delay probabilities obtained from computer simulation with targets \( \varrho = 0.1, 0.2, \ldots, 0.9 \). These delay probabilities are estimated by performing 2000 independent replications with the staffing function specified by (26) and (28). The fluctuations are largely due to the inherent discreteness associated with staffing levels; that is, the staffing levels must be integers. We observe that delay probabilities fluctuate around the target in each case., i.e., the probabilities of delay are stabilized remarkably well.

It is widely recognized that a normal approximation may not be appropriate when \( \theta \neq \mu \); see, e.g., the discussion in §6 of Feldman et al. (2008). Here we restrict our attention to the case where there is no customer abandonment (\( \theta = 0 \)) and leave the more general scenario (\( \theta \neq \mu \) and \( \theta > 0 \)) to future research. By Corollary 2, the limiting headcount process \( \hat{X} \) satisfies

\[
\hat{X}(t) = \hat{X}(0) - \mu ct - \mu \int_0^t \hat{X}(u) \, du + K(t) + \sqrt{2}B_2(t).
\]  

(29)

The proposition below spells out the stationary distribution of \( \hat{X} \), the limiting process given in (29). Its proof is deferred to §Appendix A.

**Proposition 2.** The limiting process specified by (29) has a stationary distribution whose density is given by

\[
p(x) = \begin{cases} 
\frac{c}{k} \exp\{-cx/k\} \varrho, & \text{if } x \geq 0, \\
\frac{\phi(c/k+x)}{\phi(c/k)} (1 - \varrho), & \text{if } x < 0,
\end{cases}
\]

where \( k \) is given by (27) and \( \varrho \equiv \varrho(c/k) = \left[ 1 + \frac{(c/k)\Phi(c/k)}{\phi(c/k)} \right]^{-1} = P(\hat{X}(\infty) \geq 0). \)

Proposition 2 suggests that in order to stabilize the delay probability at the target value \( \varrho \) for an \( M_{\text{CIR}}/M/s \) system, it suffices to choose \( c \) in the square-root staffing formula (26) that solves the equation

\[
1 + \frac{(c/k)\Phi(c/k)}{\phi(c/k)} = 1/\varrho.
\]  

(30)

In Figure 4, we show the delay probabilities obtained from computer simulation with targets \( \varrho = 0.25, 0.5, 0.75 \). Again, we estimate these delay probabilities by performing 2,000 independent replications with the staffing function specified by (26) and (30). We observe that for each of the three cases the delay probabilities are remarkably accurate and stable.
Figure 4: Probabilities of delay for $\varrho = 0.25, 0.5, 0.75$, with $\alpha = 100, \kappa = 1, \sigma = 2$, exponential services and abandonments with rates $\mu = 1$ and $\theta = 0$, respectively.

5. An Alternative Scaling

In general, the performance metric in a call center can be expressed in terms of the tail probability of delay (TPoD), namely,

$$\mathbb{P} (V(t) > w) \leq \varrho,$$

where $w$ is a delay target and $\varrho$ is a constant probability target. Note that when $w = 0$, the TPoD reduces to the delay probability. It is also known that with $w > 0$ the constraint (31) entails an overloaded system, in which case customer waiting times tend to be comparable to the service times; see e.g., Liu (2018); Mandelbaum and Zeltyn (2009). This gives rise to a different staffing algorithm which we discuss below.

To formally define the overloaded regime, we adopt the offered load analysis which estimates the required service capacity by estimating how much capacity would be used if there were not limit on its availability. For example, consider a single-class $M_t/G/s_t + G$ model having Poisson arrivals rate $\lambda(t)$, independent and i.i.d. service times with a general distribution $G$ (the first $G$), and i.i.d. customer abandonment following a general distribution $F$ (the $+G$). Although the $M_t/G/s_t + G$ model is complicated, the corresponding $M_t/G/\infty$ infinite-server model remains
remarkably tractable, where the number of customers (or busy servers) follows a Poisson distribution with mean

\[ m_\infty(t) \equiv \mathbb{E}[X_\infty(t)] = \int_0^t \lambda(u)G^c(t - u)du. \]  

(32)

If the objective is to stabilize the expected delay at any given point in time at a target \( w \), one will need to set the staffing levels to a modified version of (32), namely,

\[ m_{\text{DIS}}(t) \equiv \int_w^t F^c(w)\lambda(u - w)G^c(t - u)du, \]  

(33)

where we have used DIS to denote the “delayed-infinite-server approximation”, as in Liu and Whitt (2012). The effective arrival rate can be justified by the fact that, if every arrival who does not elect to abandon waits \( w \) time units, then a fraction \( F(w) \) of arrivals will abandon the queue before entering service. In other words, one can think of \( m_{\text{DIS}}(t) \) as the mean number of busy servers needed to serve all customers who are willing to wait for \( w \) time units.

Thus far we have implicitly assumed that the system starts empty at time 0 and a staffing process \( m(\cdot) \) begins at time \( w \). With this assumption, the policy should provide no staffing at all and thus allows no customer to enter service until time \( w \). By replacing \( G \) in (33) with an exponential cdf we obtain

\[ m(t) = F^c(w)\int_w^t \lambda(u - w)e^{-\mu(t-u)}du. \]  

(34)

The result of (34) was previously derived by Liu and Whitt (2012); see Eq. (13) there. Then \( s_n(\cdot) \) can be specified by (24) in which \( m \) is in the form of (34).

**Heavy-traffic scalings.** For the queue-length process \( Q_n \), define the centered and normalized version:

\[ \hat{Q}_n(\cdot) \equiv n^{-1/2} (Q_n(\cdot) - nq(\cdot)) \quad \text{for} \quad q(t) = \int_{t-w}^t F^c(t-u)\lambda(u)du. \]

Next we define the diffusion-scaled HWT and PWT processes

\[ \hat{H}_n(\cdot) \equiv \sqrt{n} (H_n(\cdot) - w) \quad \text{and} \quad \hat{V}_n(\cdot) \equiv \sqrt{n} (V_n(\cdot) - w). \]

Our next theorem establishes the FCLT result for the above centered and normalized versions.

**Theorem 2** (FCLT for the overloaded \( M_{\text{CIR}}/M/s_t + G \) model). Suppose customers arrive according to the DSPP \( A_n(\cdot) \) with intensity process \( \lambda_n(\cdot) \) given by (14) and the system is staffed according to (24) in which \( m(\cdot) \) is specified by (34). In addition, Assumptions 1 and 2 hold. Then

\[ (\hat{H}_n, \hat{V}_n, \hat{Q}_n) \Rightarrow (\hat{H}, \hat{V}, \hat{Q}) \quad \text{in} \quad \mathcal{D}^3 \quad \text{as} \quad n \to \infty, \]
jointly with (16), where \( \hat{H} \) satisfies the stochastic integral equation

\[
F^c(w)\bar{\lambda}(t-w)\hat{H}(t) = -\int_w^t f(w)\bar{\lambda}(u-w)\hat{H}(u)du - c[w,t] - \mu \int_w^t c(u)du + \mathcal{G}(t),
\]

for the noise term \( \mathcal{G}(t) \) given by

\[
\mathcal{G}(t) \equiv F^c(w)\left[ \int_0^{t-w} \sqrt{\lambda(u)}dB_\lambda(u) + F^c(w)\int_0^{t-w} \dot{\lambda}(u)du 
- \int_0^t \sqrt{\mu n(u)}dB_\mu(u) + \sqrt{F^c(w)F(w)}\int_0^{t-w} \sqrt{\lambda(u)}dB_\theta(u) \right],
\]

where \( B_\lambda, B_\mu, B_\theta \) are mutually independent Brownian motions. The limit for the PWT process is a deterministic functional of \( \hat{H} \) and satisfies \( \dot{V}(t) = \hat{H}(t+w) \); the FCLT limit for queue length \( \hat{Q} \) is the sum of four terms, namely,

\[
\hat{Q}(t) \equiv \hat{Q}_1(t) + \hat{Q}_2(t) + \hat{Q}_3(t) + \hat{Q}_4(t),
\]

where

\[
\begin{align*}
\hat{Q}_1(t) & \equiv \int_{t-w}^t F^c(t-u)\sqrt{\lambda(u)}dB_\lambda(u), \\
\hat{Q}_2(t) & \equiv \int_{t-w}^t F^c(t-u)\dot{\lambda}(u)du, \\
\hat{Q}_3(t) & \equiv \int_{t-w}^t \sqrt{F^c(t-u)F(t-u)}\lambda(u)dB_\mu(u), \\
\hat{Q}_4(t) & \equiv F^c(w)\lambda(t-w)\hat{H}(t).
\end{align*}
\]

Separation of variability. The noise term \( \mathcal{G} \) in (36) is characterized by four independent terms. The first term captures the inherent variability in the Poisson process \( \Pi_a \) given in (1); the second term stems from the stochasticity of the arrival-rate process; the third term comes from the variability of the service-completion process; and finally the fourth term accounts for the randomness of customer abandonment.

Just like what we did in Corollary 2, we obtain an important special case for the overloaded \( \text{M/M/1}/M/s_t + G \) system when the intensity process mean reverts to a constant level.

**Corollary 3** (the stationary case). Suppose that the conditions in Theorem 2 are satisfied. In addition, \( \alpha_n(t) \equiv n \), and \( c(t) \equiv c \) for some constant \( c \). Then

\[
(\mathcal{H}_n, \hat{V}_n) \Rightarrow (\hat{H}, \hat{V}) \quad \text{in} \quad \mathcal{D}^2 \quad \text{as} \quad n \to \infty,
\]

where \( \hat{V}(t) = \hat{H}(t+w) \) and \( \hat{H} \) satisfies the stochastic integral equation

\[
\hat{H}(t) = -\frac{\mu c}{F^c(w)}(t-w) - h_F \int_w^t \hat{H}(u)du + \mathcal{G}(t),
\]

for \( h_F \equiv h_F(w) \) being the value of the hazard rate function of the distribution \( F \) evaluated at the delay target \( w \) and the noise term \( \mathcal{G}(t) \) given by

\[
\mathcal{G}(t) \equiv \frac{1}{\sqrt{F^c(w)}} \left( \int_0^{t-w} dB_2(u) - \int_w^t dB_\mu(u) \right) + \int_0^{t-w} \lambda(u)du,
\]

where \( B_2 \) and \( B_\mu \) are two independent Brownian motions, independent of \( \dot{\lambda} \).
When the arrival rate is deterministic (i.e., $\hat{\lambda}$ degenerates to zero), the last term of (38) disappears and we recover the FCLT results of Aras et al. (2018) for an overloaded $M/M/s + G$ system as an important special case.

Acknowledgement

This research was partly supported by National Science Foundation (CMMI-1634133) and was completed when the author was a doctoral student at Columbia University.

Appendix A. Proofs

In this section, we provide the proofs for Lemma 1, Lemma 2, Theorem 1, Proposition 2 and Theorem 2.

Proof of Lemma 1. Dividing both sides of (14) by $n$ and in view of (13), we obtain
\begin{equation}
\bar{\lambda}_n(t) = \bar{\lambda}_n(0) + \kappa \int_0^t (\alpha(u) - \bar{\lambda}_n(u))du + \sigma_n \int_0^t \sqrt{\bar{\lambda}_n(u)}dB(u), \tag{A.1}
\end{equation}
where we have defined $\sigma_n \equiv \sigma/n$. We prove the WFLLN 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 (2013) for a formal 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{align*}
\mathbb{E} \left[ (\bar{\lambda}_n(t + u) - \bar{\lambda}_n(t))^2 \mid \mathcal{F}_t \right] & \leq 2\mathbb{E} \left[ \left( \int_t^{t+u} \kappa (\alpha(s) - \bar{\lambda}_n(s))ds \right)^2 \right] + 2\sigma_n^2 \mathbb{E} \left[ \left( \int_t^{t+u} \sqrt{\bar{\lambda}_n(s)}dB(s) \right)^2 \right] \\
& \leq 2u \mathbb{E} \left[ \int_t^{t+u} \kappa^2 (\alpha(s) - \bar{\lambda}_n(s))^2 ds \right] + 2\sigma_n^2 C_p \mathbb{E} \left[ \int_t^{t+u} \bar{\lambda}_n(s)ds \right] \\
& \leq 2u \int_0^T \kappa^2 \left( (\alpha(s))^2 + 2\alpha(s)\mathbb{E} [\bar{\lambda}_n(s)] + \mathbb{E} [\bar{\lambda}_n^2(s)] \right) ds + \sigma_n^2 C_p u^{1/4} \left( 1 + \int_0^T \mathbb{E} [\bar{\lambda}_n^2(s)] ds \right),
\end{align*}
where the first inequality follows from (A.1), the second inequality follows by applying the Cauchy-Schwarz 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-Schwarz inequality. By
Lemma 3 below we conclude that both integrals on the right-hand side approach zero as \( u \to 0 \), uniformly over all \( t \) 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 \( \{ \tilde{\lambda}_n; n \in \mathbb{N} \} \) from which the desired WFLLN follows, namely,

\[
\tilde{\lambda}_n \Rightarrow \lambda \quad \text{in} \quad D \quad \text{as} \quad n \to \infty.
\]  

(A.2)

Next subtract (3) from (A.1) and scale up both sides of the resulting equation by \( \sqrt{n} \) to get

\[
\dot{\lambda}_n(t) = \dot{\lambda}_n(0) - \kappa \int_0^t \dot{\lambda}_n(u)du + \sigma \int_0^t \sqrt{\lambda_n(u)}dB(u).
\]  

(A.3)

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

\[
y(t) = b - \kappa \int_0^t y(u)du + \sigma \int_0^t \sqrt{x(u)}dB(u) \quad \text{for} \quad t \geq 0
\]

is continuous. We can therefore invoke the continuous mapping theorem with the established weak convergence in (A.2) to obtain the desired FCLT for the sequence \( \{ \dot{\lambda}_n; n \in \mathbb{N} \} \).

\[\square\]

**Lemma 3.** If \( \bar{\lambda}_n \equiv \lambda_n/n \). Then for any finite \( T > 0 \) and \( k \in \mathbb{N} \), we have

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

**Proof of Lemma 3.** Recall that

\[
d\bar{\lambda}_n(t) = \kappa(\dot{\alpha}(t) - \bar{\lambda}_n(t))dt + \sigma_n\sqrt{\lambda_n(t)}dB(t),
\]

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

\[
\dot{\lambda}_n^k(t) = \dot{\lambda}_n^k(0) + k\kappa \int_0^t \alpha(u)\dot{\lambda}_n^{k-1}(u)du - k\kappa \int_0^t \lambda_n(u)du + k\sigma_n \int_0^t \lambda_n^{k-1/2}(u)dB(u).
\]  

(A.4)

An application of Young’s inequality yields

\[
\dot{\lambda}_n^{k-1}(u) \leq (k - 1)\dot{\lambda}_n^k(u)/k + 1/k.
\]

Substituting the above into (A.4) gives

\[
\dot{\lambda}_n^k(t) \leq \dot{\lambda}_n^k(0) + \kappa \int_0^t [(k - 1)\alpha(u) - k] \dot{\lambda}_n^k(u)du + \kappa \int_0^t \alpha(u)du + \sigma_n \int_0^t \lambda_n^{k-1/2}(u)dB(u).
\]

Taking expectation on both sides, we get

\[
\mathbb{E}\left[ \dot{\lambda}_n^k(t) \right] \leq \mathbb{E}\left[ \dot{\lambda}_n^k(0) \right] + \kappa \int_0^t [(k - 1)\alpha(u) - k] \mathbb{E}\left[ \dot{\lambda}_n^k(u) \right] du + \kappa \int_0^t \alpha(u)du.
\]

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

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

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

\[\square\]
Proof of Lemma 2. First use (19) to write
\[ \hat{A}_n(t) = \hat{A}_{n,1}(t) + \hat{A}_{n,2}(t) \]
where we defined
\[ \hat{A}_{n,1}(t) \equiv n^{-1/2} \left( A_n(t) - \int_0^t \lambda_n(u)du \right) \quad \text{and} \quad \hat{A}_{n,2}(t) \equiv \int_0^t \lambda_n(u)du. \tag{A.5} \]
The first term is a square integrable martingale with quadratic variation \( \int_0^t \lambda_n(u)du \) converging to \( \int_0^t \lambda(u)du \) as \( n \to \infty \). Appealing to the martingale FCLT, we obtain
\[ \hat{A}_{n,1}(t) \Rightarrow B_\lambda \left( \int_0^t \lambda(u)du \right) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty. \]
For the second term, we obtain the convergence \( \hat{A}_{n,2}(t) \Rightarrow \mathcal{K}(t) \) by applying the continuous mapping theorem with (16). Combining the above yields the desired result. \( \square \)

Proof of Theorem 1. The proof adopts the martingale approach for establishing the FCLT for many-server queueing systems; see, e.g. Puhalskii (2013). As the proof is fairly standard, we detail the key steps only.

First, multiple both sides of (12) by \( n \) and subtract it from (6) to get
\[ X_n(t) - nm(t) = X_n(0) - nm(0) + A_n(t) - n\Lambda(t) - \left( D_n(t) - \mu \int_0^t B_n(u)du \right) \]
\[ - \mu \int_0^t \left( B_n(u) - nm(u) \right) du - \left( R_n(t) - \theta \int_0^t Q_n(u)du \right) - \theta \int_0^t Q_n(u)du. \]
Dividing both sides by \( \sqrt{n} \) and exploiting the relations
\[ B_n(\cdot) = X_n(\cdot) \wedge s_n(\cdot) \quad \text{and} \quad Q_n(\cdot) = [X_n(\cdot) - s_n(\cdot)]^+, \]
we obtain the stochastic integral equation satisfied by the diffusion-scaled process \( \hat{X}_n \)
\[ \hat{X}_n(t) = \hat{X}_n(0) - c[0,t] - \mu \int_0^t c(u)du - \int_0^t \Psi(\hat{X}_n(u))du + \hat{A}_n(t) - \hat{D}_n(t) - \hat{R}_n(t), \tag{A.6} \]
where \( \Psi \equiv \mu x^- + \theta x^+ \), \( \hat{A}_n(t) \) is defined by (19), and
\[ \hat{D}_n(t) \equiv n^{-1/2} \left( D_n(t) - \mu \int_0^t B_n(u)du \right) \quad \text{and} \quad \hat{R}_n(t) \equiv n^{-1/2} \left( R_n(t) - \theta \int_0^t Q_n(u)du \right) \]
are two square-integrable martingales. An application of the Gronwall’s inequality to (A.6) yields the stochastic boundedness of \( \{\hat{X}_n; n \in \mathbb{N}\} \). As a consequence, the quadratic variation of \( \hat{D}_n \) and \( \hat{R}_n \) converges to the corresponding limit, which in turn implies
\[ \hat{D}_n(t) \Rightarrow \hat{D}(t) \equiv \int_0^t \sqrt{\mu m(u)}dB_\mu(u) \quad \text{in} \quad D \quad \text{and} \quad \hat{R}_n(t) \Rightarrow 0 \quad \text{in} \quad D, \tag{A.7} \]
where $B_\mu$ is a standard Brownian motion and 0 denotes the zero function. Appealing to Theorem 4.1 in Pang et al. (2007) with (20) and (A.7) yields the desired FCLT result for $\{\hat{X}_n; n \in \mathbb{N}\}$. The assertion $\hat{Q}_n \Rightarrow \hat{Q}$ is fairly straightforward and follows from the simple relation between $\hat{X}_n$ and $\hat{Q}_n$. Finally, the FCLT for $\{\hat{V}_n; n \in \mathbb{N}\}$ follows by applying the two-parameter version of Puhalskii’s invariance principle for first passage time; see Theorem 2.9 in Talreja and Whitt (2009).

Proof of Proposition 2. The proof follows from Browne et al. (1995). Note that the process $\hat{X}$ restricted to $[0, \infty)$ is a superposition of the integrated Ornstein-Uhlenbeck process $\mathcal{K}$ and a Brownian motion with infinitesimal drift $-\mu c$ and variance 2. Hence its steady-state density conditional on $\hat{X}(\infty) \geq 0$ can be easily shown to be exponentially distributed with rate $c/k$. Similarly, the process $\hat{X}$ restricted to $(-\infty, 0)$ follows the dynamics of (25). Hence its steady-state density conditional on $\hat{X}(\infty) < 0$ is the density a normal random variable with mean $-c$ and variance $k$, conditioned on having negative values only. Putting these two pieces together yields the density $p(x)$ with $\varrho = \mathbb{P}(\hat{X}(\infty) \geq 0)$. To determine the value of $\varrho$, we note that the density $p(\cdot)$ must be continuous because the infinitesimal variance is continuous. Therefore $\varrho$ can be solved by equating the limits of $p(\cdot)$ at the origin from both sides.

Proof of Theorem 2. To begin, we observe that the PWT at the time of arrival of the head-of-line customer is the HWT at time $t$ plus the additional time until the next departure, namely,

$$V_n(t - H_n(t)) = H_n(t) + a_n(t). \quad (A.8)$$

where $a_n(t)$ denotes the remain time till one of busy servers becomes idle. Since the time between two successive departures follows an exponential random variable with rate $O(n)$, the total number of departures over any compact interval $[0, T]$ is $O(n)$. This allows us to conclude that $\sup_{t \leq T} \{a_n(t)\} = O(n^{-1} \log n)$. Combining with (A.8) yields

$$V_n(t - H_n(t)) = H_n(t) + O(n^{-1} \log n). \quad (A.9)$$

Next we define the diffusion- and fluid-scaled empirical processes

$$\bar{U}_n(t, x) \equiv \frac{1}{n} \sum_{k=1}^{\lfloor nt \rfloor} \mathbb{1}_{\{X_i \leq x\}} \text{ for } t \geq 0, \quad 0 \leq x \leq 1,$$

$$\hat{U}_n(t, x) \equiv \sqrt{n} \left( \bar{U}_n(t, x) - \mathbb{E} \left[ \bar{U}_n(t, x) \right] \right) = \frac{1}{\sqrt{n}} \left( \mathbb{1}_{\{X_i \leq x\}} - x \right), \quad (A.10)$$

where $X_1, X_2, \ldots$ are i.i.d. random variables uniformly distributed on $[0, 1]$. Krichagina and Puhalskii Krichagina and Puhalskii (1997) have shown that $\hat{U}_n \Rightarrow \hat{U}$ as $n \to \infty$, where $\hat{U}$ is the standard
Following the decomposition given in (A.11) - (A.14), we can write

\[ E_n(t) = E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t), \tag{A.11} \]

where

\[ E_{n,1}(t) \equiv \sqrt{n} \int_0^{t-H_n(t)} F^c(V_n(u))d\hat{A}_n(u), \quad t \geq 0, \tag{A.12} \]

\[ E_{n,2}(t) \equiv \sqrt{n} \int_0^{t-H_n(t)} \int_0^1 \mathbb{1}_{\{y>F^c(V_n(u))\}}d\hat{U}_n(A_n(u), y) \quad t \geq 0, \tag{A.13} \]

\[ E_{n,3}(t) \equiv n \int_0^{t-H_n(t)} F^c(V_n(u))\lambda(u)du \quad t \geq 0, \tag{A.14} \]

for \( \hat{A}_n, A_n \) given in (19) and \( \hat{U}_n \) specified by (A.10). Next define

\[ \varepsilon(t) \equiv \int_0^{t-w} F^c(w)\lambda(u)du \quad \text{and} \quad \hat{E}_n(t) \equiv n^{-1/2}(E_n(t) - n\varepsilon(t)). \tag{A.15} \]

Following the decomposition given in (A.11) - (A.14), we can write

\[ \hat{E}_n(t) = \hat{E}_{n,1}(t) + \hat{E}_{n,2}(t) + \hat{E}_{n,3}(t), \tag{A.16} \]

where

\[ \hat{E}_{n,1}(t) \equiv n^{-1/2}E_{n,1}(t) = \int_0^{t-H_n(t)} F^c(V_n(u))d\hat{A}_n(u), \tag{A.17} \]

\[ \hat{E}_{n,2}(t) \equiv n^{-1/2}E_{n,2}(t) = \int_0^{t-H_n(t)} \int_0^1 \mathbb{1}_{\{y>F^c(V_n(u))\}}d\hat{U}_n(A_n(u), y), \tag{A.18} \]

\[ \hat{E}_{n,3}(t) \equiv n^{-1/2}\left(E_{n,3}(t) - n \int_0^{t-w} F^c(w)\lambda(u)du\right). \tag{A.19} \]

For (A.19) we further deduce

\[ \hat{E}_{n,3}(t) = \sqrt{n} \left( \int_0^{t-H_n(t)} F^c(V_n(u))\lambda(u)du - \int_0^{t-w} F^c(w)\lambda(u)du \right) \]

\[ = \sqrt{n} \int_w^t F^c(H_n(u))\lambda(u-H_n(u))du - \sqrt{n} \int_w^t F^c(w)\lambda(u-w)du \]

\[ - \int_w^t F^c(H_n(u))\lambda(u-H_n(u))d\hat{H}_n(u) + O(n^{-1/2}\log n) \tag{A.20} \]

\[ = - \int_w^t \left\{ f(\eta_n(u))\lambda(u-\eta_n(u)) + F^c(\eta_n(u))\lambda'(u-\eta_n(u)) \right\} \hat{H}_n(u)du \]

\[ - \int_w^t F^c(H_n(u))\lambda(u-H_n(u))d\hat{H}_n(u) + O(n^{-1/2}\log n) \],

where the second equality follows by a change of variables \( t \to t - H_n(t) \) plus the use of relation (A.9), while the third equality follows by applying the mean-value theorem with \( \eta_n(t) \) satisfying

\[ \min\{H_n(t), w\} \leq \eta_n(t) \leq \max\{H_n(t), w\}. \tag{A.21} \]
Note that the law of conservation of flow implies

\[ E_n(t) = B_n(t) + D_n(t). \]  \hfill (A.22)

By the definition of \( \varepsilon(t) \) and (34) we deduce

\[ \varepsilon(t) = m(t) + \mu \int_w^t m(u)du. \]  \hfill (A.23)

Next scaling both sides of (A.23) by \( n \) and subtracting it from (A.22) yields

\[ E_n(t) - \varepsilon(t) = B_n(t) - nm(t) + D_n(t) - \mu \int_w^t nm(u)du. \]

Then divide both sides by \( \sqrt{n} \) and make use of the relation \( B_n(t) = s_n(t) \) (which holds with arbitrarily high probability for \( n \) sufficiently large) to obtain

\[ \hat{E}_n(t) = c[w,t] + \mu \int_w^t c(u)du + \hat{D}_n(t), \]  \hfill (A.24)

where we defined centered and normalized version of the departure process

\[ \hat{D}_n(t) \equiv n^{-1/2} \left( D_n(t) - \mu \int_w^t B_n(u)du \right). \]  \hfill (A.25)

Upon plugging (A.17) - (A.18) and (A.20) into (A.24), we establish a stochastic integral equation for the diffusion-scaled process \( \hat{H}_n \), namely,

\[
\int_w^t F^c(H_n(u))\lambda(u - H_n(u))d\hat{H}_n(u) = c[w,t] + \mu \int_w^t c(u)du + \hat{E}_{n,1}(t) + \hat{E}_{n,2}(t) - \hat{D}_n(t) \\
- \int_w^t \{ f(\eta_n(u))\lambda(u - \eta_n(u)) + F^c(\eta_n(u))\lambda'(u - \eta_n(u)) \} \hat{H}_n(u)du + O(n^{-1/2} \log n).
\]  \hfill (A.26)

We can then apply the Gronwall’s inequality to (A.26) with the stochastic boundedness of \( \hat{E}_{n,1}, \hat{E}_{n,2} \) and \( \hat{D}_n \) plus the assumed properties of \( F \) and \( f \) to conclude the stochastic boundedness of the sequence \( \{ \hat{H}_n; n \in \mathbb{N} \} \). Then the FWLLN for the HWT and PWT processes follows as a directly consequence of the established stochastic boundedness of the sequence \( \{ \hat{H}_n; n \in \mathbb{N} \} \), namely,

\[ (H_n, V_n) \Rightarrow (w, w) \in \mathcal{D}^2 \quad \text{as} \quad n \to \infty, \]  \hfill (A.27)

where the joint convergence holds by applying the converging-together lemma with (A.9). Having established the FWLLN for \( (H_n, V_n) \), we can invoke the continuous mapping theorem with (20) and (A.27) to derive

\[ \hat{E}_{n,1}(t) \Rightarrow \hat{E}_1(t) \equiv F^c(w) \int_0^{t-w} \sqrt{\lambda(u)}d\mathcal{B}_\lambda(u) + F^c(w) \int_0^{t-w} \lambda(u)du. \]  \hfill (A.28)
To show the convergence of the sequence \( \{ \hat{E}_{n,2}; n \in \mathbb{N} \} \), we invoke Theorem 7.1.4 in Ethier and Kurtz (2009) to obtain
\[
\hat{E}_{n,2}(t) \Rightarrow \hat{E}_2(t) \equiv \sqrt{F^c(w)F(w)} \int_0^{t-w} \sqrt{\lambda(u)}dB_\theta(u),
\]
(A.29)
where \( B_\theta \) denotes a standard Brownian motion, independent of \( B_\lambda \) and \( \dot{\lambda} \). The detailed proof follows closely the steps in A.7.2 of Aras et al. (2018) and is hence omitted.

Towards obtaining a clean expression for \( \hat{E}_{n,3} \), we seek to simplify the right-hand side of (A.20). In view of (A.21) and (A.27), we deduce
\[
\eta_n(t) = H_n(t) + O(n^{-1/2} \log n) = w + O(n^{-1/2} \log n).
\]
(A.30)
Employing integration by parts, we get
\[
\int_w^t F^c(\eta_n(u))\dot{\lambda}(u - \eta_n(u))\dot{H}_n(u)du + \int_w^t F^c(H_n(u))\dot{\lambda}(u - H_n(u))d\dot{H}_n(u)
= F^c(\eta_n(t))\dot{\lambda}(t - \eta_n(t))\dot{H}_n(t) - \int_w^t \dot{\lambda}(u - \eta_n(u))\dot{H}_n(u)dF^c(\eta_n(u))
- \int_w^t \{ F^c(\eta_n(u))\dot{\lambda}(u - \eta_n(u)) - F^c(H_n(u))\dot{\lambda}(u - H_n(u)) \} d\dot{H}_n(u)
= F^c(w)\dot{\lambda}(t - w)\dot{H}_n(t) + O(n^{-1/2} \log n),
\]
(A.31)
where the last equality is due to (A.30). Combining (A.20) and (A.31) yields
\[
\hat{E}_{n,3}(t) = - \int_w^t f(w)\dot{\lambda}(u - w)\dot{H}_n(u)du - F^c(w)\dot{\lambda}(t - w)\dot{H}_n(t) + O(n^{-1/2} \log n).
\]
(A.32)
Finally, by a standard random-time-change argument, we get
\[
\hat{D}_n(t) \Rightarrow B_\mu \left( \mu \int_w^t m(u)du \right) \text{ in } D \text{ as } n \to \infty,
\]
(A.33)
where \( B_\mu \) denotes a standard Brownian motion independent of anything else. Now substituting (A.16) and (A.32) into (A.24), we obtain
\[
F^c(w)\dot{\lambda}(t - w)\dot{H}_n(t) = - \int_w^t f(w)\dot{\lambda}(u - w)\dot{H}_n(u)du - c[w, t] - \mu \int_w^t c(u)du
+ \hat{E}_{n,1}(t) + \hat{E}_{n,2}(t) - \hat{D}_n(t) + O(n^{-1/2} \log n).
\]

The FCLT for the HWT process then follows by applying the continuous mapping theorem (Theorem 4.1 of Pang et al. (2007)) with the convergence in (A.28), (A.29) and (A.33). Having established the FCLT for the sequence \( \{ \dot{H}_n; n \in \mathbb{N} \} \), the convergence of \( \{ \dot{V}_n; n \in \mathbb{N} \} \) follows by applying convergence-together lemma with (A.9). To establish the claimed FCLT for the queue-length process, we break the right-hand side of (8) into three pieces to get
\[
Q_n(t) = Q_{n,1}(t) + Q_{n,2}(t) + Q_{n,3}(t) + Q_{n,4}(t),
\]
26
where

\[ Q_{n,1}(t) \equiv \sqrt{n} \int_{t-H_n(t)}^t F^c(t-u) d\hat{A}_{n,1}(u), \]
\[ Q_{n,2}(t) \equiv \sqrt{n} \int_{t-H_n(t)}^t F^c(t-u) d\hat{A}_{n,2}(u), \]
\[ Q_{n,3}(t) \equiv \sqrt{n} \int_{t-H_n(t)}^t \int_0^1 \mathbb{1}_{\{x>F(t-u)\}} d\hat{U}_n(\hat{A}_n(u), x), \]
\[ Q_{n,4}(t) \equiv n \int_{t-H_n(t)}^t F^c(t-u) \hat{\lambda}(u) du, \]

for \( \hat{A}_{n,1} \) and \( \hat{A}_{n,2} \) being defined by (A.5). Accordingly we can decompose the diffusion-scaled queue-length process into four components

\[ \hat{Q}_n(t) = \hat{Q}_{n,1}(t) + \hat{Q}_{n,2}(t) + \hat{Q}_{n,3}(t) + \hat{Q}_{n,4}(t), \]

where

\[ \hat{Q}_{n,1}(t) \equiv \int_{t-H_n(t)}^t F^c(t-u) d\hat{A}_{n,1}(u) \Rightarrow \int_{t-w}^t F^c(t-u) \sqrt{\hat{\lambda}(u)} dB_{\lambda}(u), \]  \hspace{1cm} (A.34)
\[ \hat{Q}_{n,2}(t) \equiv \int_{t-H_n(t)}^t F^c(t-u) d\hat{A}_{n,2}(u) \Rightarrow \int_{t-w}^t F^c(t-u) \hat{\lambda}(u) du, \]  \hspace{1cm} (A.35)
\[ \hat{Q}_{n,3}(t) \equiv \int_{t-H_n(t)}^t \int_0^1 \mathbb{1}_{\{x>F(t-u)\}} d\hat{U}_n(\hat{A}_n(u), x) \Rightarrow \int_{t-w}^t F^c(t-u) F(t-u) \hat{\lambda}(u) dB_\theta(u), \]  \hspace{1cm} (A.36)
\[ \hat{Q}_{n,4}(t) \equiv \sqrt{n} \int_{t-H_n(t)}^{t-w} F^c(t-u) \hat{\lambda}(u) du \Rightarrow F^c(w) \hat{\lambda}(t-w) \hat{H}(t). \]  \hspace{1cm} (A.37)

Here the proof for the convergence in (A.34) - (A.36) is very similar to that for (A.28) and (A.29), and the proof for (A.37) follows from a straightforward application of the continuous mapping theorem.


